

الجمهورية الجزائرية الديمقراطية الشعبية

La république algérienne démocratique et populaire

وزارة التعليم العالي و البحث العلمي

Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITÉ BADJI MOKHTAR
- ANNABA -



جامعة باجي مختار - عنابة

Année / 2018

Faculté des sciences
Département de chimie

THÈSE

Présentée en vue de l'obtention du diplôme de Doctorat en sciences

Option : Chimie Analytique et Environnement

THÈME

**Prédiction des températures d'ébullition et des points d'éclair
de différentes classes de solvants : relation quantitative
structure propriété**

Présentée par : M^{me} DIDI Mabrouka

Devant le jury :

Présidente :	M^{me} HAIOUR -BIDJOU Chahra	Professeur	U. B. M -Annaba
Directeur de thèse :	M. MESSADI Djelloul	Professeur	U.B. M - Annaba
Examineur :	M. DJELLAL Ahmed	Professeur	U.B.M- Annaba
Examineur	M. MERDES Rachid	Professeur	U. 8 mai Guelma
Examineur :	M.CHAWKI Mourad	M.C.A	U-Ouargla
Examineur :	M. ZENATI Noureddine	M.C.A	U- Souk- Ahras

REMERCIEMENTS

Ce travail de thèse a été réalisé au sein du laboratoire de sécurité environnementale et alimentaire du département de chimie, faculté des sciences, université d'ANNABA.

*Je tiens à remercier tout particulièrement le Professeur **MESSADI Djelloul**, d'avoir accepté d'être mon encadreur. Je vous remercie Monsieur pour votre disponibilité, votre générosité, votre soutien permanent, votre pédagogie exceptionnelle, votre œil critique et le partage de vos connaissances. Je sais que j'ai pu vivre une thèse exceptionnelle grâce à votre encadrement.*

Permettez-moi de vous dire que plus qu'un encadrant ou un chef, je crois avoir trouvé en vous un grand frère qui m'a aidé aussi bien dans le travail que dans la vie lorsque j'en avais besoin

*Je tiens aussi à remercier Madame **HAIOUR -BIJOU Chahra** d'avoir accepté de présider ma thèse et j'espère que vous trouvez ici l'expression de ma plus vive considération.*

*Je remercie également les membres de mon jury : monsieur **DJELLAL Ahmed**, monsieur **MERDES Rachid**, monsieur **CHAWKI Mourad** et monsieur **ZENATI Nouredine** ; pour avoir accepté d'en faire partie.*

Merci à toute l'équipe du Laboratoire LASEA pour leurs conseils, leurs compétences et leurs amitiés.

Dédicace

Je dédie ce modeste travail à :

** Mes parents, ma source de joie et de courage avec un grand merci pour leur présence, leur soutien et leur amour.*

** Mes sœurs et frères.*

** Mes nièces et neveux.*

** Ma grande famille.*

** Tous mes amis du lycée et de l'université sans exception.*

** Enfin, toute l'équipe du labo **LASEA**.*

Mabrouka DIDJ

المُلخَص

تم تطبيق طريقة QSPR للتنبؤ لدرجات التلألؤ و درجات الغليان لمجموعة من المذيبات المنتمية لأقسام مختلفة و ذلك بالاعتماد على نموذج التراجع المتعدد الخطي (RLM) و طريقة آلات ناقلات الدعم (SVM).

تم استخدام منهجية التراجع المتعدد الخطي (RLM) من اجل التنبؤ بدرجات التلألؤ لمجموعتين من المركبات العضوية : مجموعة من الفحوم الهيدروجينية غير المشبعة مكونة من 173 مركب و مجموعة من الألكانات العادية مكونة من 92 مركب حيث تم تقسيمهما بطريقة تلقائية إلي مجموعتين : مجموعة لحساب النموذج و مجموعة للتأكد الخارجي للنموذج.

تم تطبيق دراسة أخرى QSPR للتنبؤ لدرجات حرارة الغليان لخليط من فئات مختلفة من المذيبات مكونة من 111 مركبا ثم تقسيمها باستخدام خوارزمية CADEX إلي مجموعتين فرعيتين: مجموعة من 89 مركبة لحساب النموذج و 22 مركب لاختبار النموذج. النموذج الخطي الأول الذي تم الحصول عليه مع 5 متغيرات طريق تعظيم الفرق Kxy-Kxx. أما النموذج الثاني غير الخطي (SVM) فقد تم تطويره على نفس المجموعة باستخدام نفس المتغيرات الخمسة السابقة المستعملة في النموذج الخطي.

تم تطبيق تحليل البواقي للكشف عن القيم المتطرفة يليها تشخيص التأثير لاستكمال التحليل السابق.

تم التحقق من المتانة والقدرة التنبؤية للنماذج باستخدام المعاملات الإحصائية للتحقق الداخلي (عبر التحقق من صحة Q^2_{LOO}) و كذلك التحقق من متانة و أداء التنبؤية للنماذج المقترحة على الصعيدين الداخلي و الخارجي (Q^2_{ext}).

بالنسبة للنموذج غير الخطي ، استخدمنا طريقة " آلات ناقلات الدعم " باستغلال دالة الأساس الشعاعي RBF من أجل القيم المثلى لمعاملات ($C = 116$ ؛ $\gamma = 0.075$ ؛ و $\epsilon = 0.115$) ؛ فإن النموذج الناتج يؤدي إلى قدرات تنبؤية داخلية وخارجية جيدة.

تشهد قيم المعلمات الإحصائية (R^2 و Q^2_{ext} و $SDEP$ و $SDEP_{ext}$) على أهمية النماذج المطورة، مع وجود تفوق طفيف لنموذج SVM لكن هذا الفارق الصغير بين نتائج النموذجين يشجعنا على اختيار نموذج MLR الذي يسمح لنا بإعطاء تفسير مقبول وسهل.

الكلمات الدالة:

نقطة التلألؤ - درجة حرارة الغليان - QSPR - الواصفات الجزئية - التراجع المتعدد الخطي (MLR) - آلات ناقلات الدعم (SVM).

SOMMAIRE

	PAGES
SYMBOLES ET ABREVIATIONS	
LISTE DES TABLEAUX	
LISTE DES FIGURES	
INTRODUCTION GENERALE	2

CHAPITRE I : Généralités sur les solvants

I.1. Introduction	7
I.2. Définition d'un solvant	7
I.3. Propriétés physico-chimiques	8
I.3.1. Densité	8
I.3.2. Points d'ébullition	9
I.3.3. Tension de vapeur	9
I.3.4. Chaleur d'évaporation	9
I.3.5. Taux d'évaporation	10
I.3.6. Viscosité	10
I.3.7. Tension superficielle	10
I.3.8. Paramètre de solubilité	11
I.3.9. Point d'éclair	11
I.3.10. Point d'inflammation	12
I.3.11. Limite d'inflammabilité	12
I.3.12. Explosivité	13
I.3.13. Auto-inflammation	14
I.4. Comment mesurer le point d'éclair	15
I.4.1. Domaine d'application	16
I.5. Classification des solvants selon la CLP	16
I.5.1. Comprendre la CLP	16
I.5.2. Classification des liquides inflammables selon le règlement CLP	17
I.5.2.1. Définition	17
I.5.2.2. Critère de classification	17
I.5.2.3. Évolution de l'étiquetage des produits chimiques	17
I.5.2.4. Résumé et comparaison des méthodes d'évaluation	21
I.6. Principales catégories des solvants	24
I.6.1. Hydrocarbures	24
I.6.1. Solvants halogénés	24
I.6.3. Solvants oxygénés	24

I.6.4 .Autres solvants	24
I.7. La toxicité et les maladies professionnelles des solvants organiques	25
I.8. Risque pour l'environnement	25
I.9. Quelques accidents	27
RÉFÉRENCES BIBLIOGRAPHIQUES	29

CHAPITRE II : Étude théorique

II.1. La modélisation moléculaire	32
II.2. Optimisation des molécules	32
II.2.1. La méthode HFR	32
II.2.1.1. Energie d'un micro système représenté par un déterminant de Slater	32
II.2.1. 2. Détermination des orbitales ou équations de Hartree-Fock	35
II.2.1. 3. Equation de Roothaan et Hall	36
II.2.1.4. Quelques remarques sur les processus de résolution des équations de Hartree-Fock-Roothaan	37
II.2.1.5. Détermination des intégrales de la méthode de HFR	38
II.2.2. Méthodes semi-empiriques	39
II.2.2.1.. Définition du semi-empiriques	40
II.2.2.2. Quelques théories semi-empiriques	40
II.2.2.3. Limites et avantages des méthodes semi-empiriques	43
II.2.3. Analyse de distributions de charges	47
II.2.3.1. Analyse de population de Mullikan	47
II.2.3.2. Calcul du moment dipolaire	48
II.2.3.3. Application	50
II.3. La mécanique moléculaire	52
II.3.1. Pas de calcul de champ de force sans définition préalable des types d'atome	52
II.3.2. Forme fonctionnelle des champs de force courants	53
II.3.3. Quelques exemples	54
II.4. Génération des descripteurs moléculaires	55
II.5. Méthodes appliquées pour la sélection d'échantillons	56
II.5.1. Sélection aléatoire des échantillons	56
II.5.2. Algorithme CADEX pour la sélection d'échantillons	57
II.6. Sélection d'un sous ensembles de descripteurs significatifs	57
II.6.1.Principe de sélection par Algorithme génétique	58
II.6.2. Initialisation aléatoire du modèle	58
II.6.3. Étape de croisement	58
II.6.4. Étape de mutation	59
II.6.5. Conditions d'arrêt	59
II.7. Développement des modèles QSPR/QSAR	59
II.7.1. La régression linéaire multiple (RLM)	60
II.7.2. Machine à vecteur support (SVM)	61
II.7.3. Paramètres statistiques d'évaluation d'un modèle QSAR/QSPR	62

II.7.4. Analyse des résidus	66
II.7.5. Diagnostic d'influence	67
II.7.6. Statistique $DFBETAS_{j,i}$	69
II.7.7. $COVRATIO_i$	69
II.7.8. Le domaine d'application du modèle QSPR	70
II.7.9. Test de randomisation	70
RÉFÉRENCES BIBLIOGRAPHIQUES.	71

CHAPITRE III : Résultats et discussions

III.1. Modélisation et prédiction des points d'éclair des hydrocarbures non-saturés en utilisant l'approche hybride algorithme génétique / régression linéaire multiple	77
III .1.1. Introduction	77
III .1. 2. Méthodologies	78
III .1. 2.1. La collecte des donnés	78
III.1.3. Résultats et discussion	82
III.1.3.1.Développement et validation de modèle	82
III.1.3.2.Analyse des résidus	87
III.1.3.3. Diagnostics d'influence	100
III .1. 3.4. Le test de randomisation	111
III .1. 3.5. Étude de contribution des descripteurs au modèle	112
III .1. 3.6. Domaine d'application	113
III .1. 4. Conclusion	114
III.2. Modélisation des points d'éclair d'un ensemble de n-alcane	116
III.2.1.Introduction	116
III.2.2. Données et méthodes de recherche	117
III.2.3. Résultats et discussion	119
III.2.3.1. Calcul du modèle	119
III.2.3.2. Analyse des résidus	122
III.2.3.2. Diagnostic d'influence	130
III.2.3.3. Domaine d'applicabilité	134
III.2.3.4. Test de randomisation	135
III.2.3.5. Validation statistique externe.	135
III.2.4. conclusion.	137
III.3. Modélisation des températures d'ébullition d'un mélange de différentes classes de solvants	138
III.3.1.Données expérimentales et calcul des descripteurs	138
III.3.2. Résultats et discussions	139
III.3.2.1.Choix du modèle linéaire	139
III.3.2.1.A. Analyse des résidus et validation du modèle	146
III.3.2.1.B. Diagnostic d'influence	155

III.3.2.1.C. Évaluation du modèle	162
III.3.2.1.D. Vérification de la qualité d'ajustement	163
III.3.2.1.E. Validation externe	163
III.3.2.1.F. Diagramme de Williams	164
III.3.2.1.G. Test de randomisation	165
III.3.2.1.H. Autres analyses des erreurs	166
III.3.2.2. Modèle non- linéaire (machine à support vecteur)	170
III.3.2.3. Comparaison entre paramètres statistiques des deux modèles	174
III.3.2.4. Comparaison des droites d'ajustement	174
III.3.2.5. Comparaison des distributions des erreurs	175
III.3.3. Conclusion	175
RÉFÉRENCES BIBLIOGRAPHIQUES	177

Conclusion générale 181

ANNEXE I : Présentations des données 184

ANNEXE II : Article publié



*Symboles et
abrégations*

SYMBOLES ET ABREVIATIONS

AG:	Algorithme génétique.
AM1 :	Austin Model 1.
CAS :	Chemical Abstracts Service.
CI :	Configuration Interaction.
CGS :	Centimètre-Gramme-Seconde.
CLOA:	Combinaison Linéaire des Orbitales Atomiques.
CLP :	Classification, emballage et étiquetage (En anglais : Classification, Labelling, Packaging).
CMR :	Molécules Cancérogènes, Mutagènes et Reprotoxiques.
CNDO :	Complete Neglect of Differential Overlap.
COC :	Cleveland Open Cup (pour, Coupelle ouverte).
COV :	Composés Organiques Volatils.
DFITS :	Statistique permettant de mesurer l'influence d'une observation i sur la valeur ajustée.
Di :	Distance de COOK.
d :	Statistique de Durbin-Watson.
di :	Résidu standardisé.
DNA :	DeoxyriboNucleic Acid.
DPD :	Directive Préparations Dangereuses.
DSP :	Directive Substances Dangereuses.
ECOSOC :	Conseil économique et social des Nations unies.

EQM:	Ecart quadratique moyen.
EQMC:	Ecart quadratique moyen calculé sur l'ensemble de calibrage.
EQMP	Ecart quadratique moyen de prédiction.
EQMP _{ext} :	Ecart quadratique moyen calculé sur l'ensemble de validation externe.
ERENAV :	Entreprise Nationale d'Entretien et de Réparation Navale.
e_i :	Résidu : différence entre les valeurs observée (y_i) et estimée (\hat{y}_i).
F :	Statistique de Fisher.
FIV:	Facteur d'inflation de la variance.
INDO:	Intermediale Neglect of Differential Overlap.
GA:	Genetic Algorithm (pour, Algorithme génétique).
HF:	Hartree –Fock.
HFR:	Hartree -Fock-Roothan.
hii :	Eléments diagonaux de la matrice chapeau.
HOMO:	Highest Occupied Molecular Orbital.
INERIS :	Institut National de l'Environnement Industriel et des RISques.
LII :	Limite inférieure d'inflammabilité.
LMO:	Cross-validation by leave-many-out: Validation croisée par omission d'un ensemble d'observations.
LOO:	Cross-validation by leave-one-out: Validation croisée par omission d'une observation.
LSI :	Limite supérieure d'inflammabilité.
LUMO:	Lowest Unoccupied Molecular Orbital.

MM : Mécanique Moléculaire.

n: Dimension de la population (échantillon).

n-p : Nombre de degrés de liberté.

NDDO: Neglect of Diatomic Differential Overlap

OM: Orbitales Moléculaires.

PLS(ou MCP): Moindres carrés partiels.

PMCC : Pensky –Martens Closed Cup (pour, Coupelle fermée).

PM3 : Parametrization Method 3.

PRESS : Somme des carrés des erreurs de prédiction.

p : Nombre de descripteurs en comptant la constante (Nombre de paramètres).

PPP : Pople-Pariser-Parr.

QSAR : Quantitative Structure/ Activity Relationships.
(Relations Quantitatives Structure/ Activité).

QSPR : Quantitative Structure/ Property Relationships.
(Relations Quantitatives Structure/ Propriété).

Q_{boot}^2 : Coefficient de prédiction par la technique du bootstrap.

Q_{LOO}^2 : Coefficient de prédiction par leave one out.

RBF : Radial Basis Function (fonction de base radiale).

REACH : en **R**egistrement, **E**valuation et **A**utorisation des produits **C**himiques.

RLM (MLR): Régression linéaire multiple.

RMSE: Racine de l'écart quadratique moyen (Root Mean Squared Error).

RNA: Réseaux de neurones artificiels.

R^2 : Coefficient de détermination.

r_i :	Résidu studentisé interne.
S :	Erreur standard.
SCE :	Somme des carrés des écarts.
SCF :	Self Consistent Field.
SCT :	Somme des carrés totale.
SGH :	Système général harmonisé.
SVM :	Support Vector Machine (machine à vecteur support)
t :	t de Student.
t_i :	Résidu studentisé externe.
y_i :	Valeur observée.
\hat{y}_i :	Valeur estimée.
$\hat{y}_{(i)}$:	Valeur prédite.



*Liste des
tableaux*

Liste des tableaux

Chapitre	Tableaux	Page
I	Tableau I.1: Critères applicables aux liquides inflammables.	17
	Tableau I.2: Les nouveaux pictogrammes selon le règlement CLP et leur signification.	18
	Tableau I.3: Résumé et comparaison des méthodes d'évaluation.	23
	Tableau I.4: Toxicité de quelques solvants organiques appartenant à différentes familles.	25
III	Tableau III.1: Nomenclature et valeurs des points d'éclair des composés étudiés.	78
	Tableau III.2: Présentation des caractéristiques des descripteurs sélectionnées par le modèle.	84
	Tableau III.3: Matrice de corrélation entre les descripteurs et la température d'éclair.	85
	Table III.4: Les paramètres statistiques du modèle développé.	86
	Tableau III.5: Résidus caractéristiques et valeurs estimées des points d'éclair.	90
	Tableau III.6 : Diagnostics d'influence.	101
	Tableau III.7: Composés étudiés et valeurs des températures d'éclair expérimentales.	117
	Tableau III.8: Classe et définition des deux descripteurs moléculaires.	121
	Tableau III.9: Résidus caractéristiques et valeurs estimées des points d'éclair.	125
	Tableau III.10: Diagnostics d'influence.	130
	Tableau III.11: Valeurs des paramètres statistiques pour tout l'ensemble étudié.	136
	Tableau III.12: Valeurs des paramètres statistiques pour le modèle inverse.	136
	Tableau III.13: Paramètres statistique des modèles obtenus.	140
	Tableau III.14: Les composés étudiés, les températures d'ébullition et les descripteurs sélectionnés.	143
	Tableau III.15: Paramètres statistiques obtenus.	146
	Tableau III.16: Résidus caractéristiques et valeurs estimées des températures d'ébullition des solvants étudiés.	149
	Tableau III.17: Diagnostics d'influence.	156
	Tableau III.18: Les paramètres statistiques pour modèle RLM.	162
	Tableau III.19: Températures d'éclair expérimentales et prédites et les valeurs des erreurs.	166

Liste des tableaux (suite)


Chapitre	Tableaux	Page
III	Tableau III.20: Classes des solvants, leurs distributions dans chaque intervalle d'EAR% et les EAR % associées.	169
	Tableau III.21: Valeurs des températures d'ébullition estimées et prédites pour l'ensemble de calibrage.	170
	Tableau III.22: Les paramètres statistiques pour les deux modèles.	174



*Liste des
figures*

Liste des figures

Chapitre	Figure	Page
I	Figure I.1: Les six conditions pour une explosion	14
	Figure I.2 : Signification d'une étiquette	20
II	Figure II.1: Déterminants de Slater excités générés à partir d'une référence HF.	45
	Figure II.2: Les indices électroniques de la méthode des orbitales moléculaires et leurs application.	51
III	Figure III.1: Droite d'ajustement des Tec prédites en fonction des Tec expérimentales pour les ensembles de calibrage et de test.	87
	Figure III.2: Test de randomisation associé au modèle QSPR.	112
	Figure III.3: Contributions relatives des descripteurs sélectionnés dans le modèle RLM.	112
	Figure III.4: Diagramme de Williams.	114
	Figure III.5: Droite d'ajustement des Tec prédites en fonction des Tec expérimentales pour les ensembles de calibrage et de test.	122
	Figure III.6: Diagramme de Williams	134
	Figure III.7: Test de randomisation	135
	Figure III.8: Choix de la taille du modèle.	141
	Figure III.9: Droite d'ajustement des Teb prédites en fonction des Teb expérimentales pour les ensembles de calibrage et de test.	163
	Figure III.10: Diagramme de Williams.	164
	Figure III. 11: Test de randomisation.	165
	Figure III. 12: Droites d'ajustement de modèles RLM et SVM.	174
	Figure III. 13: Distributions des résidus en fonction des valeurs prédites : (A)-RLM, (B)-SVM.	175



*Introduction
générale*

Introduction générale

Un danger chimique peut se mesurer par la toxicité du produit ou par sa capacité à produire des dégâts et des effets néfastes pour la santé humaine et l'environnement. Ces effets sur la santé peuvent concerner aussi bien le travailleur qui les produit, ou les utilise, que le consommateur final. Plus généralement, c'est l'ensemble de la population qui peut être exposée via le relargage de substances dans l'environnement.

Chaque année on note l'apparition d'un million de nouvelles molécules. Vingt deux millions seulement sont répertoriées auprès de la banque de données du Chemical Abstracts Service (CAS), cent mille sont commercialisées dont seulement 5% ont des propriétés (physicochimiques, activités biologiques, etc...) connues [1].

Afin de mieux connaître et maîtriser les risques liés à l'utilisation de ces produits chimiques, plusieurs réglementations, essentiellement d'origine communautaire, les encadrent en fonction de leurs usages. On distingue les règlements européens, adoptés au niveau de l'Union Européenne (UE) et applicables directement, les directives, transposées en droit français pour être applicables, d'autres enfin sont purement nationales.

Le règlement, enregistrement, évaluation et autorisation des substances chimiques (REACH), entré en vigueur le 1er juin 2007 en Europe dans le but de recueillir un grand nombre d'informations sur les propriétés des substances chimiques produites ou importées en quantité supérieure à 1 tonne/an. Ce règlement complexe avec de fortes obligations vis-à-vis des industriels, constitue un outil fondamental pour les pouvoirs publics et la société civile pour améliorer à long terme le bien-être de la population en termes de santé et d'environnement.

Une autre mesure réglementaire importante concerne la classification, l'étiquetage et l'emballage des produits chimiques : il s'agit du règlement européen 1272/2008 CE dit CLP (en anglais *Classification, Labelling, Packaging*). Entré en vigueur le 20 janvier 2009, ses dispositions étaient prévues pour être entièrement applicables au 1er décembre 2010 pour les substances et au 1er juin 2015 pour les mélanges. Ce règlement européen est basé sur les dispositions établies par le "Système général harmonisé" (SGH) promu par le Conseil économique et social des Nations unies (ECOSOC) en juillet 2003 [2].

En Algérie [3], l'arrêté interministériel du 13 Safar 1437 correspondant au 25 novembre 2015 fixant la liste et la classification des matières et produits chimiques dangereux (article.1, article. 2, article. 3). Le présent arrêté a pour objet de fixer la liste et la classification des matières et produits chimiques dangereux. La liste citée indique pour chaque matière ou produit chimique dangereux son numéro d'identification selon la classification de

l'Organisation des Nations-Unies (ONU) et sa classe de risque principal telle que définie à l'article 3.

Les propriétés physico-chimiques sont utilisées pour caractériser les produits chimiques qu'ils soient naturels ou de synthèse. En prévention, elles permettent d'évaluer la dangerosité d'un produit chimique (inflammabilité, volatilité, pouvoir corrosif, pouvoir oxydant...) et les risques qui en découlent (que ce soit lors de sa manipulation, son stockage ou son élimination) tels que les risques d'incendie, les risques d'explosion, corrosivité, réactivité chimique et instabilité...

La connaissance des caractéristiques des produits (température d'ébullition, point d'éclair, pression de vapeur saturante, température d'auto-inflammation, domaine d'inflammabilité ou d'explosivité) est essentielle pour la maîtrise du risque. De nombreux exercices de mises en situation permettent d'appliquer immédiatement ces acquis théoriques.

Dans la plupart des cas, il est excessivement cher d'obtenir de telles informations expérimentalement et le recours à l'expérience pour pouvoir les identifier devient impossible. Par conséquent, les compagnies et les agences régulatrices se tournent vers la prédiction de ces propriétés à travers l'usage des relations quantitatives structure / propriété (QSPR).

Les Relations Quantitatives Structure- Activité/ Propriété (QSAR/QSPR) sont devenues un puissant outil théorique, alternatif à la mécanique quantique, pour la description et la prédiction des propriétés des systèmes moléculaires complexes dans différents environnements.

Les relations entre les structures des molécules et leurs propriétés ou activités sont généralement établies à l'aide de méthodes de modélisation par apprentissage statistique.

Les techniques usuelles reposent sur la caractérisation des molécules par un ensemble de descripteurs, nombres réels mesurés ou calculés à partir des structures moléculaires. Il est alors possible d'établir une relation entre ces descripteurs et la grandeur modélisée [4].

Dans le présent travail, nous nous sommes intéressés à la modélisation de deux propriétés physico- chimiques: la température d'éclair et la température d'ébullition de différentes classes de solvants. Nous avons appliqué les techniques QSPR les plus courantes pour établir des modèles, en utilisant soit une analyse de régression multilinéaire (RLM), soit une régression non linéaire par Machine à Vecteur Support (SVM pour Support Vector Machine).

Notre mémoire de thèse comporte en plus d'une introduction, d'une conclusion générale, et des annexes, trois chapitres :

- Dans le premier chapitre, nous avons présenté des généralités sur les solvants

- Le deuxième chapitre expose la base théorique exploitée. Nous avons décrit la modélisation moléculaire et ses principes. Nous y avons également développé les connaissances théoriques de base utilisées tout au long de ce travail: algorithmes génétiques, régression multilinéaire, machine à vecteurs de support, et les paramètres statistiques utilisés pour l'évaluation de la qualité des modèles obtenus.
- Dans le troisième chapitre, nous présentons et nous discutons les quatre modèles développés :
 - Un modèle AG/MLR développé pour modéliser la température d'éclair d'une série d'hydrocarbures non-saturés.
 - Un modèle AG/MLR développé pour modéliser la température d'éclair d'une série de n-alcane.
 - Deux modèles : un modèle linéaire AG/MLR et l'autre non linéaire AG/SVR pour la modélisation d'une série de solvants organiques formés de différentes classes, avec une comparaison des résultats obtenus selon les deux modèles.

Une analyse des résidus a été effectuée pour les trois modèles linéaires pour les trois ensembles étudiés dans le but de vérifier la validité du modèle d'une part et pour repérer les observations aberrantes et les observations qui jouent un rôle important dans la détermination de la régression.

RÉFÉRENCES BIBLIOGRAPHIQUES

- [1] Gramatica, P., Applicazione delle metodologie QSAR a problematiche ambientali di inquinanti organici, Università degli Studi di Bologna - Dottorato in Chimica Ind –2004.
- [2] Journal officiel de l'Union européenne, REGLEMENT (CE) N° 1272/2008 DU PARLEMENT EUROPÉEN ET DU CONSEIL du 16 décembre 2008.
- [3] JOURNAL OFFICIEL DE LA REPUBLIQUE ALGERIENNE N° 07. 28 Rabie Ethani 1437. 7 février 2016. pp21-22
- [4] Aurélie, G.S.A., Thèse de doctorat, Une nouvelle méthode d'apprentissage de données structurées : applications à l'aide à la découverte de médicaments, Université de Paris 6, 2008.



Chapitre I



**I : Généralités sur les
solvants**

I.1.Introduction

Le champ d'application des solvants est des plus étendu car l'économie moderne y a recours pour la préparation d'une foule de produits nouveaux dont l'utilisation déborde dans la vie courante.

Chaque année le nombre de ces substances mises sur le marché et à la disposition des industriels augmente à travers le monde.

Utilisés dans de nombreux domaines, les solvants organiques entrent dans la composition de divers produits tels que les peintures, les encres, les pesticides, les dégraissants, la production des textiles artificiels, les diluants et les colles. Parmi les matières premières utilisées au cours de la fabrication, les adhésifs, notamment les adhésifs solides et liquides naturels et les solutions adhésives préparées à partir de solvants organiques, représentent les plus importants risques professionnels [I.1]. Depuis longtemps, les procédés d'extraction ont employé les solvants comme agent d'extraction.

I.2.Définition d'un solvant

Ayant à l'esprit le cas de l'utilisation des solvants en vue de la fabrication des laques cellulosiques, Durrans [I.2] propose la définition suivante : « Le solvant est un moyen de transfert d'un solide d'une place à une autre de manière convenable. Lorsque le transfert a été réalisé, le solvant ne présente plus d'intérêt ; on l'élimine donc aussi rapidement et aussi complètement que possible ».

Mais cette définition a le défaut de viser surtout une application particulière. Une définition technique plus large pourrait être celle-ci: « Un solvant est un composé volatil liquide, capable de se charger d'autres substances en ne formant qu'une seule phase liquide, ledit composé étant capable de restituer, par simple évaporation, ces substances inaltérées et amenées éventuellement sous un état physique plus convenable [I.3-p3].

En terme général, un solvant est une substance qui sert à dissoudre une autre substance. Dans le contexte industriel, on se limite aux solvants organiques, c'est-à-dire ceux qui contiennent au moins un atome de carbone dans leurs structures moléculaires. D'après Cohr [I.4], un solvant organique est un composé chimique ou mélange qui est liquide entre 0° C et 250°C approximativement, qui est volatil et relativement inerte chimiquement.

Les solvants sont utilisés industriellement pour extraire, dissoudre ou suspendre des substances généralement insolubles dans l'eau (l'eau n'est donc pas un solvant organique) ou pour modifier les propriétés physiques d'un matériau.

Le concept de solvant organique ne doit pas être confondu avec celui des corps organiques volatils (COV) que l'on retrouve dans la réglementation environnementale visant à protéger la qualité de l'atmosphère. Le terme solvant a une dimension utilitaire alors que les COV sont définis en terme de réactivité photochimique dans l'atmosphère et de tension de vapeur minimale, généralement autour de 13,3 Pa (soit 0,1 mm Hg) à 25°C [I.5].

I.3. Propriétés physico-chimiques

En plus du coût, les propriétés physico-chimiques, liées à la performance technique, sont les paramètres cruciaux qui déterminent les types d'utilisation industrielle des solvants et leur mise en œuvre. Elles permettent également de prévoir une partie de leur comportement environnemental, ces propriétés sont liées spécifiquement aux dangers de manipulation, d'explosion et d'incendie.

I.3.1 Densité

La densité d'une substance est égale à la masse volumique de la substance divisée par la masse volumique du corps de référence à la même température. Pour les liquides et les solides, l'eau est utilisée comme référence, pour les gaz, la mesure s'effectue par rapport à l'air. Elle est notée **d** et n'a pas d'unité (grandeur physique sans dimension).

$$d = \frac{\rho_{\text{substance}}}{\rho_{\text{H}_2\text{O}}} \quad (\text{I.1})$$

La masse volumique de l'eau est mesurée à la température de 4°C, qui correspond à une température où sa masse volumique passe par un maximum. On indique cette température de référence en mettant 4 en indice. La notation devient alors d_4 . Pour des raisons pratiques, la mesure de la masse volumique de la substance s'effectue à la température ambiante et généralement à 20°C. Il est donc usuel de noter la densité d'un solide ou d'un liquide en indiquant les 2 températures: d_4^{20} qui signifie donc « densité de la substance à 20°C par rapport à celle de l'eau à 4°C ».

À l'exception des solvants halogénés la plupart des solvants sont plus légers que l'eau. Ceci explique pourquoi la majorité des feux de solvants ne peuvent être étouffés par l'eau.

I.3.2: Point d'ébullition

Le point d'ébullition est la température à laquelle la pression de vapeur du liquide est égale à celle de la pression atmosphérique normale (101,325 kPa), soit la température à laquelle le solvant passe de l'état liquide à l'état gazeux. L'unité de mesure de cette variable dans le système international est le Kelvin mais, en pratique, on utilise le degré Celsius.

Le contrôle de la température d'ébullition à la pression ordinaire est facile à faire au laboratoire et ne demande que quelques instants.

Les différents solvants sont classés suivant leur ordre de volatilité en 3 catégories : solvants légers ($T_{eb} < 100^{\circ}\text{C}$); solvants moyens ($100^{\circ}\text{C} < T_{eb} < 150^{\circ}\text{C}$); solvants lourds ($T_{eb} > 150^{\circ}\text{C}$). Il faut souligner que la présence de certaines impuretés est capable de modifier le comportement d'un solvant lors de la distillation; elles peuvent en particulier abaisser son point d'ébullition et le faire ainsi passer d'une classe à la classe inférieure [I.3-p31].

I.3.3. Tension de vapeur

La tension ou pression de vapeur saturante est la pression exercée par la vapeur lorsqu'elle est à l'équilibre avec le liquide. L'unité de mesure de cette variable est le kilopascal (kPa) et on utilise souvent le millimètre de mercure (mmHg ou Torr, $1\text{kPa} = 7,500\text{ mmHg}$). La tension de vapeur est rapportée le plus souvent à 25°C ; plus la tension de vapeur est élevée et plus le solvant a une tendance naturelle à s'évaporer.

La détermination des tensions de vapeur d'un liquide nécessite un appareillage spécial, et demande beaucoup plus de soin et de temps que celle d'une température d'ébullition. On opère le plus souvent avec une chambre barométrique à mercure. Il faut une parfaite uniformisation de la température au moment de la mesure de la pression de vapeur et il est souhaitable d'opérer sur des échantillons de produits purs [I.3-p22].

D'une manière générale il est évident que les liquides sont d'autant plus dangereux au point de vue de l'inflammabilité que leur tension de vapeur est plus forte

I.3.4. Chaleur de vaporisation

La chaleur de vaporisation ou chaleur latente de vaporisation d'un solvant est la quantité de chaleur requise pour vaporiser une quantité définie de solvant.

Ses unités sont le kilojoule par mole dans le système international (kJ/mol) et la kilocalorie par mole dans le système CGS (kcal/mol). Ce paramètre est utile à connaître notamment pour la comparaison des solvants quant à leur exigence énergétique dans le domaine du dégraissage à la vapeur des surfaces métalliques.

I.3.5. Taux d'évaporation

Le taux d'évaporation absolu d'un solvant est la quantité de matière qui s'évapore d'une surface par unité de temps. Dans la littérature le taux d'évaporation relatif est rapporté à un solvant de référence, soit l'éther éthylique ou l'acétate de butyle normal.

Cette variable qui dépend des conditions environnementales n'a pas d'unité. Les taux d'évaporation des solvants sont très utilisés dans la formulation des peintures, des adhésifs et des encres.

I.3.6. Viscosité

Le coefficient de viscosité ou viscosité dynamique est défini comme étant la force nécessaire au déplacement d'une surface plane de liquide de 1 cm^2 avec une vitesse de 1 cm/s par rapport à une autre surface plane du même liquide qui lui est parallèle à une distance de 1 cm [4]. L'unité de la viscosité dynamique dans le système international est le Pascal seconde (Pa.s). En pratique on utilise également la poise ($1\text{P} = 0,1 \text{ Pa.s} = 0.1 \text{ N.s/m}^2 = 1 \text{ dyne.s/cm}^2$) et la centipoise (cP). La viscosité est importante lors de la formation d'un mélange ainsi que lors du choix du mode d'application, par exemple par pulvérisation.

I.3.7. Tension superficielle

La tension superficielle est définie comme la résultante des forces intermoléculaires s'exerçant sur les molécules à la surface du liquide. On peut imaginer que les molécules situées à l'intérieur du liquide sont soumises à des forces différentes de celles subies par sa surface: en effet, dans la phase liquide contrairement à la phase gazeuse, les molécules se touchent. Bien souvent, ces molécules ne sont pas électriquement neutres. Il s'ensuit que la résultante de ces forces ne peut pas être nulle.

Lorsque deux liquides dissemblables sont en contact, ces forces intermoléculaires modifieront la forme de l'interface jusqu'à ce que l'énergie potentielle de tout le système moléculaire atteigne un minimum.

Ses unités de mesure sont : dans le système international le newton par mètre (N/m), et dans le système CGS la dyne par centimètre (1 dyne/cm = 0,001N/m).

La tension superficielle est un paramètre utile à connaître pour évaluer la tendance que possède un liquide à s'étaler sur une surface.

I.3.8. Paramètre de solubilité

La capacité de solubilisation d'un solvant pour un soluté donné, également appelé son pouvoir solvant, est une donnée essentielle d'un solvant.

Suite à une étude faite en 2005 [I.6], le concept de paramètre de solubilité a été introduit dans des travaux théoriques développés pendant la première moitié du vingtième siècle à partir d'une base thermodynamique par Joël Hildebrand. Il définit le paramètre de solubilité global d'une substance comme étant la racine carrée de l'énergie molaire de cohésion du système par unité de volume (volume molaire).

Le paramètre de solubilité est exprimé en $(\text{J}/\text{cm}^3)^{1/2}$, c'est une fonction de l'énergie molaire de vaporisation (E) et du volume molaire (V) du liquide.

Un solvant dissout bien le soluté lorsque leurs paramètres de solubilité sont identiques ou ayant un pouvoir de solubilisation très proche.

Il faut observer qu'au point de vue incendie, mieux vaut avoir à faire à des liquide solubles dans l'eau qu'à des corps non miscibles. En effet les premiers donnent naissance à des foyers qu'on peut attaquer efficacement avec la lance à eau classique et la dilution qui résulte de cet arrosage abondant contribue directement à diminuer la puissance de la flamme. Au contraire, lorsqu'il s'agit de liquides insolubles dans l'eau, comme les hydrocarbures par exemple, même l'envoi d'une grande quantité d'eau ne suffit pas à éteindre l'incendie. Au contraire, le plus souvent, l'arrosage à la lance ne fait qu'accroître la surface du liquide enflammé en contact avec l'air, tout en risquant d'augmenter le danger de propagation du sinistre par écoulement des liquides enflammés dans les caniveaux de l'usine.

I.3.9. Point d'éclair

Le point d'éclair est la température la plus basse, corrigée pour une pression de 101,325 kPa, à laquelle le liquide d'essai dégage des vapeurs, dans les conditions définies dans la

méthode d'essai, en quantité telle qu'il en résulte dans le récipient d'essai un mélange vapeur/air inflammable [I.7].

C'est le phénomène qui se produit lorsqu'un liquide organique déterminé a été porté à une température suffisante pour que les vapeurs émises puissent être enflammées mais sans que la flamme persiste ou que le liquide s'enflamme au contact de l'air ambiant par l'approche d'une petite flamme. Le point d'éclair caractérise en quelque sorte la première manifestation d'un liquide dans l'aptitude à l'inflammation. Sa détermination se ramène à la mesure d'une température.

Au point d'éclair le mélange solvant - air contient deux fois plus d'oxygène que celui que nécessiterait la combustion complète [I.3-p55].

I.3.10. Point d'inflammation

Le point d'inflammation est la température la plus basse à laquelle un liquide émet suffisamment de vapeurs pour former avec l'air ambiant un mélange inflammable dont la combustion une fois débutée puisse s'entretenir d'elle-même après retrait de la source d'allumage. Il est supérieur au point d'éclair de quelques degrés [I.3].

I.3.11. Limite d'inflammabilité

Avant de définir la limite d'inflammabilité, nous allons rappeler quelques définitions [I.8]:

a- Combustible

Toute substance susceptible de brûler, c'est-à-dire pouvant être partiellement ou totalement détruite par le feu, est considérée comme combustible. Les solides et les liquides ne brûlent pas en tant que tels. Ce sont les gaz et les vapeurs qu'ils émettent qui brûlent.

b- Comburant

Un comburant est le corps qui provoque et entretient la combustion du combustible; le plus souvent, le comburant est constitué par l'oxygène présent dans l'air ambiant ; la réaction de combustion est alors une oxydation, mais il existe d'autres comburants (halogènes, soufre, phosphore); si l'oxygène est le comburant, sa concentration diminue très rapidement dans l'atmosphère (par phénomène de consommation oxydative) et expose les victimes au risque d'asphyxie. L'oxygène peut se trouver soit à l'état pur, soit en mélange avec d'autres gaz, soit lors de la décomposition de certains produits chimiques. Dans la plupart des cas, le

comburant est l'oxygène de l'air ambiant (environ 21 % d'oxygène 79 % d'azote). Pour que l'air soit un comburant efficace, il faut qu'il contienne plus de 15 % d'oxygène.

c- Combustion

La combustion est un processus d'oxydation qui se produit par réaction chimique entre deux corps un combustible et un comburant pour donner naissance à un ou plusieurs corps différents des premiers « les produits de combustion ». Il s'agit d'une réaction chimique s'accompagnant d'un dégagement de chaleur. On parle de combustion lente lorsque l'élévation de température devient perceptible, mais sans atteindre une température donnant une lumière visible. Les combustions vives correspondent à une réaction provoquant des températures élevées. À l'extrême, quand la vitesse de propagation devient extrêmement grande, on parle d'explosion.

Les concentrations limites d'un gaz ou d'une vapeur combustible, dans l'air ou dans tout autre comburant en aval et en amont desquelles la propagation de la flamme n'est pas possible, sont appelées «limites d'inflammabilité». Si le mélange est trop pauvre en combustible, l'inflammation ne se produit pas. Le pourcentage est alors au-dessous de la limite inférieure d'inflammabilité (LII). Au-dessus de ce seuil, le mélange combustible-comburant pourra brûler tant que l'on n'aura pas dépassé un taux maximum de combustible au-delà duquel le mélange serait trop pauvre en comburant; ce second seuil est la limite supérieure d'inflammabilité (LSI).

L'intervalle entre la limite inférieure et la limite supérieure d'inflammabilité s'appelle «domaine d'inflammabilité ou intervalle d'inflammabilité ». Celui-ci varie fortement selon les gaz ou vapeurs combustibles, la température, le taux d'oxygène et la pression. Une matière combustible sous forme gazeuse ou de vapeur ne peut exploser que si elle est mélangée à de l'air avec une concentration comprise entre LII et LSI [I.8]. Ces valeurs sont généralement exprimées en pourcentage du volume de gaz inflammable dans le volume total du mélange.

I.3.12. Explosivité

Les explosions peuvent être soit d'origine physique (par exemple, éclatement d'un récipient dont la pression intérieure est trop grande), soit d'origine chimique, cette dernière est la résultante d'une réaction chimique.

Une réaction d'origine chimique est une réaction rapide de combustion ou de décomposition entraînant une élévation de température et /ou de pression. Les six conditions à remplir pour une explosion sont réunies et symbolisées par un hexagone sont (Figure I.1) :

- La présence d'un comburant(en général l'oxygène de l'air) ;
- La présence d'un combustible ;
- La présence d'une source d'inflammation ;
- Un combustible sous forme gazeuse, d'aérosol ou poussières en suspension ;
- L'obtention d'un domaine d'explosivité (domaine de concentrations de combustible dans l'air comprises entre la LII et LSI à l'intérieur duquel les explosions sont possibles.
- Un confinement suffisant (en l'absence d'un confinement, on obtient un phénomène de combustion rapide sans effet notable de pression, type boule de feu [I.9].

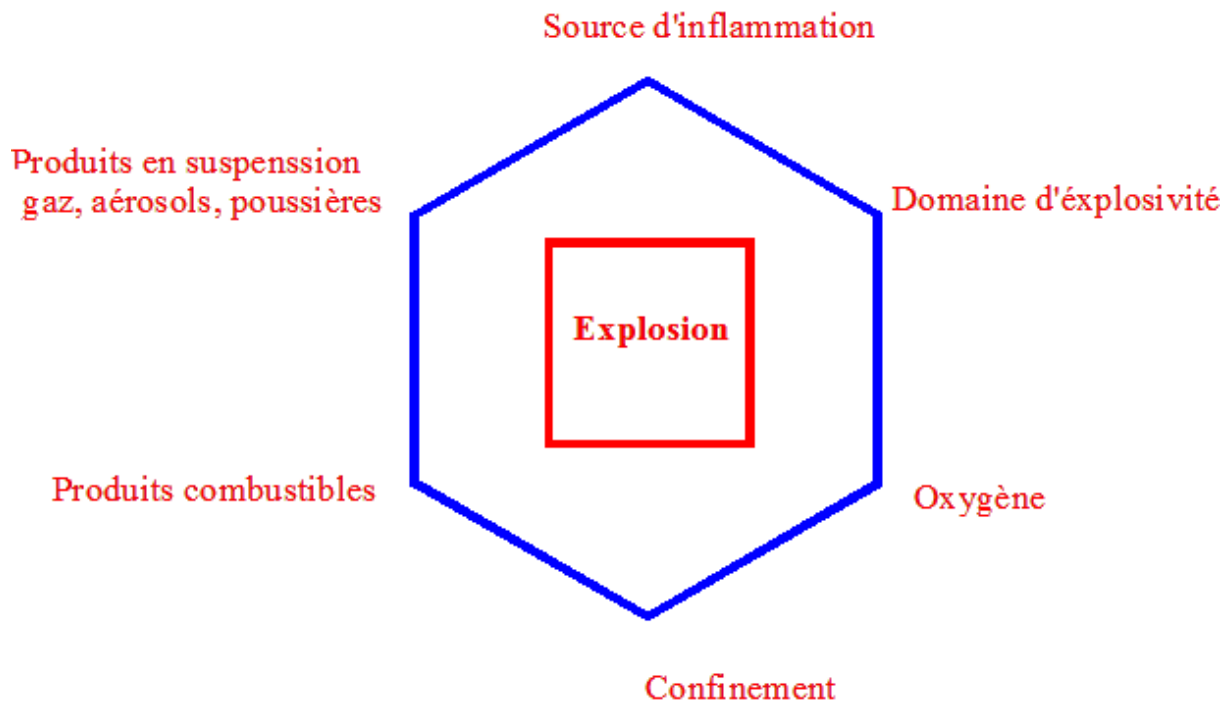


Figure II.1: Les six conditions pour une explosion.

I.3.13. Auto-inflammation

Un mélange d'air et de vapeur d'une substance inflammable porté à une température assez élevée peut déclencher une combustion spontanée en l'absence d'une flamme ou d'une étincelle. Ce phénomène est appelé « auto-inflammation ».

Ici, il est important de noter la grande influence de la nature des parois sur le déclenchement de cette réaction d'oxydation. Le verre a généralement une influence plus hâtive que les métaux [I.10].

I.4. Comment mesurer le point d'éclair ?

La sécurité dans l'utilisation et le stockage de produits inflammables et des mélanges liquides combustibles est bien nécessaire à cause des accidents dramatiques qui se produisent régulièrement sous forme d'une série d'explosions d'huiles essentielles. L'importance de la sécurité du transport des liquides inflammables et des combustibles ainsi que le risque d'explosion des liquides se caractérise principalement par leur point d'éclair ou température d'éclair (Tec). Le point d'éclair est un concept pétrolier et les premiers appareils qui permettent de définir un point d'éclair ont été décrits dans des normes pétrolières. Il faut donc toujours préciser l'appareil qui a été utilisé quand on donne une valeur de point d'éclair. La valeur dépend non seulement de l'appareil mais également de la bonne vue de l'opérateur qui doit déceler le début d'inflammation. On dénombre deux catégories de test : test à coupelle ouverte (COC: Cleveland Open Cup), et test à coupelle fermée.

Le point d'éclair en coupelle fermée se mesure à l'aide d'un appareil dit de Pensky-Martens (PMCC : Pensky –Martens *Closed Cup*). Un appareil semi-automatique de ce type est constitué d'une coupelle que l'on peut remplir du liquide dont on désire connaître le point d'éclair. On referme ensuite la coupelle. Le couvercle est muni d'un thermomètre dont l'embout se situe au-dessus du liquide dans les vapeurs. L'appareil dispose d'un chauffage qui permet d'élever la température degré par degré. Chaque fois que la température atteint un degré supérieur, une flamme est plongée dans les vapeurs. S'il y a inflammation, c'est que le point d'éclair est atteint, dans le cas contraire l'appareillage continue d'augmenter la température du liquide [I.11].

Le règlement sur les matières dangereuses, prévoit un test d'inflammabilité pour des échantillons liquides contenant ou non des solides en suspension ou en solution. L'inflammabilité d'une matière s'évalue en mesurant la température la plus basse à laquelle les vapeurs du liquide s'enflamment en présence d'une flamme.

Selon le règlement sur les matières dangereuses, toute matière liquide ou semi-liquide, autre qu'une boisson alcoolisée, est considérée comme inflammable si le point d'éclair est égal ou inférieur à 60 °C. De plus, le règlement sur les matières dangereuses permet l'utilisation d'huile usée à des fins énergétiques pour autant que le point d'éclair soit de 38 °C ou plus élevé.

I.4.1. Domaine d'application

Cette méthode s'applique à la détermination du point d'éclair d'une huile usée, d'un liquide ou d'un liquide contenant des solides en suspension ou en solution.

Le domaine d'application se situe entre 25 °C et 80 °C. Un point d'éclair peut être mesuré à plus basse température, en abaissant la température de l'échantillon à l'aide de glace sèche par exemple et en chauffant par la suite.

I.5. Classification des solvants selon le règlement CLP

I.5.1. Comprendre le règlement CLP

CLP est l'abréviation de l'expression anglaise «Classification, Labelling and Packaging» ou «classification, étiquetage et emballage». Le règlement CLP est entré en vigueur en janvier 2009 et la méthode de classification et d'étiquetage des produits chimiques qu'il introduit repose sur le règlement du Système Général Harmonisé (SGH) des Nations unies.

Le règlement remplace progressivement deux actes législatifs antérieurs, à savoir la directive «Substances dangereuses» (dSD) et la directive «Préparations dangereuses» (dPD). La période de transition a pris fin en 2015.

Le règlement CLP a pour objet d'assurer que les dangers présentés par les substances chimiques soient clairement communiqués aux travailleurs et aux consommateurs de l'Union européenne grâce à la classification et à l'étiquetage des produits chimiques.

Avant de procéder à la mise sur le marché de produits chimiques, l'industrie doit déterminer les risques potentiels de ces substances et mélanges pour la santé humaine et l'environnement et les classer conformément aux dangers identifiés. Les produits chimiques dangereux doivent aussi être étiquetés selon un système normalisé de sorte que les travailleurs et les consommateurs soient informés de leurs effets avant de les manipuler.

Grâce à ce processus, les dangers des produits chimiques sont communiqués en recourant à des mentions types et à des pictogrammes imprimés sur les étiquettes et les fiches de données de sécurité. Ainsi, lorsqu'un fournisseur identifie une substance comme présentant une «toxicité aiguë de catégorie 1 (oral)», l'étiquetage inclura la mention de danger «mortel en cas d'ingestion», le mot «Danger» et le pictogramme comportant une tête de mort et deux tibias [I.12].

I.5.2. Classification des liquides inflammables selon le règlement CLP

I.5.2.1. Définition

Selon les prescriptions relatives à la classification et à l'étiquetage des substances et mélanges dangereux, on entend par « liquide inflammable », un liquide ayant un PE $\leq 60^{\circ}\text{C}$ [I.12].

I.5.2.2. Critères de classification

Les critères de classification de la classe des liquides inflammables reposent comme dans le cas du système préexistant sur la mesure du PE et du point initial d'ébullition (T_{eb}). On distingue trois catégories dans cette classe en fonction des différentes valeurs indiquées dans le tableau I.1.

Tableau I.1: Critères applicables aux liquides inflammables

Catégorie	Critères
1	Le point d'éclair est $< 23^{\circ}\text{C}$ et le point initial d'ébullition est $\leq 35^{\circ}\text{C}$
2	Le point d'éclair est $< 23^{\circ}\text{C}$ et le point initial d'ébullition est $> 35^{\circ}$
3	Le point d'éclair: $23^{\circ}\text{C} < \text{PE} \leq 60^{\circ}$ (*)

(*) Aux fins de ce règlement, les gazoles, les carburants diesel et huiles de chauffage légères dont le point d'éclair compris entre 55°C et 75°C peuvent être considérés comme relevant de la catégorie 3.

1.5.2.3. Évolution de l'étiquetage des produits chimiques

Depuis l'arrêté du 20 avril 1994 modifié, les symboles et indications de danger utilisés pour l'étiquetage des substances et préparations dangereuses sont définis par la réglementation française. Ces symboles, noirs sur un fond carré jaune-orangé, vont être progressivement remplacés par un nouveau système.

Le règlement (CE), dit "règlement CLP [I.12] définit de nouvelles règles européennes de classification, d'étiquetage et d'emballage des produits chimiques. Entré en vigueur le 20 janvier 2009, ce règlement prescrit de nouveaux pictogrammes de danger.

Les anciens et nouveaux systèmes coexisteront durant une période transitoire. La mise en application du nouveau règlement est devenue obligatoire à partir du 1er décembre 2010 pour les substances et du 1er juin 2015 pour les mélanges. Ces pictogrammes ont la forme carrée debout sur la pointe et comportent un symbole en noir sur fond blanc dans un cadre rouge

suffisamment épais pour être clairement visible. Ils possèdent chacun un code différent composé de SGH suivi d'un numéro (01, 02, ...,09). On lira, sous le pictogramme, l'un ou l'autre de ces mots : « DANGER » ou « ATTENTION » (voir le tableau I.2).

Tableau I.2: Les nouveaux pictogrammes selon le règlement CLP et leur signification [I.12].





Symbole	Signification
 SGH01	Ces produits peuvent exploser au contact d'une flamme , d'une étincelle, d'électricité statique, sous l'effet de la chaleur, d'un choc, de frottements...
 SGH02	Ces produits peuvent s'enflammer , suivant le cas: * au contact d'une flamme, d'une étincelle, d'électricité statique... ; * sous l'effet de la chaleur, de frottements... ; * au contact de l'air ; * au contact de l'eau, s'ils dégagent des gaz inflammables (certains gaz s'enflamment spontanément, d'autres au contact d'une source d'énergie , flamme, étincelle...).
 SGH03	Ces produits peuvent provoquer ou aggraver un incendie, ou même provoquer une explosion s'ils sont en présence de produits inflammables. On les appelle des produits comburants .
 SGH04	Ces produits sont des gaz sous pression contenus dans un récipient. Certains peuvent exploser sous l'effet de la chaleur : il s'agit des gaz comprimés, des gaz liquéfiés et des gaz dissous. Les gaz liquéfiés réfrigérés peuvent, quant à eux, être responsables de brûlures ou de blessures liées au froid appelées brûlures et blessures cryogéniques.
 SGH05	Ces produits sont corrosifs , suivant les cas : * ils attaquent ou détruisent les métaux * ils peuvent ronger la peau et/ou les yeux en cas de contact ou de projection.

Tableau I.2 (Suite)





Symbole	Signification
 SGH06	<p>Ces produits rentrent dans une ou plusieurs de ces catégories :</p> <ul style="list-style-type: none"> * produits cancérogènes : ils peuvent provoquer le cancer ; * produits mutagènes : ils peuvent modifier l'ADN des cellules et peuvent alors entraîner des dommages sur la personne exposée ou sur sa descendance (enfants, petits-enfants...) ; * produits toxiques pour la reproduction: ils peuvent avoir des effets néfastes sur la fonction sexuelle, diminuer la fertilité ou provoquer la mort du fœtus ou des malformations chez l'enfant à naître ; * produits qui peuvent modifier le fonctionnement de certains organes comme le foie, le système nerveux... Selon les produits, ces effets toxiques apparaissent si l'on a été exposé une seule fois ou bien à plusieurs reprises ; * produits qui peuvent entraîner de graves effets sur les poumons et qui peuvent être mortels s'ils pénètrent dans les voies respiratoires (après être passés par la bouche ou le nez ou bien lorsqu'on les vomit) ; * produits qui peuvent provoquer des allergies respiratoires (asthme, par exemple).
 SGH07	<p>Ces produits empoisonnent rapidement, même à faible dose. Ils peuvent provoquer des effets très variés sur l'organisme : nausées, vomissements, maux de tête, perte de connaissance ou d'autres troubles plus importants entraînant la mort.</p>
 SGH08	<p>Ces produits chimiques ont un ou plusieurs des effets suivants :</p> <ul style="list-style-type: none"> ils empoisonnent à forte dose ; ils sont irritants pour les yeux, la gorge, le nez ou la peau ; ils peuvent provoquer des allergies cutanées (eczémas) ; ils peuvent provoquer une somnolence ou des vertiges.

Tableau I.2 (Suite et fin)	
Symbole	Signification
 <p>SGH09</p>	<p>Ces produits provoquent des effets néfastes sur les organismes du milieu aquatique (poissons, crustacés, algues, autres plantes aquatiques...).</p>

Un exemple d'étiquette représenté par la figure I.2, donne plus de détails et d'informations sur le pictogramme présenté.

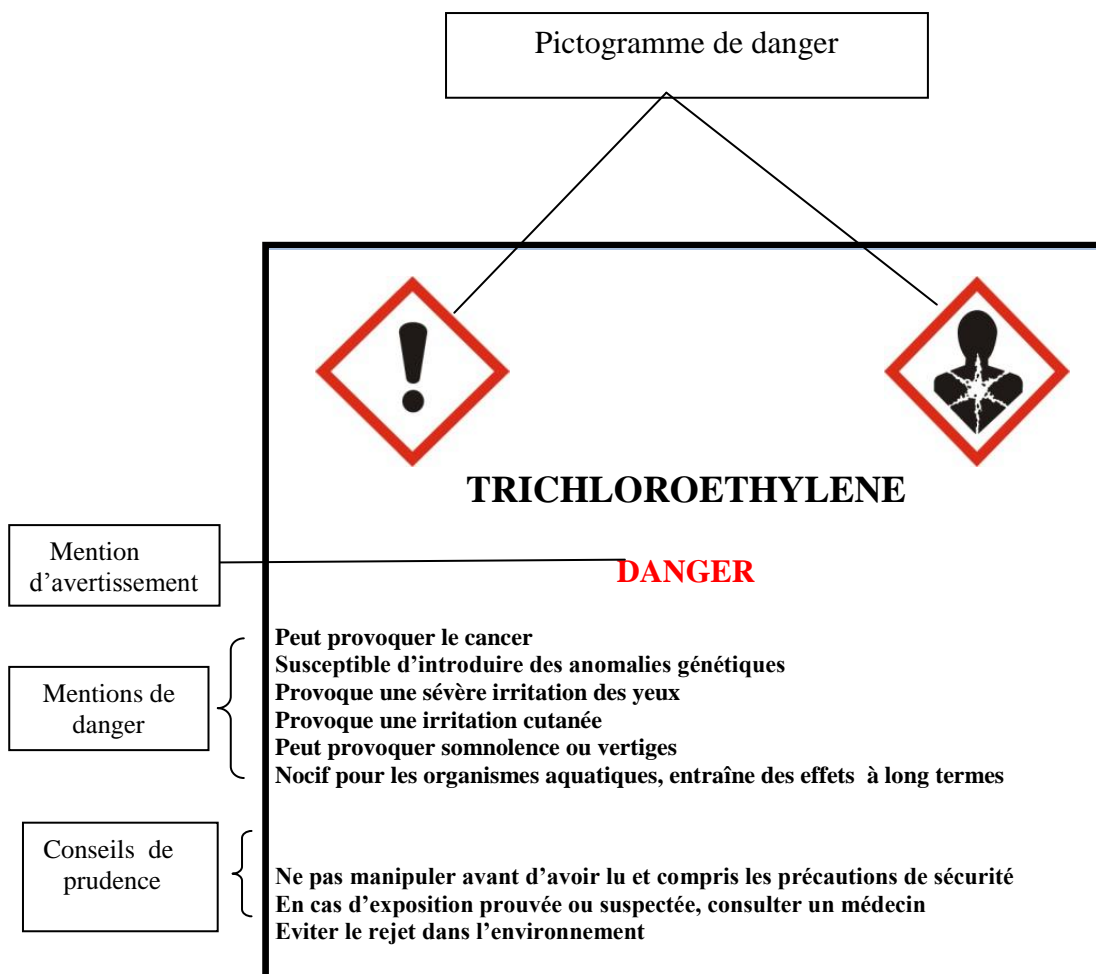


Figure I.2: Signification d'une étiquette.

I.5.2.4. Résumé et comparaison des méthodes d'évaluation

Comme illustré par le tableau I.3 (page 23), la classification d'un liquide inflammable selon l'arrêté du 20 avril 1994 modifié, est basée sur les résultats d'essai alors que celle du règlement CLP est basée sur les résultats d'essai ou sur l'application d'une méthode de calcul de type QSAR/QSPR (Quantitative Structure Activity /Property Relationship) permettant de prédire de manière qualitative les propriétés physico-chimiques, biologiques et environnementales à partir de la connaissance de leur structure chimique.

Les symboles et indications de danger, pictogrammes, phrases de risques, catégories de danger, mentions d'avertissement et mentions de danger sont aussi illustrés. Par exemple, pour une substance dont la phrase de risque R10 avec $23^{\circ}\text{C} \leq \text{PE} \leq 60^{\circ}\text{C}$ se retrouve en catégorie 3 du règlement CLP (voir tableau I.3).

Les éléments d'étiquetage des trois catégories de liquides inflammables sont menés par des phrases de risque et des conseils de sécurité dont les significations sont comme suit [I.12] :

R10 : Inflammable

R11 : Facilement inflammable

R12 : Extrêmement inflammable

H224 : Liquide et vapeurs extrêmement inflammables

H225 : Liquide et vapeurs très inflammables

H226 : Liquide et vapeurs inflammables

P210 : Tenir à l'écart de la chaleur/ des étincelles/des flammes nues/des surfaces chaudes/Ne pas fumer.

P233 : Maintenir le récipient fermé de manière étanche

P240 : Mise à la terre/liaison équipotentielle du récipient et du matériel de réception

P241 : Utiliser du matériel électrique/de ventilation/d'éclairage/.../ antidéflagrant

P242 : Ne pas utiliser d'outils produisant des étincelles

P243 : Prendre des mesures de précaution contre les décharges électrostatiques

P280 : Porter des gants de protection/des vêtements de protection/un équipement de protection des yeux/du visage

Conseil de prudence et d'intervention :

P303 + P361 + P353 : En cas de contact avec la peau (ou les cheveux) : enlever immédiatement les vêtements contaminés. Rincer à l'eau/se doucher.

P370 + P378 : En cas d'incendie: utiliser ... pour l'extinction.

Conseil de prudence et de stockage






P403 + P235 : Stocker dans un endroit bien ventilé. Tenir au frais.

Conseil de prudence et d'élimination

P501 : Éliminer le contenu/réceptacle dans ...

Des indications sur les dimensions minimales des étiquettes et des pictogrammes sont données par le règlement CLP par rapport à la contenance de l'emballage.

Tableau I.3: Résumé et comparaison des méthodes d'évaluation [I.12].

	Arrêté du 20 avril 1994 modifié			Règlement CLP		
Symbole et indication de danger, pictogramme, phrases de risque, catégorie de danger, mentions d'avertissement et mentions de danger	 F+ - Extrêmement inflammable	 F - Extrêmement inflammable	R10	 Catégorie 1 : Danger H224	 Catégorie 1 : Danger H 225	 Catégorie 1 : Attention H 226
	R12	R11				
Critères et méthodes	Règlement (CE) n°440/2008 A.2 et A.9 PE < 0°C Teb ≤ 35°C	Règlement (CE) n°440/2008 A.2 et A.9 PE ≤ 21°C	Règlement (CE) n°440/2008 A.9 21°C ≤ PE ≤ 55°C	Règlement CLP (idem TMD) PE < 23°C Teb ≤ 35°C	Règlement CLP (idem TMD) PE < 23°C Teb > 35°C	Règlement CLP (idem TMD) 23°C ≤ PE ≤ 60°C
	Classification sur la base des résultats d'essais			Classification sur la base des résultats d'essais ou de l'application d'une méthode de calcul		
	Méthodes d'essai identiques					

I.6. Principales catégories de solvants

Pour les besoins de l'hygiène du travail et pour la substitution des solvants en particulier, il est utile de classer les solvants organiques en quatre grandes familles : les hydrocarbures, les solvants halogénés, les hydrocarbures oxygénés et les autres solvants.

I.6.1. Hydrocarbures

Les hydrocarbures sont constitués uniquement de carbone et d'hydrogène. La plupart des hydrocarbures sur le marché sont issus de la pétrochimie. Cette famille comprend les Hydrocarbures aliphatiques, Hydrocarbures cycliques, Hydrocarbures aromatiques, et Mélanges complexes.

Quelques hydrocarbures appelés terpènes sont extraits de végétaux, par exemple le d-limonène provenant des pelures d'agrumes.

I.6.2. Solvants halogénés

Les solvants halogénés sont des hydrocarbures où l'on a remplacé un ou plusieurs atomes d'hydrogène par des atomes d'halogènes (Brome, Chlore, Fluor, Iode). Cette structure leur confère des propriétés sécuritaires intéressantes telles que l'inflammabilité et même l'incombustibilité, en plus d'un pouvoir de dissociation incomparable, d'où leur usage répandu en milieu de travail.

I.6.3. Solvants oxygénés

Les solvants oxygénés contiennent des atomes d'oxygènes dans leur structure moléculaire en plus d'atomes de carbone et d'hydrogène. Cette caractéristique les rend amphiphiles. Cette famille comporte les alcools, les cétones, les esters, les éthers et les éthers de glycol.

I.6.4. D'autres solvants

Il existe de nombreux autres solvants qui n'entrent pas dans les catégories listées précédemment. Plusieurs sont utilisés comme solvants réactionnels en synthèse chimique. Ils sont aussi appelés les solvants particuliers tels que les hydrocarbures nitrés, d'autres composés azotés, les dérivés soufrés et les hydrocarbures complexes. Certains solvants introduits récemment ont fait l'apparition sur le marché, mais leur utilisation est limitée notamment par leur coût élevé [I.13-Ed 4229].

I.7. La toxicité et les maladies professionnelles des solvants organiques

Aucun solvant n'est inoffensif et certains solvants très toxiques doivent être évités. Beaucoup de solvants organiques ont été reconnus comme susceptibles de provoquer des maladies professionnelles lors de l'exposition des travailleurs aux solvants, dans le cadre de leur activité professionnelle.

Le pouvoir toxique des solvants se manifeste en général par une atteinte neurologique centrale et parfois périphérique, des atteintes muco-cutanées, des atteintes de la reproduction et des glandes endocrines, des troubles digestifs, hépatiques, rénaux, etc...

Certains peuvent provoquer des transformations du patrimoine génétique (effets mutagène et cancérogène) ainsi que des atteintes du système reproducteur (perturbation de la reproduction et atteinte de la descendance). Les femmes enceintes sont particulièrement vulnérables car la plupart de ces solvants franchissent la barrière placentaire. Aujourd'hui, ces dernières molécules, dites CMR (Cancérogènes, Mutagènes et Reprotoxiques) sont unanimement reconnues comme étant responsables de problèmes graves de santé. Elles doivent être éliminées de façon prioritaire pour être remplacées par des solvants de toxicité plus faible [I.14].

Tableau I.4: Toxicité de quelques solvants organiques appartenant à différentes familles [I.14].

Famille	Solvant pris pour exemple	Toxicité
Hydrocarbures saturés	Hexane (Inflammable)	Très toxique, puissantes propriétés neurotoxiques périphériques (polynévrite) et certainement centrales (maladie de Parkinson)
Hydrocarbures aromatiques	Benzène (Inflammable)	Extrêmement toxique : cancérogène pour l'Homme. Neurotoxique central Effets cumulatifs

Famille	Solvant pris pour exemple	Toxicité
Hydrocarbures halogénés	1,2 -Dichloroéthane (Inflammable)	Très toxique, reconnu cancérogène chez les animaux (rongeurs), hépatotoxique et dangereux pour les reins, le système nerveux et le cœur. Irritant de la peau, des yeux et du système respiratoire. Groupe 2B : cancérogène possible chez l'Homme, doit être éliminé.
Alcools	Méthanol (Inflammable)	Très neurotoxique pour le nerf optique et la rétine (cécité). Provoque à forte dose de l'acidose (confusion mentale jusqu'au coma).
Cétones	Propanone (Très Inflammable)	Neurotoxique modéré (narcotique). Irritant des yeux, de la peau et du tractus respiratoire.
Esters	Acétate de méthyle (Facilement inflammable)	Libère par hydrolyse du méthanol et de l'acide acétique. Irritant des yeux, des muqueuses, de la peau. Atteinte à long terme du nerf optique.

I.8. Risque pour l'environnement

Les solvants pétroliers sont tous des composés organiques volatils. Leur émission dans l'atmosphère contribue à la production d'ozone dans la troposphère par réaction photochimique, augmentant ainsi les risques pour les personnes asthmatiques ou souffrant d'insuffisance respiratoire.

En cas de rejet dans un milieu aquatique les solvants pétroliers surnageront à la surface. Il peut être envisagé de stopper leur progression par des barrages flottants et, éventuellement, de récupérer cette pollution au moyen d'absorbants par exemple.

La biodégradation est faible, variable selon leur nature: les solvants pétroliers à forte teneur en hydrocarbures aromatiques sont plus toxiques pour les organismes aquatiques.

Le risque d'incendie et d'explosion est l'un des risques majeurs lors de l'utilisation des solvants pétroliers dû à leur inflammabilité. Cependant l'utilisation à chaud de ceux qui sont non inflammables retrouvent les mêmes caractéristiques d'inflammabilité que les autres solvants inflammables car les vapeurs émises par ces substances peuvent former avec l'air des mélanges explosifs [I.13-Ed4224].

I.9. Quelques accidents

- ❖ Le « Soummam 937 » est un navire-école des Forces navales algériennes, acquis en 2006, et dont il est la plus grosse unité en dotation. Il est conçu spécifiquement pour l'entraînement. Le 10 mai 2017 [I.15], un incendie suivi d'une explosion à l'intérieur du navire au niveau du Dock flottant de Béjaia. L'explosion est survenue au moment de peindre, causant deux morts et quatre blessés des employés de l'Entreprise Nationale d'Entretien et de Réparation Navale (ERENAV).
- ❖ La catastrophe de GL1K de SKIKDA, le 21/01/2004 [I.16], qui s'est traduite par la destruction presque totale du complexe et la mort de 27 agents ainsi qu'un nombre important de blessés, de traumatisés avec des répercussions jusqu'à la ville de SKIKDA à environ 3 km.
- ❖ Explosion d'un bac d'hydrocarbure, s'est produite le 20 Février 2001 en France [I.17]. Le bac concerné par l'accident avait une capacité de 5090 m³. Ce jour là, le bac était vide mais contenait habituellement de l'essence de type supercarburant. Les travaux, réalisés par des employés d'une entreprise extérieure, consistaient en un raclage du sol pour en retirer des dépôts résiduels. Les deux intervenants ont été brûlés dont un grièvement.

L'hypothèse d'une étincelle provoquée par un équipement porté par l'un des intervenants conduisant à l'explosion de l'atmosphère explosive.

RÉFÉRENCES BIBLIOGRAPHIQUES

- [I.1] INRS (Institut national de la recherche scientifique). Ventilation des ateliers d'encollage de petits objets (chaussures). Guide pratique de ventilation n° 5. ED 672, 1987 : 28 .
- [I.2] Durrans, T. H., (1938), Solvents Ed4., P238. Van Nostrand (New-York).
- [I.3] Guinot, H., (1948), solvants et plastifiants ouvrage de la collection « Matériaux et synthèse ».
- [I.4] Cohr, K.H., (1985), "Definition and Practical Limitation of the Concept of Organic Solvents ". In: Chronic Effects of Organic Solvents on the Central Nervous System and Diagnostics Criteria. World Health Organization - Regional Office for Europe, Copenhagen, pp. 43-55.
- [I.5] Environnement Canada, (1992). PCE-12-89 : Peintures : Peintures à base de solvant. Environnement Canada; Programme Choix environnemental, Ottawa.
- [I.6] De Lanty, P., (2005), " Paramètres de solubilité ", OCL (Oilseeds and fats, Crops and Lipids). 12 (4). pp. 299-301
- [I.7] INERIS (Institut National de l'Environnement Industriel et des RISques), DRA-09-103185-12091D, 22 décembre 2009.
- [I.8] CUSSTR (Commission Universitaire de Sécurité et Santé au Travail Romande), Danger incendie, Version 1, 2005. p. 3
- [I.9] Florian, M., Benoit, S., (2013), " Point des connaissances ", INRS, ED 5001, 2013 : pp.1-2
- [I.10] Thomson, N, J., (1929), Auto-Ignition Température of Flammable Liquids,(1929), Ind. Eng.Chem.21(2),pp 133-139.
- [I.11] Centre d'expertise en analyse environnementale du Québec. *Détermination de la température du point d'éclair selon la technique Pensky-Martens (vase clos)*. MA. 108 – P.E. 1.1, Ministère du Développement durable, de l'Environnement, de la Faune et des Parcs du Québec, 2012, p 10.
- [I.12] Règlement (CE) n° 1272/2008 du Parlement européen et du Conseil du 16 décembre 2008 relatif à la classification, à l'étiquetage et à l'emballage des substances et des mélanges, modifiant et abrogeant les directives 67/548/CEE et 1999/45/CE et modifiant le règlement (CE) n° 1907/2006, Annexe I : Prescriptions relatives à la classification et à l'étiquetage des substances et mélanges dangereux, Partie 2 : Dangers physiques, Paragraphe 2.6 : Liquides inflammables.
- [I.13] Boust, C., (2011), Fiche solvants " les solvants particuliers", INRS, ED 4229/Ed4224.

[I.14] Palmade-le Dantec, N., Picot, A., " La prévention du risque : le remplacement des solvants les plus toxiques par des solvants moins toxiques ". Actes du colloque " Conservation-Restauration et Sécurité des personnes ", 3-5 février 2010, Draguignan; ISBN 978-2-9531978-1-5.

[I.15] Outenzabt, M., Article publié dans le Matin d'Algérie, le 12 mai 2017.

[I.16] Guerzi, C., mémoire de Magister, Scénarii d'Incendie –Explosion au niveau du complexe pétrochimique de SKIKDA, Etude de cas: Unité Ethylène, Université d'Oran, 2011

[I.17] Séminaire retour d'expérience – IMPEL (European Union Network for the Implementation and Enforcement of Environmental Law) / Inspecteurs des installations classées– Bordeaux, N°ARIA 19979, juin 2002.



Chapitre II



II: Étude théorique

II.1. La modélisation moléculaire

La modélisation moléculaire peut être considérée comme un ensemble de techniques informatiques basées sur des méthodes de chimie théorique et les données expérimentales qui peuvent être utilisées pour la description et la prédiction des propriétés des systèmes moléculaires complexes dans différents environnements.

Cette approche procède de l'hypothèse d'une correspondance univoque entre n'importe quelle propriété physique, affinité chimique, ou activité biologique d'un composé chimique et sa structure moléculaire.

La stabilité de la structure tri-dimensionnelle d'une molécule est déterminée par les interactions intramoléculaires et les interactions avec le milieu extérieur (solvant). La recherche des conformations stables d'une molécule consiste à déterminer les minima de l'énergie globale d'interaction. Cette énergie peut être calculée par des méthodes quantiques *ab initio* ou semi-empiriques généralement longues et onéreuses. Pour faciliter les calculs, on considère habituellement que le terme variable de cette énergie dépend de la construction de la molécule et de l'arrangement de ses atomes : c'est le principe des méthodes empiriques (mécanique moléculaire, dynamique moléculaire). Dans la plupart de ces méthodes, il n'est pas tenu compte des interactions avec le solvant, mais uniquement des interactions entre les atomes constitutifs de la molécule. La recherche d'une conformation consiste alors à faire une minimisation de l'énergie intramoléculaire. Cette énergie potentielle est fractionnée en un certain nombre de termes additifs indépendants. Chacun de ces termes est représenté par une fonction analytique simple justifiée par des calculs quantiques et incluant des paramètres empiriques.

II.2. Optimisation des molécules

II.2.1. La Méthode de HARTREE-FOCK-ROOTHAAN (Méthode de HFR)

II.2.1.1. Energie d'un micro système représenté par un déterminant de Slater

Les calculs quanto-mécaniques courants sont basés sur le modèle de l'électron indépendant où l'on suppose les orbitales soit vides soit garnies de deux électrons au plus.

Dans le cadre de ce modèle, la fonction d'onde polyélectronique peut s'écrire sous la forme d'un produit anti-symétrisé de spin-orbitales :

$$\psi(1,2, \dots, n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \psi_1(1) & \bar{\psi}_1(1) & \dots & \dots & \dots & \bar{\psi}_n(1) \\ \psi_1(2) & \bar{\psi}_1(2) & \dots & \dots & \dots & \bar{\psi}_n(2) \\ \vdots & \vdots & & & & \vdots \\ \vdots & \vdots & & & & \vdots \\ \psi_1(n) & \bar{\psi}_1(n) & \dots & \dots & \dots & \bar{\psi}_n(n) \end{vmatrix} \quad (\text{II.1})$$

Les spin-orbitales sont obtenues en multipliant chaque orbitale par l'une des deux fonctions de spin possibles :

$$\psi_m(n) = \varphi_m(n)\alpha(n) \quad (\text{II.2})$$

$$\bar{\psi}_m(n) = \varphi_m(n)\beta(n)$$

Nous considérerons le cas des systèmes à couches complètes (gaz inertes, molécules courantes dans l'état fondamental...) pour lesquels $n=2m$.

La fonction déterminantale $\psi(1, 2, 3, \dots, n)$ est appelée *déterminant de Slater*.

L'hamiltonien du système est l'hamiltonien résultant, à l'approximation de Born-Oppenheimer.

$$H(1, 2, \dots, n) = \sum_{i=1}^n h_{(i)}^c + \sum_{i<j} \frac{e^2}{r_{ij}} \quad (\text{II.3})$$

$h_{(i)}^c$: est l'hamiltonien monoélectronique de cœur ; le symbole $\sum_{i<j}$ désigne une sommation sur couples ordonnés.

Comme ψ est normé à l'unité (constante de normalisation $1/\sqrt{n!}$), l'énergie du système est donnée par :

$$E = \langle \psi | H | \psi \rangle \quad (\text{II.4})$$

Lorsqu'on développe cette intégrale on arrive [II.1] au résultat :

$$E = \sum_{i=1}^m 2h_{ii}^c + \sum_{i=1}^m \sum_{j=1}^m (2J_{ij} - K_{ij}) \quad (\text{II.5})$$

L'écriture $\sum_{i=1}^m$, signifie que l'on somme sur toutes les orbitales occupées.

$$h_{ii}^c = \langle \psi_i(\mu) | h_{(\mu)}^c | \psi_i(\mu) \rangle \quad (\text{II.6})$$

est l'intégrale monoélectronique moléculaire de cœur, intégrale triple qui porte sur les coordonnées d'un seul électron : le $\mu^{\text{ème}}$ dans ce cas.

$$J_{ij} = \iint \psi_i^*(\mu)\psi_i(\mu) \frac{e^2}{r_{\mu\nu}} \psi_j^*(\nu)\psi_j(\nu) d\tau_\mu d\tau_\nu \quad (\text{II.7})$$

est l'intégrale monoélectronique moléculaire coulombienne, parce qu'elle représente une somme de termes d'interactions coulombiennes, intégrale sextuple qui porte sur les coordonnées de deux électrons.

$$K_{ij} = \iint \psi_i^*(\mu)\psi_i^*(\nu) \frac{e^2}{r_{\mu\nu}} \psi_j(\mu)\psi_j(\nu) d\tau_\mu d\tau_\nu \quad (\text{II.8})$$

est l'intégrale biélectronique moléculaire d'échange ; elle représente également une somme de répulsions entre charges élémentaires, l'électron occupant deux orbitales moléculaires ψ_i et ψ_j .

$r_{\mu\nu}$ représente la distance entre les deux électrons μ et ν .

Remarques :

1)- Dans l'expression de l'énergie E , nous trouvons deux termes :

*- E^c , qui est l'énergie de l'ensemble des électrons évoluant dans le champ des noyaux sans interactions les uns avec les autres.

*- E^{RE} , qui est l'énergie de répulsion électronique.

$$E = E^c + E^{RE} \quad (\text{II.9})$$

Evidemment si l'on suppose qu'il n'existe pas d'interactions entre électrons, le second terme disparaît complètement.

2)- Si on a à traiter une molécule, il faut ajouter un terme supplémentaire de répulsion nucléaire.

$$E_T = E + \sum_{N < L} \frac{Z_K Z_L e^2}{R_{KL}} \quad (\text{II.10})$$

Z_K et Z_L sont les charges des noyaux K et L et R_{KL} la distance entre ces noyaux.

La relation (II.5) est équivalente à :

$$E = \sum_{i=1}^m \{h_{ii}^c + (h_{ii}^c + \sum_{j=1}^m (2J_{ij} + K_{ij}))\} \quad (\text{II.11})$$

Le terme :

$$e_i = h_{ii}^c + \sum_{j=1}^m (2J_{ij} + K_{ij}) \quad (\text{II.12})$$

correspond à ce qu'on appelle l'énergie des orbitales moléculaires.

E se réduit donc à :

$$E = \sum_{i=1}^m (h_{ii}^c + e_i) \quad (\text{II.13})$$

Remarque : Dans les méthodes approchées, comme la méthode de Slater par exemple, on prend :

$$E = \sum_{i=1}^m 2 e_i \quad (\text{II.14})$$

Dans la méthode de Hatree-Fock-Roothaan ceci n'est plus vrai : l'énergie des micro-systèmes n'étant pas égale à la somme des énergies des orbitales moléculaires.

Pour qu'il en soit ainsi, il faudrait que $h_{ii}^c = e_i$ ce qui n'est pas vrai.

Les orbitales moléculaires ne sont pas connues. Le déterminant de Slater n'est connu que par rapport à un jeu de $\{\psi_i\}$ dont on ne sait rien, à part qu'elles sont orthogonales.

Le problème est de déterminer le jeu d'orbitales qui permet de construire le système de Slater.

II.2.1.2. Détermination des Orbitales ou équations de Hartree-Fock

On construit le système de Slater à partir d'un jeu de $\{\psi_i\}$.

Quelles propriétés doivent posséder les ψ_i pour être acceptables au sens de la mécanique ondulatoire, et qu'elles puissent s'adapter au système particulier envisagé ?

Il faut que le déterminant de Slater soit une solution approchée de l'équation de Schrödinger totale :

$$H(1, 2, \dots, n)\psi(1, 2, \dots, n) = E\psi(1, 2, \dots, n) \quad (\text{II.15})$$

La propriété la plus fondamentale des solutions de l'équation de Schrödinger est leur stabilité : c'est-à-dire que si on fait subir à la fonction d'onde déterminantale une perturbation du premier ordre, il s'ensuit une perturbation du premier ordre de l'énergie nulle. Il faut donc réaliser absolument cette condition.

Comme la variation du déterminant de Slater s'exprime par la variation du jeu des $\{\psi_i\}$, il faudrait avoir, pour une variation première du jeu d'orbitales choisies, une variation première de l'énergie totale nulle, et pour cela il faut que les ψ_i soient solutions des équations de Hartree-Fock [II.2-II.4]:

$$\{\delta\psi_i\} \rightarrow \delta E^1 = 0 \quad (\text{II.16})$$

Ces deux conditions contiennent les équations de Hartree-Fock :

$$F_{(\mu)}\psi_i(\mu) = e_i\psi_i(\mu) \quad (\text{II.17})$$

L'équation de Hartree-Fock est une équation intégra-différentielle qui, contrairement à une équation de Schrödinger mono-électronique, fait intervenir un opérateur F qui dépend des fonctions inconnues ψ_i .

Opérateur de Hartree-Fock :

$$F_{(\mu)} = [h_{(\mu)}^c + \sum_{i=1}^m 2J_i(\mu) - K_i(\mu)] \quad (\text{II.18})$$

J_i et K_i sont, respectivement, les opérateurs coulombien et d'échange relatifs à chaque orbitale doublement occupée ψ_i .

II.2.1.3. Equations de Roothaan et Hall

Découlent de la méthode de Hartree-Fock lorsqu'on introduit la condition CLOA (Combinaison Linéaire d' Orbitales Atomiques).

Chaque orbitale moléculaire ψ_i se présentera sous la forme :

$$\psi_i(\mu) = \sum_{p=1}^N C_{pi} \varphi_p(\mu) \quad (\text{II.19})$$

L'ensemble des orbitales atomiques $\{\varphi_p\}$ étant supposé connu, la détermination des ψ_i se ramène à la détermination des C_{pi} .

Les équations de Hartree-Fock prennent, en tenant compte de (II.19), une expression vectorielle assez simple :

$$\sum_{p=1}^N C_{pi} [F_{pq} - e_i S_{pq}] = 0 \quad , \quad q \in [1, N] \quad (\text{II.20})$$

Les coefficients :

$$S_{pq} = \int \varphi_p^* \varphi_q d\tau \quad (\text{II.21})$$

$$F_{pq} = \int \varphi_p^* (F\varphi_q) d\tau$$

sont les intégrales de recouvrement sur la base des fonctions φ_p et les éléments matriciels de l'opérateur de Hartree-Fock F , et les valeurs propres sont les énergies orbitales ϵ_i .

L'équation (II.20) est un système linéaire homogène (N équations à N inconnues) qu'on peut écrire sous la forme matricielle :

$$[\mathbf{F} - \epsilon_i \mathbf{S}] \mathbf{C}_i = \mathbf{0} \quad (\text{II.22})$$

Où \mathbf{F} est la matrice $[F_{pq}]$; \mathbf{S} est la matrice $[S_{pq}]$; \mathbf{C}_i est la matrice $[C_{pi}]$.

$$F_{pq} = h_{pq}^c + \sum_{l=1}^N \sum_{m=1}^N p_{lm} [\langle pq | lm \rangle - \frac{1}{2} \langle pm | lq \rangle] \quad (\text{II.23})$$

-* h_{pq}^c = intégrale monoélectronique sur les orbitales atomiques de base.

$$h_{pq}^c = \left\langle \varphi_p(\mu) \left| h_{(\mu)}^c \right| \varphi_q(\mu) \right\rangle \quad (\text{II.24})$$

$$-* \langle pq | lm \rangle = \iint \varphi_p(\mu) \varphi_q(\mu) \frac{e^2}{r_{\mu\nu}} \varphi_l(\nu) \varphi_m(\nu) d\tau_\mu d\tau_\nu \quad (\text{II.25})$$

$$-* \langle pm | lq \rangle = \int \varphi_p(\mu) \varphi_m(\mu) \frac{e^2}{r_{\mu\nu}} \varphi_l(\nu) \varphi_q(\nu) d\tau_\mu d\tau_\nu \quad (\text{II.26})$$

$$-* p_{lm} = \sum_{i=1}^N 2 C_{li} C_{mi} = \text{éléments de la matrice densité} \quad (\text{II.27})$$

$$-* \mathbf{P} = [p_{lm}] = \text{matrice densité} \quad (\text{II.28})$$

II.2.1.4. Quelques remarques sur les processus de résolution des équations de Hartree-Fock-Roothaan.

L'équation de Hartree-Fock-Roothaan sous forme matricielle est :

$$[\mathbf{F} - \epsilon_i \mathbf{S}] \mathbf{C}_i = \mathbf{0} \quad (\text{II.29})$$

Löwdin [II.5] a proposé un procédé qui permet de se ramener dans tous les cas au calcul des valeurs propres et vecteurs propres d'une matrice moyennant une transformation de la base des orbitales atomiques (**orthogonalisation de Löwdin**).

Multiplions à gauche les deux membres de (II.22) par la matrice $\mathbf{S}^{-1/2}$, qui n'est jamais singulière puisque \mathbf{S} ne l'est pas ; il vient successivement:

$$\begin{aligned} \mathbf{S}^{-1/2} \mathbf{F} \mathbf{C}_i &= e_i \mathbf{S}^{-1/2} \mathbf{S} \mathbf{C}_i \\ \left[\mathbf{S}^{-1/2} \mathbf{F} \mathbf{I} \mathbf{S}^{-1/2} \right] \mathbf{S}^{-1/2} \mathbf{C}_i &= e_i \mathbf{S}^{-1/2} \mathbf{C}_i \end{aligned}$$

Soit en posant :

$$\left[\mathbf{S}^{-1/2} \mathbf{F} \mathbf{I} \mathbf{S}^{-1/2} \right] = \bar{\mathbf{F}} \quad \text{et} \quad \mathbf{S}^{-1/2} \mathbf{C}_i = \bar{\mathbf{C}}_i \quad (\text{II.29})$$

$$\bar{\mathbf{F}} \bar{\mathbf{C}}_i = e_i \bar{\mathbf{C}}_i \quad \text{c'est-à-dire} \quad [\bar{\mathbf{F}} - e_i \mathbf{I}] \bar{\mathbf{C}}_i = \mathbf{0} \quad (\text{II.30})$$

Les équations de Hatree-Fock-Roothaan sont résolues selon un procédé itératif qui se fait sur l'ensemble orthogonalisé.

$$\bar{\mathbf{F}} \bar{\mathbf{C}} = e_i \bar{\mathbf{C}}_i \quad (\text{II.30})$$

On peut toujours initialiser le problème en choisissant a priori une matrice densité, obtenue en négligeant la matrice des interactions électroniques (problème d'ordre zéro). Le nombre d'itérations dépend du problème à résoudre.

II.2.1.5. Détermination des intégrales de la méthode de Hartree-Fock-Roothaan (HFR)

Le très gros problème dans la méthode HFR est la détermination des intégrales.

❖ **Intégrales monoélectroniques atomiques de cœur :**

$$h_{pq}^c = \langle \varphi_p(\mu) | h_{\mu}^c | \varphi_q(\mu) \rangle \quad (\text{II.31})$$

Il existe deux types d'intégrales de ce genre : **monocentres** lorsque φ_p et φ_q appartiennent au même atome R, **bicentres**, lorsque φ_p et φ_q appartiennent à des atomes différents.

Les intégrales monoélectroniques de cœur monocentres comprennent : les intégrales de cœur coulombiennes (même orbitale atomique des deux côtés) et les intégrales de cœur d'échange (les deux orbitales atomiques sont différentes)

$$h_{pq}^c = \underbrace{-\frac{\hbar^2}{2m} \int \varphi_p(\mu) \Delta(\mu) \varphi_q(\mu) d\tau_\mu}_{\text{Intégrales cinétiques}} - \underbrace{\sum_k Z_K \int \varphi_p(\mu) \frac{e^2}{r_{k\mu}} \varphi_q(\mu) d\tau_\mu}_{\text{intégrales d'attractions nucléaires}} \quad (\text{II.32})$$

Les intégrales d'attractions nucléaires peuvent être monocentres, bicentres ou tricentres (très compliquées à calculer).

❖ Intégrales bi-électroniques

$$G_{pq} = \sum_l \sum_m p_{lm} \left[\langle pq|lm \rangle - \frac{1}{2} \langle pm|lq \rangle \right] \quad (\text{II.33})$$

$\langle pq|lm \rangle$, $\langle pm|lq \rangle$ et p_{lm} sont respectivement définis par les relations (II.25), (II.26) et (II.27).

On a plusieurs types d'intégrales :

- **monocentres**, lorsque, $\varphi_p, \varphi_q, \varphi_l, \varphi_m$ appartiennent au même atome.
- **bicentres**, lorsque parmi, $\varphi_p, \varphi_q, \varphi_l, \varphi_m$ il y en a qui appartiennent à deux atomes différents.
- **tricentres**, parmi, $\varphi_p, \varphi_q, \varphi_l, \varphi_m$ il y en a qui appartiennent à trois atomes différents.
- **tétracentres**, chaque orbitale appartient à un atome différent.

Le calcul des intégrales biélectroniques prend le plus grand temps, et il n'est pas possible, en prenant des orbitales de Slater (II.34) d'en donner des expressions analytiques.

$$\varphi_{n,l,m}(k, \vec{r}) = N r^{n-1} e^{-kr} y_{l,m}(\theta, \varphi) \quad (\text{II.34})$$

$y_{l,m}(\theta, \varphi)$ étant les harmoniques sphériques.

On décompose alors chaque orbitale de Slater en orbitales gaussiennes dont la partie radiale est de la forme e^{-kr^2} , ce qui permet de ramener un problème d'analyse numérique à un problème d'algèbre.

II.2.2. Les méthodes semi-empiriques

Dans le précédent chapitre, nous avons exposé la théorie des orbitales moléculaires d'un point de vue *ab-initio*, déterminant une fonction d'onde qui nécessite le calcul d'un certain nombre d'intégrales et l'utilisation d'une procédure algébrique auto-cohérente.

Dans le cadre de cette théorie, une approche plus approximative est développée, ce qui permet d'éviter l'évaluation difficile de beaucoup d'intégrales et de sélectionner les valeurs de certaines autres en tenant compte des données expérimentales.

Les approches semi-empiriques, qui traitent des électrons de valence, sont désignées par des sigles dont les lettres correspondent aux approximations admises dans le recouvrement différentiel des orbitales.

II.2.2.1. Définition du semi-empirisme

Une méthode est semi-empirique si elle admet le cadre de Hatree-Fock-Roothan, en y incorporant un certain nombre de simplifications.

On arrive ainsi à réduire considérablement le nombre d'intégrales. En particulier on élimine les intégrales biélectroniques à 3 et 4 centres, qui sont très faibles.

Une fois le cadre HFR simplifié, on évalue empiriquement les intégrales restantes en ajustant la méthode sur des molécules bien connues.

II.2.2.2. Quelques théories semi-empiriques

La première théorie semi-empirique, ou théorie de Pople-Pariser-Parr (PPP), introduite en 1953 par Pariser et Parr [II.6-II.7], et utilisée la même année par Pople [II.8], permet d'étudier les systèmes conjugués sans tenir compte du squelette σ .

La première théorie des orbitales moléculaires semi-empirique tri-dimensionnelle est l'approximation au recouvrement différentiel nul (CNDO pour : Complete Neglect of Differential Overlap), introduite par Pople, Santry et Segal [II.9], pour être appliquée à tous les électrons de valence de molécules quelconques organiques ou minérales.

L'approximation utilisée dans CNDO, et dans de nombreuses approximations subséquentes, pour traiter des interactions électron-électron est connue comme :

- Approximation du champ moyen ;
 - Théorie du champ auto-cohérent (SCF : Self Consistent Field)
- et
- Théorie de Hartree- Fock (HF).

De ces appellations, l'approximation du champ moyen est probablement la plus expressive, mais c'est le terme SCF qui est le plus courant.

Comme le problème du calcul de l'énergie d'interaction électron-électron dans un système poly-électronique ne peut avoir de solution exacte, on doit utiliser des approximations. La théorie SCF traite chaque électron comme s'il interagissait (au cours du temps) avec le champ moyen de tous les autres électrons de la molécule. Ce qui signifie que les électrons restants de la molécule ne réagissent pas avec l'électron considéré dans sa position instantanée. Ainsi, le calcul de l'énergie de chaque électron pris individuellement devient un problème mono-électronique auquel nous avons à ajouter l'effet du champ causé par les électrons restants. Cette approximation néglige le fait que les mouvements des électrons sont corrélés de manière à réduire leurs répulsions mutuelles (c'est-à-dire que chaque électron réagit aux positions instantanées de tous les autres). Ainsi, la théorie SCF rend la tâche computationnelle gérable au prix d'une surestimation de l'énergie de répulsion électron-électron.

Cependant, en 1965, les ressources computationnelles nécessaires pour l'approche SCF complète n'étaient pas encore disponibles. La pratique des théories des orbitales moléculaires nécessitaient donc encore des approximations. Le principal problème réside dans le calcul et le stockage des intégrales tétracentres notées $\langle \mu\nu|\lambda\sigma \rangle$, nécessaires pour le calcul des interactions électron-électron dans le cadre de l'approximation SCF. Les indices μ, ν, λ et σ dénotent quatre centres d'orbitales atomiques de sorte que le nombre de telles orbitales à calculer croît proportionnellement à N^4 , où N est le nombre d'orbitales atomiques. En fait, le nombre de telles intégrales n'est pas exactement égal à la puissance quatrième du nombre de fonctions de base parce que beaucoup d'entre elles sont reliées par symétrie. Ce qui était une tâche très difficile en 1965 ; ainsi Pople, Santry et Segal [II.9] ont introduit l'approximation que seules les intégrales pour lesquelles $\mu = \nu$ et $\lambda = \sigma$ c'est-à-dire : $\langle \mu\mu|\nu\nu \rangle$ seront prises en compte et que, de plus, toutes les orbitales atomiques seront traitées de la même façon (comme si elles étaient des orbitales s), de sorte que l'équation (II.35) s'applique, où μ est centrée sur l'atome A et λ sur l'atome B et ainsi γ_{AB} ne dépend que des identités de A et B, et peut être traité comme paramètre.

$$\langle \mu\mu|\lambda\lambda \rangle = \gamma_{AB} \quad (\text{II.35})$$

Une première approximation, due à Pariser et Parr [II.7] consiste à traiter le terme mono-centre γ_{AA} comme différence entre le potentiel d'ionisation PI_A et l'affinité électronique AE_A de A (Equation(II.36)):

$$\gamma_{AA} = PI_A - AE_A \quad (\text{II.36})$$

Les termes di-centres sont alors donnés par l'équation(II.37):

$$\gamma_{AB} = \frac{\gamma_{AA} + \gamma_{BB}}{2 + r_{AB}(\gamma_{AA} + \gamma_{BB})} \quad (\text{II.37})$$

Ce qui conduit à : $\gamma_{AB} = (\gamma_{AA} + \gamma_{BB})/2$ pour une distance interatomique, r_{AB} , nulle et $\gamma_{AB} \approx 1/r_{AB}$ pour des distances interatomiques plus grandes. Ces expressions (Equations (II.35) et (II.37)) montrent la simplicité de la technique CNDO, qui a été utilisée pour calculer des propriétés électroniques comme les moments dipolaires ou les énergies d'excitation, généralement à partir des géométries expérimentales. Il y a eu beaucoup de modifications des eqs. ((II.36) et (II.37)), mais elles restent d'une simplicité comparable. Pareillement, des expressions simplifiées ont aussi été utilisées pour les intégrales mono-électroniques.

Cependant, la méthode CNDO montra des insuffisances systématiques directement imputées aux simplifications ébauchées précédemment, aussi fut-elle remplacée par la méthode **INDO (Intermediale Neglect of Differential Overlap)**, introduite en 1967 par Pople, Beveridge et Dobosh [II.10]. L'approximation qui conduit à l'équation (II.35) s'étant avérée très sévère, elle fut remplacée par des valeurs individuelles pour les différents types d'interactions entre deux orbitales atomiques. Ces valeurs individuelles, souvent désignées par G_{ss} , G_{sp} , G_{pp} et G^2_{pp} dans la littérature, peuvent être ajustées pour donner un accord avec l'expérience meilleur que celui obtenu avec la méthode CNDO. Cependant, en INDO les termes di-centres sont maintenus du même type que ceux apparaissant dans les équations. (II.36) et (II.37). Cette approximation conduit à des affaiblissements systématiques, comme par exemple dans le traitement des interactions entre doublets isolés.

Pour surmonter ces carences, Pople et collaborateurs revinrent à une approche plus complète que celle qu'ils proposèrent initialement en 1965[II.9]: l'approximation au recouvrement différentiel diatomique nul (NDDO : Neglect of Diatomic Differential Overlap).

Dans la NDDO, toutes les intégrales tétracentres $\langle \mu\nu | \lambda\sigma \rangle$ dans lesquelles μ et ν sont sur le même centre, comme le sont λ et σ (mais pas nécessairement sur le même comme le sont μ et ν) sont prises en compte. De plus, les intégrales pour lesquelles les deux centres atomiques sont différents sont traitées de manière analogue que les intégrales mono-centres en INDO, entraînant, une amélioration de la description des interactions (doublet isolé)-(doublet isolé)

par rapport aux méthodes précédentes. La NDDO forme la base de presque toutes les autres méthodes semi-empiriques qui, à quelques exceptions ont été développées par MJS Dewar et son école.

Les premières techniques semi-empiriques développées par Dewar et son groupe ont été désignées par MINDO/1-3 et ont été basées sur INDO. Beaucoup d'approximations d'intégrales de l'INDO originale ont été remplacées et les méthodes paramétrées pour reproduire un large intervalle de données expérimentales, particulièrement les énergies et les géométries.

Les méthodes MINDO sont maintenant largement obsolètes.

La méthode avantageuse pour la plupart des techniques modernes d'orbitales moléculaires semi-empiriques est la MNDO, qui a été publiée par Dewar et Thiel en 1977 [II.11-II.12]. La MNDO est une méthode NDDO dans laquelle Dewar et Thiel ont introduit un formalisme basé sur les multipôles pour le calcul des intégrales bi-électroniques. Elle a été paramétrée pour reproduire les chaleurs de formation expérimentales, les géométries, les moments dipolaires et les potentiels d'ionisation. Elle s'avéra très supérieure aux méthodes MINDO pour la plupart des grandeurs calculées. Cependant la MNDO présente une faiblesse qui limite sévèrement son utilité ; elle ne reproduit pas la liaison hydrogène. Cette faiblesse a été surmontée de façon pragmatique par Bustein et Isaev [II.13] qui modifièrent simplement le potentiel de répulsion cœur-cœur par addition de fonctions gaussiennes en vue d'obtenir des liaisons hydrogène. Ce « fixe » a été adopté par le groupe Dewar pour leur méthode suivante AM1 [II.13-II.15] qui est par ailleurs identique à la MNDO. AM1, en retour, s'avéra présenter une faiblesse dans le traitement des composés nitrosés et hypervalents. Ces faiblesses ont été abordées par Stewart dans une nouvelle paramétrisation nommée PM3 [II.16-II.17] qui est par ailleurs identiques à AM1. Cependant, MNDO, MNDO/H, AM1 et PM3 sont pour l'essentiel identiques du point de vue quanto-mécanique. Leurs différences se limitent à la « correction » classique des potentiels entre atomes et pour laquelle les paramètres sont traités comme variables dans la procédure de paramétrisation.

II.2.2.3. Limites et avantages des méthodes semi-empiriques [II.18]

La négligence de toutes les intégrales bi-électroniques tri et tétracentres réduit la matrice de Fock d'un ordre formel M^4 à M^2 . Toutefois, le temps requis pour la diagonalisation de la matrice F croît comme le cube de la dimension de la matrice. La diagonalisation d'une

matrice devient importante lorsque la dimension dépasse $\sim 10\,000 \times 10\,000$. De nombreuses itérations sont nécessaires pour la résolution des équations SCF, et habituellement la géométrie est également optimisée, nécessitant de nombreux calculs pour différentes géométries. Ce qui situe la limite actuelle des méthodes semi-empiriques à environ 1000 atomes. Il est à noter que la méthode classique de résolution des équations HF par diagonalisation de la matrice de Fock s'impose rapidement comme l'étape limitante réelle dans les méthodes semi-empiriques. Des développements ultérieurs se sont ainsi focalisés sur la formulation de méthodes alternatives pour l'obtention d'orbitales SCF sans passer par la diagonalisation [II.19-II.20]. De telles méthodes utilisent des ajustements (combinaisons) linéaires avec le nombre d'atomes, ce qui permet d'effectuer des calculs pour des systèmes comprenant plusieurs milliers d'atomes.

La paramétrisation de MNDO/AM1/PM3 est réalisée en ajustant les constantes impliquées dans les différentes méthodes de façon à ce que les résultats des calculs HF ajustent les données expérimentales aussi près que possible. Ce qui est faux dans un sens. On sait que la méthode HF ne peut conduire au résultat correct, même à la limite d'un ensemble de base infini et sans approximations. Les résultats HF ne reproduisent pas la corrélation électronique, mais les données expérimentales impliquent naturellement de tels effets. Ceci peut être considéré comme un avantage, les effets de corrélation électronique sont implicitement pris en compte dans la paramétrisation, et il n'est pas besoin d'exécuter des calculs compliqués pour surmonter les déficiences de la procédure HF. Cependant, il y a réellement problème quand la fonction d'onde HF ne peut décrire le système correctement, même qualitativement, comme par exemple avec les bi-radicaux et les états excités.

Une flexibilité additionnelle peut être introduite dans la fonction d'onde d'essai en ajoutant davantage de déterminants de Slater, par exemple par l'intermédiaire d'une procédure d'interaction de configuration (CI : pour Configuration Interaction). Seulement la corrélation électronique est prise en compte deux fois, une première fois lors de la paramétrisation au niveau HF, et une seconde fois explicitement par le calcul CI.

Remarque : l'interaction de configuration CI résout le problème de la corrélation électronique en considérant plus d'un schéma d'occupation des orbitales moléculaires (OM) et en combinant les micro-états obtenus par permutation des positions électroniques sur toutes les OM disponibles. Dans sa forme la plus simple, un calcul CI consiste en un calcul SCF préliminaire qui fournit les OM qui seront utilisées telles quelles tout au long du reste du

traitement. Des micro-états sont alors construits en déplaçant les électrons des orbitales occupées à celles vacantes selon des schémas pré-établis. La matrice CI est alors calculée, ses éléments diagonaux représentent les énergies des micro-états et les éléments non diagonaux leurs interactions. Cette matrice est diagonalisée en vue d'obtenir les énergies des différents états (fondamentaux et excités) de la molécule comme combinaisons linéaires des micro-états. De nouveau les énergies sont fournies par les valeurs propres et les coefficients de la combinaison linéaire par les vecteurs propres. Cette procédure conduit à la stabilisation de l'état fondamental, et fournit également les énergies et les fonctions d'onde des états excités. Le problème est que si l'on doit considérer chacun des arrangements possibles de tous les électrons dans toutes les OM (CI complète), les calculs deviennent pas trop importants même pour des molécules de taille moyenne avec un ensemble de base pas trop important (parce qu'il y a de trop nombreuses orbitales virtuelles).

Aussi, deux types de restrictions sont habituellement utilisées ; seul un nombre limité d'OM autour de l'intervalle des orbitales frontières (HOMO-LUMO) est inclus dans CI, et seuls certains types de réarrangements (excitations) des électrons sont utilisés.

La forme la plus économique est celle pour laquelle seuls les micro-états dans lesquels un électron est promu de l'état fondamental à une orbitale virtuelle (excitations simples) sont utilisés. Ce qu'on désigne, dans une forme abrégée, par CIS. En ajoutant toutes les excitations doubles (pour les quelles deux électrons sont promus) on est conduit à CISD, et ainsi de suite (Figure. II.1).

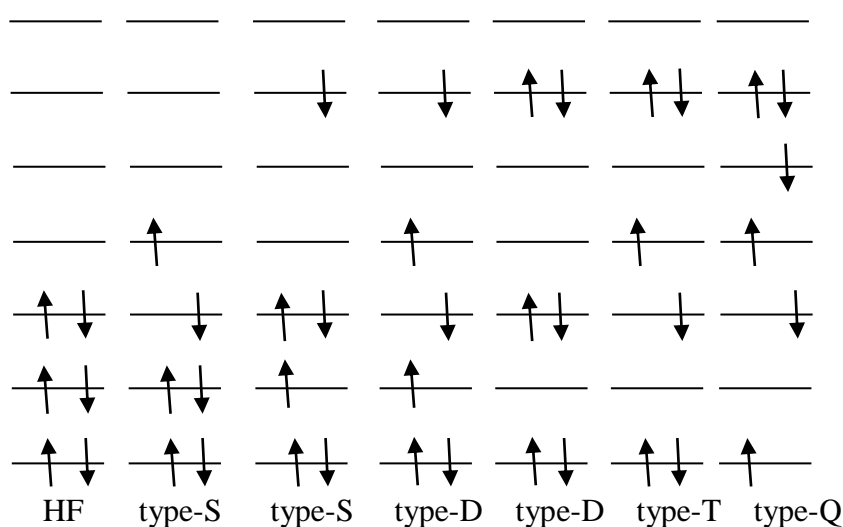


Figure II.1: Déterminants de Slater excités générés à partir d'une référence HF.

Les déterminants sont désignés par simples (S), doubles (D), Triples (T), quadruples (Q) etc...

La fonction d'onde avec interaction de configuration (ψ_{CI}) peut être représentée par l'équation suivante :

$$\psi_{CI} = a_0 \phi_{SCF} + \sum_{\text{Simples (S)}} a_S \phi_S + \sum_{\text{Doubles (D)}} a_D \phi_D + \dots = \sum_{i=0} a_i \phi_i \quad (\text{II.38})$$

La méthode des multiplicateurs indéterminés de Lagrange [II.21] est ensuite appliquée pour minimiser l'énergie :

$$E = (\langle \psi | \hat{H} | \psi \rangle / \langle \psi | \psi \rangle) \quad (\text{II.39})$$

Les méthodes semi-empiriques partagent les avantages/désavantages des méthodes de champ de force (cf : II.3), elles sont davantage performantes avec les systèmes pour lesquels on dispose de données expérimentales en quantités, mais il leur est impossible de faire des prédictions pour des types de composés totalement inconnus. La dépendance des données expérimentales n'est pas aussi sévère que pour la méthode du champ de force, à cause de la forme complexe de la fonctionnelle du modèle. Les méthodes NDDO nécessitent uniquement des paramètres atomiques, et nullement des paramètres di-, tri- et tétra-atomiques comme dans les méthodes de champ de force. Une fois un atome donné paramétré, tous les types de composés possibles contenant cet élément peuvent être traités. Le plus petit nombre de paramètres et la forme plus complexe de la fonctionnelle ont l'inconvénient, par rapport aux méthodes de champ de force, qu'il est très difficile de « réparer » un problème spécifique par reparamétrisation.

Les méthodes semi-empiriques sont de dimension nulle, tout comme les méthodes de champ de force. Il n'y a aucun moyen d'évaluer la fiabilité d'un résultat donné dans les limites de la méthode. Cela est dû à la sélection d'un ensemble de base fixe (minimum). La seule façon de juger les résultats est de comparer la précision d'autres calculs sur des systèmes similaires avec des données expérimentales.

Les méthodes semi-empiriques fournissent une méthode de calcul de la fonction d'onde électronique, qui peut être utilisée pour la prévision d'une variété de propriétés. Il n'y a rien qui entrave le calcul, par exemple, de la polarisabilité d'une molécule, bien qu'il soit connu des calculs *ab-initio* que l'obtention de bons résultats nécessite un grand ensemble de base polarisé incluant des fonctions diffuses. Les méthodes semi-empiriques comme AM1 ou PM3

n'ont qu'une base minimale (absence de polarisation et de fonctions diffuses), la corrélation électronique n'est qu'implicitement incluse par les paramètres et aucune donnée de polarisabilité n'a été utilisée pour dériver ces paramètres. Il est douteux que de tels calculs puissent conduire à des résultats comparables à ceux fournis par l'expérience, et ils nécessitent, pour le moins, un calibrage soigné [II.18]. Encore une fois, il convient de souligner que la capacité d'effectuer un calcul ne garantit pas la fiabilité des résultats obtenus.

II.2.3. Analyse des distributions de charges

Plutôt que de décrire la distribution électronique d'une molécule par des cartes d'isodensité, on préfère caractériser cette distribution, dans le voisinage d'un atome ou d'une liaison, par des nombres simples ou indices. Cette procédure, qui entraîne une perte d'information, est avantageuse dans les études comparatives.

La caractérisation d'une molécule par un tel ensemble d'indices est appelée son **analyse de population**.

Il existe une famille d'analyses de population, parmi lesquelles nous citerons celles de Coulson et Longuet-Higgins [II.22], exprimée en termes de charges (ou « densités de charge ») et d'ordres de liaison, celle de Mulliken [II.23], que nous rappellerons brièvement, et qui fait intervenir les populations atomiques et de recouvrement.

II.2.3.1. Analyse de population de Mulliken

Mulliken introduit le concept important de population de recouvrement, c'est-à-dire de population électronique non localisée sur un atome mais répartie dans la liaison entre deux atomes. Ce concept permet une représentation très nuancée de la liaison chimique.

Dans l'analyse de population électronique qu'il propose, Mulliken définit les grandeurs [II.23] :

$$P_v = \sum_k^{OM.occupées} N_k C_{kv}^* C_{kv} \quad (\text{II.40})$$

où N_k est la population de l'O.M. ψ_k ; P_v est la population électronique localisée dans l'O.A. $\varphi_{\mu\nu}$, que l'on appelle la population nette de l'O.A. φ_ν , dans la molécule.

$$R_{\mu\nu} = 2 \sum_k^{OM.occupées} C_{k\mu}^* C_{kv} S_{\mu\nu} \quad (\text{II.41})$$

$R_{\mu\nu}$ est la population électronique localisée ni dans φ_μ , ni dans φ_ν , mais répartie entre ces deux O.A, que l'on appelle population de recouvrement entre les O.A φ_μ et φ_ν .

En désignant par N le nombre total d'électrons, on a :

$$\sum_{\mu} R_{\mu\nu} = \sum_{\mu} \sum_{\nu} P_{\mu\nu} S_{\mu\nu} = N \quad [\text{Décomposition sur les OA}] \quad (\text{II.42})$$

$$\int \psi^* \psi d\tau = N \quad [\text{Décomposition sur les OM}] \quad (\text{II.43})$$

Posons :

$q_{\mu} = \sum_{\nu} P_{\mu\nu} S_{\mu\nu} =$ Quantité d'électricité qui peut être attribuée à la $\mu^{\text{ème}}$ orbitale atomique de base.

Alors, la quantité d'électricité qui peut être attribuée à l'atome M , dans la molécule, est la somme des $q_{\mu}(M)$ ($\mu \in M$), soit :

$$Q_M = \sum_{\mu(M)} q_{\mu}(M) \quad (\text{II.44})$$

q_{μ} = densité électronique de l'orbitale μ ;

Q_M = densité électronique de l'atome M .

On peut ainsi déterminer la **charge (formelle) de l'atome M , dans la molécule, soit δ_M** :

$$\delta_M = Z_M - Q_M \quad (\text{II.45})$$

Z_M = nombre d'électrons de l'atome isolé ; Q_M = quantité d'électricité qu'il possède dans la molécule.

II.2.3.2. Calcul du moment dipolaire

Le moment dipolaire d'une molécule peut être décomposé, de façon unique, en trois composantes : une composante atomique ou d'hybridation, une composante de recouvrement, et une composante de transfert de charge (qui permet de définir les charges atomiques nettes), chacune étant définie de façon univoque dans le cadre du schéma OM-CLOA.

Dans ce schéma, l'expression en u.a du moment dipolaire d'une molécule, dans la convention des chimistes, est [II.24] :

$$\vec{\mu} = \sum_P \sum_Q \sum_{r \in P} \sum_{s \in Q} P_{rs}^{PQ} \int \varphi_r^* \vec{r} \varphi_s d_s d_r - \vec{\mu}_{nucl} \quad (\text{II.46})$$

avec :

$$P_{rs}^{PQ} = \sum_i n_i C_{ir} C_{is} \quad (\text{II.47})$$

n_i = taux d'occupation de l'OM ψ_i , C_{ir} et C_{is} , coefficients des orbitales φ_r et φ_s appartenant respectivement, aux atomes P et Q , dans l'approximation CLOA des ψ_i . Le vecteur position d'un électron en général et le vecteur position d'un atome P (mesurés en u. a par rapport à la même origine arbitraire) seront notés \vec{r} et \vec{r}_p , alors que np désignera le nombre d'électrons de l'atome P engagés dans la formation de la molécule.

On peut alors faire les substitutions suivantes :

$$\vec{r} = \vec{r}_p + \vec{\xi}, \text{ dans les termes tels que } P = Q \quad (\text{II.48})$$

$$\vec{r} = \frac{1}{2}(\vec{r}_p + \vec{r}_q) + \vec{\chi}, \text{ dans les termes tels que } P \neq Q$$

Evidemment $\vec{\xi}$ est le rayon vecteur qui a pour origine la position de l'atome P , $\vec{\chi}$ est le rayon vecteur dont l'origine coïncide avec le milieu du segment PQ . En tenant compte de l'orthogonalité des deux orbitales φ_r et $\varphi_{r'}$, centrées sur le même atome P , en appelant S_{rs}^{PQ} l'intégrale de recouvrement des orbitales centrées sur des atomes P et Q différents, et en posant :

$$\vec{\xi}_{rr'}^P = \int \varphi_r^* \vec{\xi} \varphi_{r'} d\tau ; \vec{\chi}_{rs}^{PQ} = \frac{\int \varphi_r^* \vec{\chi} \varphi_s d\tau}{S_{rs}^{PQ}} \quad (\text{II.49})$$

Le moment dipolaire (51) devient [II.23]:

$$\vec{\mu} = \sum_p \delta_p \vec{r}_p + \vec{\mu}_{\text{hybrid}} + \vec{\mu}_{\text{recouvr}} \quad (\text{II.50})$$

Avec :

$$\vec{\mu}_{\text{hybrid}} = \sum_p \sum_{r,r' \in P} P_{rr'}^{PP} \vec{\xi}_{rr'}^P \quad (\text{II.51})$$

Et :

$$\vec{\mu}_{\text{recouvr}} = \sum_p \sum_{r,r' \in P} \sum_Q \sum_{s \in Q} P_{rs}^{PQ} S_{rs}^{PQ} \vec{\chi}_{rs}^{PQ} \quad (\text{II.52})$$

II.2.3.3. Application

Nous avons réuni dans la figure II.2 quelques applications [II.25] des indices électroniques de la méthode des orbitales moléculaires.

Sur la base des charges atomiques partielles on peut calculer des descripteurs électrostatiques simples qui peuvent servir pour le développement d'équations QSXR [Relations Quantitatives Structures –X ; où X= P (propriété) – A (activité) – R (rétention chromatographique) – T (toxicité)...].

- Les charges partielles minimale (la plus négative) et maximale (la plus positive) dans la molécule (q_{\min} , q_{\max}).
- Les charges partielles minimale et maximale pour les atomes particuliers (C, O etc...).
- Un paramètre de polarité simple (q_{\max} , q_{\min}) ou pondéré par une fonction de la distance r_{\max} entre les atomes portant les charges partielles minimale et maximale.

$$P_f = \frac{q_{\max} - q_{\min}}{F(r_{\max})} \quad (\text{II.53})$$

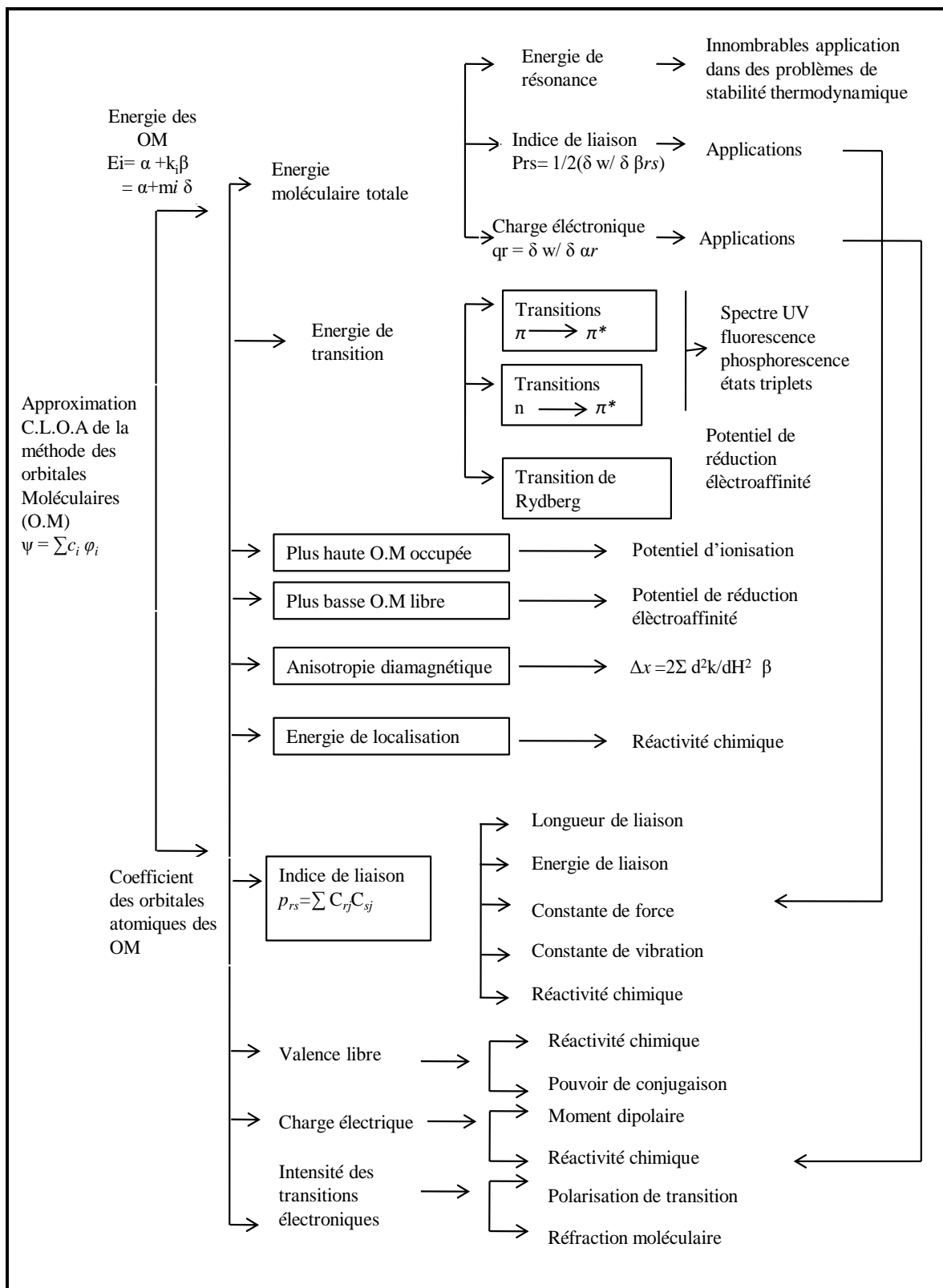


Figure II.2: Les indices électroniques de la méthode des orbitales moléculaires et leurs applications, d'après [II.25].

II.3. La mécanique moléculaire.

Si une molécule est trop grosse pour subir un traitement semi-empirique, il est toujours possible de modéliser son comportement en évitant complètement la mécanique quantique. Les méthodes désignées par mécanique moléculaire, établissent une expression algébrique simple de l'énergie d'un composé, sans avoir à calculer une fonction d'onde ou une densité électronique totale [II.26]. L'expression de l'énergie consiste en des équations classiques simples, comme l'équation de l'oscillateur harmonique, dans le but de décrire l'énergie associée à l'étirement de liaison, de flexion, de rotation, et aux forces intermoléculaires, telles que les interactions de Van Der Waals et de liaison hydrogène. Toutes les constantes apparaissant dans ces équations doivent être obtenues à partir de données expérimentales ou d'un calcul *ab initio*.

Dans une méthode de mécanique moléculaire, la base de données des composés utilisés pour paramétrer la méthode (un ensemble de paramètres et de fonctions est appelé un champ de force) est cruciale pour son succès. La méthode de mécanique moléculaire peut être paramétrée à partir d'une classe spécifique de molécules, telles que des protéines, des molécules organiques, organo-métalliques, etc...

La mécanique moléculaire permet la modélisation de très grosses molécules, comme les protéines et des segments de DNA, la faisant le premier outil de la biochimie computationnelle. Le défaut de cette méthode est qu'il y a beaucoup de propriétés chimiques qui n'y sont pas définies, comme par exemple les états électroniques excités. De plus, pour travailler avec des systèmes très grands et très compliqués, les logiciels doivent être très puissants et faciles dans l'utilisation des interfaces graphiques.

II.3.1. Pas de calculs de champ de force sans définition préalable des types d'atomes.

La géométrie de la molécule traitée (caractérisée par les coordonnées internes ou les coordonnées cartésiennes), le numéro atomique de chaque noyau, et l'état général de charge et de spin, constituent le nombre minimal d'entrées préalable à un calcul par mécanique moléculaire. Les informations concernant les distributions des électrons, en terme de densité électronique ou de fonction d'onde, ou les charges atomiques partielles, sont mieux interprétées sur la base de la géométrie moléculaire. Dans le contexte de la méthodologie du champ de force, l'entrée de la charge totale et du spin d'une molécule n'est pas obligatoire car ces types de calculs ne traitent pas des électrons. Pour représenter l'aspect électrostatique, il

n'est même pas besoin des charges atomiques partielles si l'on utilise, par exemple, des dipôles de liaisons. Au contraire de la mécanique quantique, la mécanique moléculaire nécessite plus d'informations que le numéro atomique seul. En fait, chaque atome doit être décrit de manière plus détaillée.

Le concept de types d'atomes permet une différenciation en termes d'environnement local, d'état d'hybridation, ou de conditions spécifiques telles que la tension dans les systèmes comportant un petit anneau. Allinger et ses co-auteurs, qui ont développé les champs de force MM2, MM3, et MM4 pour les « petites molécules » (cf : III.3.3) ont défini dans la paramétrisation de MM3 plus de 15 types d'atomes différents pour le seul carbone. A savoir, alcanes sp^3 , alcènes sp^2 , cyclopropanes sp^2 , carbonyles sp^2 , alcynes sp etc..., tous nécessaires pour rendre MM3 applicable (ce qui signifie l'obtention de résultats raisonnables) pour un ensemble de molécules diverses. On peut constater immédiatement la difficulté de cette approche : le plus d'atomes définis, le plus de paramètres de contribution à la fonction énergie potentielle (liaisons, angles, dièdres...) doivent être développés. Des champs de force plus généraux affecteront donc, un seul type d'atome de carbone générique sp^2 , sacrifiant en faveur d'une application générale. Une autre tendance consiste à utiliser pour les champs de force de classes spécifiques des types d'atomes plus importants en nombres, qu'on ne le ferait dans le cas de paramétrisations pour une application générale.

II.3.2. Forme fonctionnelle des champs de force courants.

Un champ de force ne consiste pas uniquement en une expression mathématique qui décrit l'énergie d'une molécule en fonction des coordonnées atomiques. La deuxième partie indispensable est le jeu de paramètres lui-même. Deux champs de force différents peuvent présenter la même forme fonctionnelle, mais utilisent un paramétrage complètement différent. D'un autre côté, différentes formes fonctionnelles peuvent conduire à des résultats presque identiques, en fonction des paramètres mis en jeu. Cette comparaison montre que les champs de force sont empiriques : il n'y a pas de forme « correcte ».

Parce que certaines formes fonctionnelles donnent de meilleurs résultats que d'autres, la plupart des implémentations dans les logiciels disponibles (académiques et commerciaux) sont très similaires.

Une hypothèse importante est que le champ de force déterminé à partir d'un ensemble de molécules est transférable à d'autres molécules.

Notons qu'un ensemble de paramètres développés et testés sur un nombre relativement petit de cas est encore applicable à une plus large gamme de problèmes. En outre, les paramètres développés à partir de données relatives à de petites molécules peuvent être utilisés pour étudier des molécules plus grandes telles que les polymères.

II.3.3. Quelques exemples

Parmi les champs disponibles nous citerons les plus répandus largement utilisés pour le traitement de petites molécules.

-**MM2, MM3, et MM4** : (<http://europa.chem.uga.edu/allinger/mm2mm3.html>).

Introduit par Allinger *et al.* [II.27-II.30], largement utilisé pour le traitement de petites molécules.

-**AMBER** : (Assisted Method Building and Energy Refinement) (<http://amber.scripps.edu>)

Introduit par Mac kerell *et al.*, [II.31], très largement utilisé dans le traitement des protéines et des acides nucléiques.

-**CHARMM** : (Chemistry at Harvard Molecular Modeling) (<http://yuri.harvard.edu>)

Développé par Mac kerell *et al.*, 1995/1998 ; Brook *et al.*, 1983; [II.31-II.33] qui est largement utilisé pour la simulation de petites molécules et jusqu'aux complexes solvatés de grandes macromolécules biologiques.

CHARMm est une version commerciale disponible de CHARMM qui est également applicable aux petits composés organiques [II.34].

-**MMFF**: (Merck Molecular Force Field)

Développé par Halgren [II.35-II.39], il est similaire à MM3 dans la forme, mais en diffère par son application focalisée sur les processus de condensation de phases (en dynamique moléculaire). Il reproduit l'exactitude MM3 pour les petites molécules, et est applicable aux protéines et autres systèmes d'importance biologique.

Les différents champs de force se distinguent par trois aspects principaux :

- 1) La forme de la fonction de chaque terme énergétique.
- 2) Le nombre de termes croisés (qui reflètent le couplage entre coordonnées internes) inclus.
- 3) Le type d'information utilisé pour ajuster les paramètres.

II.4. Génération des descripteurs moléculaires.

La représentation numérique de la structure chimique (descripteurs moléculaires) est une étape importante de l'investigation QSPR. Les performances du modèle élaboré et la précision des résultats sont étroitement liées au mode de détermination de ces descripteurs.

Nous avons utilisé le logiciel de modélisation moléculaire Hyperchem 6.03 [II.40] pour représenter et optimiser les molécules. Tous les calculs ont été menés dans le cadre du formalisme RHF sans interaction de configuration. Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak- Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,01 kcal/mol. Les géométries finales ont été obtenues à l'aide de la méthode semi-empirique PM3 (Parametric Method 3). Les géométries ainsi optimisées ont été transférées dans le logiciel DRAGON [II.41] pour le calcul de plus de 1600 descripteurs appartenant aux 20 classes différentes suivantes :

1. Descripteurs constitutionnels
2. Descripteurs topologiques
3. Descripteurs Walk and path counts (dénombrements de pas et chemins)
4. Descripteurs de connectivité
5. Descripteurs indices d'information
6. Descripteurs autocorrélation 2D
7. Descripteurs indices d'adjacence
8. Descripteurs valeur propre de Burden
9. Descripteurs indices de charge topologique
10. Descripteurs indices basés sur des valeurs propres
11. Descripteurs profil moléculaire de Randic
12. Descripteurs géométriques
13. Descripteurs RDF
14. Descripteurs Morse - 3D
15. Descripteurs WHIM
16. Descripteurs GETAWAY
17. Descripteurs nombre de groupes fonctionnels
18. Descripteurs fragments des atomes centraux
19. Descripteurs de charge
20. Propriétés moléculaires

En utilisant les options correspondantes du logiciel DRAGON, nous avons d'abord éliminé les descripteurs à valeurs constantes (écarts types inférieurs à 0,001) qui n'apportent aucune information, ensuite ceux qui sont hautement corrélés ($R \geq 0,95$) et qui véhiculent une information redondante. Pour chaque paire de descripteurs corrélés, est éliminé automatiquement celui qui présente les plus hautes corrélations croisées avec les autres descripteurs.

II.5. Méthodes appliquées pour la sélection d'échantillons

La sélection d'échantillons représentatifs est une étape importante dans une procédure d'élaboration de modèles QSAR/ QSPR. En effet, si les jeux d'étalonnage et de validation ne couvrent pas les mêmes domaines de variation, la validation du modèle ne sera pas correcte. Les échantillons d'étalonnage doivent donc répondre à certains critères ; on a identifié 3 règles d'optimalité pour les échantillons de calibrage:

- les échantillons retenus doivent présenter une variabilité maximale;
- la plage de variation des valeurs doit être la plus grande possible, mais limitée aux valeurs rencontrées dans la pratique;
- les échantillons doivent être uniformément répartis.

Différentes méthodes de sélection d'échantillons sont appliquées dans les études QSAR/QSPR, chacune avec ses avantages et ses inconvénients. Parmi ces méthodes de sélection d'échantillons, nous avons appliqué les deux méthodes suivantes :

II.5.1. Sélection aléatoire des échantillons

L'échantillonnage aléatoire simple (au hasard) est la méthode la plus courante pour le fractionnement des données dans le développement des modèles, où les données sont sélectionnées avec une probabilité uniforme. L'échantillonnage au hasard simple est facile à réaliser et peut être efficacement exécuté dans un seul passage sur les données en utilisant des algorithmes tels que l'algorithme de Knuth [II.42].

Cependant, le problème avec cette approche est qu'il y'a une chance que la sélection de données souffre de la variance, ou de partialité, en particulier lorsque les données ne sont pas réparties uniformément [II.43].

Nous avons appliqué le choix aléatoire pour la sélection des ensembles de calibrage et de test lors de la modélisation des points d'éclair des hydrocarbures non-saturés et de n-alcanes.

II.5.2. Algorithme CADEX pour la sélection d'échantillons

L'algorithme de Kennard et Stone [II.44] est une méthode séquentielle d'échantillonnage qui maximise les distances euclidiennes entre les nouveaux échantillons sélectionnés. Elle commence par situer les deux échantillons les plus éloignés l'un de l'autre, qui sont retirés de la base de données initiale et affectés à l'ensemble de calibrage.

Pour chaque échantillon non sélectionné (éch i), l'algorithme :

- calcule la distance vers chaque échantillon déjà sélectionné ;
- attribue à (éch i) la plus petite des distances.

L'échantillon (éch i) associé à la plus grande distance est donc le plus éloigné de tous les échantillons ; c'est donc lui qui est sélectionné.

La procédure est répétée jusqu'à l'obtention du nombre d'échantillons désirés pour l'ensemble de calibrage. Le fait de sélectionner les échantillons les plus éloignés les uns des autres introduit une grande diversité dans l'ensemble de calibrage ; l'obtention d'une répartition uniforme est un autre avantage de cette technique. L'algorithme de Kennard et Stone est parmi les meilleures méthodes de construction des ensembles de calibrage et de test pour les différentes bases des données [II.44].

Cette méthode a été appliquée pour sélectionner les deux sous ensembles de calibrage et de test lors de la modélisation de la température d'ébullition des différentes classes de solvants.

II.6. Sélection d'un sous-ensemble de descripteurs significatifs

Dans toutes les méthodes QSPR, les descripteurs (variables explicatives) sont disposés en colonnes et les molécules en lignes, ce qui aboutit à une matrice. L'objectif consiste à trouver une corrélation entre un ensemble réduit de descripteurs et la propriété physico-chimique ou bien l'activité biologique des molécules. Les algorithmes génétiques constituent une méthode de choix pour la sélection de sous-ensembles de variables explicatives.

II.6.1. Principe de sélection par algorithme génétique

Dans la terminologie des algorithmes génétiques, le vecteur binaire I , appelé "chromosome", est un vecteur de dimension p où chaque position (un "gène") correspond à une valeur (1 si elle figure dans le modèle, 0 sinon). Chaque chromosome représente un modèle basé sur un ensemble de variables explicatives.

On commence par définir le paramètre statistique à optimiser (par exemple maximiser Q^2 en utilisant la validation croisée par "leave-one-out"), avec la taille P de la population du modèle (par exemple, $P = 100$), et le nombre maximum de variables L permises pour le modèle (par exemple, $L = 10$) ; le minimum de variables permises est généralement supposé égal à 1. De plus, une probabilité de croisement p_c (habituellement élevée, $p_c > 0,9$), et une probabilité de mutation p_M (habituellement faible, $p_M < 0,1$) doivent être également définies.

Après définition des principaux paramètres, la mise en œuvre de l'algorithme génétique est démarrée, son évolution comprend trois étapes principales.

II.6.2. Initialisation aléatoire du modèle

La population est constituée au départ de modèles aléatoires avec des variables comprises entre 1 et L , puis les modèles sont ordonnés eu égard au paramètre statistique sélectionné – la qualité du modèle – (le meilleur modèle est en première position, le plus mauvais en position P) ;

II.6.3. Étape de croisement

A partir de la population, on sélectionne des paires de modèles (aléatoirement, ou avec une probabilité proportionnelle à leur qualité). Puis, pour chaque paire de modèles on conserve les caractéristiques communes, c'est-à-dire les variables exclues dans les 2 modèles restent exclues, et les variables intégrées dans les 2 modèles sont conservées. Pour les variables sélectionnées dans un modèle et éliminées dans l'autre, on en essaye un certain nombre au hasard que l'on compare à la probabilité de croisement p_c : si le nombre aléatoire est inférieur à la probabilité de croisement, la variable exclue est intégrée au modèle et vice versa. Finalement, le paramètre statistique du nouveau modèle est calculé : si la valeur de ce paramètre est meilleure que la plus mauvaise pour la population, le modèle est intégré à la population, à la place correspondant à son rang ; dans le cas contraire, il n'est pas pris en compte. Cette procédure est répétée pour de nombreuses paires (100 fois par exemple).

II.6.4. Étape de mutation

Pour chaque modèle de la population (c'est-à-dire pour chaque chromosome) p nombres aléatoires sont éprouvés, et, un à la fois, chacun est comparé à la probabilité de mutation, p_M , définie : chaque gène demeure inchangé si le nombre aléatoire correspondant excède la probabilité de mutation, dans le cas contraire, on le change de 0 à 1 ou vice versa.

Les faibles valeurs de p_M permettent uniquement peu de mutations, conduisant à de nouveaux chromosomes peu différents des chromosomes générateurs.

Après obtention du modèle transformé, on en calcule le paramètre statistique : si cette valeur est meilleure que la plus mauvaise de la population, le modèle est intégré à la population, à la place correspondant à son rang ; dans le cas contraire il n'est pas pris en compte. Cette procédure est répétée pour tous les chromosomes, c'est-à-dire P fois.

II.6.5. Conditions d'arrêt

Les étapes précédentes (croisement et mutation) sont répétées jusqu'à la rencontre d'une condition d'arrêt (par exemple un nombre maximum d'itérations défini par l'utilisateur), ou qu'il est mis fin arbitrairement au processus.

Une caractéristique importante de la sélection d'un ensemble réduit de variables par algorithme génétique est qu'on n'obtient pas nécessairement un modèle unique, mais le résultat consiste habituellement en une population de modèles acceptables ; cette caractéristique, parfois considérée comme un désavantage, fournit une opportunité pour procéder à une évaluation des relations avec la réponse à différents points de vue.

La sélection des descripteurs a été réalisée par algorithme génétique, dans la version MOBYDIGS de Todeschini [II.45], en maximisant le coefficient de prédiction Q_{LOO}^2 .

II.7. Développement des modèles QSPR/QSAR

Les techniques les plus courantes pour établir des modèles QSPR/QSAR utilisent l'analyse de régression (dans les cas où la propriété étudiée est disponible sur une échelle continue), comme la régression linéaire multiple : RLM et la régression par les moindres carrés partiels (MCP ou PLS). Les méthodes de classification, ou en utilisant la régression non-linéaire telle que la régression par les Machines à Vecteurs Supports (SVM) sont également utilisées en général.

II.7.1. La régression linéaire multiple (RLM)

La régression linéaire multiple (RLM) est largement utilisée dans les méthodes QSPR pour sa simplicité, sa transparence, sa reproductibilité, et sa facilité d'interprétation. Une RLM consiste en une équation (multi) linéaire obtenue par régression des données expérimentales en fonction d'un ensemble de descripteurs pré-sélectionnés, en utilisant la méthode des moindres carrés ordinaires (MCO) [II.46].

$$Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + a_3 \cdot X_3 + \dots + a_n \cdot X_n \quad (\text{II.54})$$

Y est la réponse (propriété physique ou activité biologique d'intérêt), $X_1, X_2, X_3, \dots, X_n$ sont les descripteurs ou variables explicatives présentes dans le modèle, $a_1, a_2, a_3, \dots, a_n$ sont les coefficients de la régression et a_0 est le terme constant du modèle.

Supposons qu'on ait mesuré sur n individus et p variables représentées par des vecteurs de \mathfrak{R}^n : y, x_1, x_2, \dots, x_k ; y est la variable dépendante ou à expliquer et les x_j les variables explicatives ou encore prédicteurs (descripteurs moléculaires). On cherche alors à reconstruire y au moyen des x_j par une formule linéaire.

On pose :

$$Y = \beta_0 1 + X(j)\beta(j) + \varepsilon_{(j)} \quad (\text{II.55})$$

Y est un vecteur de dimension n contenant la propriété étudiée des composés chimiques considérés, 1 est un vecteur unité, c'est-à-dire une matrice colonne formée d'éléments égaux à 1, $X(j)$ indique la matrice ($n \times j$), et $\varepsilon(j)$ correspond aux résidus qui doivent suivre une distribution normale; ils peuvent être définis comme la différence entre les valeurs observées et prédites de y notées \hat{y}_i .

$$\varepsilon_i = y_i - \hat{y}_i \quad (\text{II.56})$$

Les estimateurs $\{\beta\}$ sont calculés en utilisant la technique des moindres carrés ordinaires.

II.7.2. Machines à vecteurs supports SVM

Plusieurs générations de machines d'apprentissage ont vu le jour dans le but de classer, de catégoriser ou de prédire des structures particulières dans les données. Mais la plupart de ces techniques éprouvent de grandes difficultés à traiter les données de très haute dimension. Pour surmonter ce problème, on procède souvent par la sélection d'une partie des attributs des données pour réduire la dimension de l'espace d'entrée. Mais dans ce cas on aura besoin d'utiliser des hypothèses simplificatrices qui ne se vérifient pas toujours en pratique.

Par ailleurs, une méthode issue récemment d'une formulation de la théorie de l'apprentissage statistique due en grande partie à l'ouvrage de Vapnik en 1995 intitulé « *The nature of learning statistical theory* » [II.47] surmonte ce problème en imposant que le nombre de paramètres soit linéairement lié au nombre des données d'apprentissage. Cette technique est appelée machines à vecteurs support (SVM : Support Vector Machines, en anglais)

SVM est un modèle discriminant qui tente de minimiser les erreurs d'apprentissage tout en maximisant la marge séparant les données des classes. La maximisation de la marge est une méthode de régularisation qui réduit la complexité du classifieur [II.48].

Cela revient à ne pas considérer les erreurs inférieures à ε et à interdire celles supérieures à ε [II.49].

Maximiser la platitude de la fonction permet de minimiser la complexité du modèle qui influe sur ses performances en généralisation. En effet, la théorie de l'apprentissage [II.48] permet de borner l'erreur de généralisation par une somme de deux termes : l'un dépendant de la complexité du modèle et l'autre dépendant de l'erreur sur les données d'apprentissage [II.50].

Les méthodes SVM sont basées sur le contrôle de la complexité du modèle lors de l'apprentissage.

Dans la méthode SVM, différents hyperparamètres apparaissent : C , qui représente le compromis entre la complexité du modèle et l'erreur sur les données d'apprentissage; ε , qui correspond à la largeur du tube d'insensibilité ; les éventuels paramètres de la fonction noyau k (σ, γ, \dots). Ces hyperparamètres sont en général réglés en fonction d'une estimation de

l'erreur de généralisation qui peut être évaluée sur un jeu indépendant de données de validation ou par validation croisée [II.51].

Cela implique de réaliser l'apprentissage pour différentes valeurs et d'estimer leur performance. Dans le cas d'une estimation de l'erreur de généralisation par validation croisée, cette procédure peut se révéler très coûteuse en temps de calcul.

II.7.3. Paramètres statistiques d'évaluation d'un modèle QSAR/QSPR

L'analyse de régression linéaire multiple (RLM) et la sélection des variables ont été effectuées avec le logiciel MobyDigs [II.45] en utilisant la méthode de régression par des moindres carrés ordinaire (OLS) et l'algorithme génétique (GA-VSS) pour la sélection des sous-ensembles [II.52].

Deux paramètres sont couramment utilisés comme critères pour le choix du modèle linéaire :

- **Le coefficient de détermination multiple:** Le coefficient de détermination R^2 est une mesure de la précision de l'ajustement de la droite de régression. Il augmente de façon monotone avec l'introduction de nouvelles variables même si celles-ci sont peu corrélées avec la variable expliquée Y . Son utilisation n'est donc pas recommandée sauf dans le cas de modèle à nombres de variables identiques. Il est donné par l'équation suivante :

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\text{SCE}}{\text{SCT}} \quad (\text{II.57})$$

Où

\hat{y}_i est la valeur estimée de la propriété étudiée de l'ensemble de calibrage, et \bar{y} la moyenne des valeurs observées de cet ensemble.

SCE est la somme des carrés des résidus ($e_i = y - \hat{y}_i$) et SCT la somme des carrés totale (= $\sum (y_i - \bar{y})^2$)

- La racine de l'erreur quadratique moyenne de prédiction (désignée également par SDEP) :

$$\sigma_N = \text{EQMP} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2} = \sqrt{\frac{\text{PRESS}}{n}} \quad (\text{II.58})$$

PRESS (Predictive Residual Sum of Squares) dans l'équation précédente est une mesure de la dispersion des estimations.

- La racine de l'écart quadratique moyen (RMSE) calculée sur les ensembles de calibrage (EQMC), et sur l'ensemble de validation externe (EQMP_{ext}), sont deux paramètres à considérer, également. Ces deux paramètres sont définis par les équations suivantes :

$$\text{EQMC} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (\text{II.59})$$

$$\text{EQMP}_{\text{ext}} = \sqrt{\frac{\sum_{i=1}^{n_{\text{ext}}} (y_i - \hat{y}_{(i)})^2}{n_{\text{ext}}}} \quad (\text{II.60})$$

- La stabilité des modèles a été explorée en utilisant la "validation croisée par omission d'une observation" (LOO : cross-validation by leave-one-out) [II.53]. Elle consiste à recalculer le modèle sur (n-1) composés, le modèle obtenu servant alors à estimer la valeur de la propriété du composé écarté noté $\hat{y}_{(i)}$. On répète le procédé pour chacun des composés étudiés.

$$Q_{\text{LOO}}^2 = \frac{\text{SCT} - \text{PRESS}}{\text{SCT}} \quad (\text{II.61})$$

Contrairement à R^2 qui augmente avec le nombre de paramètres du modèle, le facteur Q_{LOO}^2 affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande importance au

coefficient Q_{LOO}^2 . Une valeur de $Q_{LOO}^2 > 0,5$ est considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [II.53,II.54].

Si de petites valeurs de Q_{LOO}^2 indiquent des modèles peu robustes, caractérisés par de faibles capacités prédictives internes, le contraire n'est pas nécessairement vrai. En fait, si une forte valeur de Q_{LOO}^2 est une condition nécessaire de robustesse et d'une possible capacité prédictive élevée d'un modèle, cette condition seule n'est pas suffisante, et peut conduire à une surestimation de la capacité prédictive du modèle lorsqu'il est appliqué à des composés réellement externes.

- La technique de validation du bootstarp est une validation interne. Elle consiste à simuler k échantillons de taille n , la même que celle de l'échantillon initial. Le tirage est fait au hasard. Le modèle obtenu pour l'ensemble des objets sectionnés sera utilisé pour prédire l'ensemble exclu. Le bootstrapping a été répété 500 fois pour la validation de chaque modèle [II.55].

Obtenir un modèle robuste ne donne aucune information réelle sur son pouvoir prédictif. Cela sera évalué par la prédiction du modèle en utilisant l'ensemble de test. Le paramètre Q_{ext}^2 [II.56] qui a pour principe de valider un nombre d'observations qui n'ont pas contribué à la construction du modèle, est déterminé par l'équation suivante :

$$Q_{ext}^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2 / n_{tr}} \quad (II.62)$$

Où

y_i et $\hat{y}_{i/i}$ sont respectivement les valeurs mesurées et prédites (sur l'ensemble de prédiction) de la variable dépendante, et \bar{y} est la valeur moyenne de la variable dépendante pour l'ensemble de calibrage. n_{tr} et n_{ext} sont respectivement le nombre des observations dans l'ensemble de calibrage et l'ensemble de test.

- D'autres paramètres utiles sont : l'écart-type de prédiction externe ($SDEP_{ext}$), défini comme suit:

$$SDEP_{\text{ext}} = \sqrt{\frac{1}{n_{\text{ext}}} \sum_{i=1}^{n_{\text{ext}}} (y_i - \bar{y})^2} \quad (\text{II.63})$$

Où la somme porte sur l'ensemble des objets tests (n_{ext}).

- Selon Golbraikh et Tropsha [II.57], un bon modèle QSPR doit satisfaire aux critères suivants:

$$R^2_{\text{CV}_{\text{ext}}} > 0.5 \quad (\text{II.64})$$

$$r^2 > 0.6 \quad (\text{II.65})$$

$$(r^2 - r_0^2) / r^2 < 0,1 \text{ or } (r^2 - r_0'^2) / r^2 < 0,1 \quad (\text{II.66})$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15 \quad (\text{II.67})$$

$$Ab = |r^2 - r_0'^2| < 0.3 \quad (\text{II.68})$$

Où:

$$r = \frac{\sum (y_i - \tilde{y}_i)(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (\text{II.69})$$

$$r_0^2 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^0)}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \quad (\text{II.70})$$

$$r_0'^2 = 1 - \frac{\sum (y_i - y_i^0)^2}{\sum (y_i - \bar{y})^2} \quad (\text{II.71})$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \quad (\text{II.72})$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad (\text{II.73})$$

$$T1 = \frac{(r^2 - r_0^2)}{r^2} \quad (\text{II.74})$$

$$T2 = \frac{(r^2 - r_0'^2)}{r^2} \quad (\text{II.75})$$

Ces critères de validation externe sont appliqués uniquement à l'ensemble de test et recommandés pour évaluer la prédictivité du modèle QSPR.

r^2 est le coefficient de corrélation entre les valeurs calculées et expérimentales de l'ensemble de test; r_0^2 (valeurs calculées par rapport à celles observées) et $r_0'^2$ (valeurs observées par rapport à celles calculées) sont les coefficients de détermination; k et k' sont les pentes des droites de régressions passant par l'origine pour les valeurs calculées par rapport aux valeurs observées et observées par rapport aux calculées, respectivement; $y_i^{r_0}$ et $\tilde{y}_i^{r_0}$ sont donnés respectivement par : $y_i^{r_0} = k \tilde{y}_i$ et, $\tilde{y}_i^{r_0} = k' y_i$; et les sommations portent sur tous les échantillons dans l'ensemble de test.

II.7.4. Analyse des résidus pour la détection des observations aberrantes :(Commande « regstasts » du logiciel Matlab [II.58])

Les tests statistiques utilisés pour la détection des observations aberrantes en analyse de régression sont réunis dans les tableaux III.5 et III.6 pour les hydrocarbures non saturés, les tableaux III.9 et III.10 pour les n-alcanes et dans les tableaux III.16 et III.17 pour la série des solvants.

- Les résidus ordinaires e_i , différences entre les valeurs observées (y_i) et estimées par le modèle (\hat{y}_i).

$$e_i = y_i - \hat{y}_i \quad (\text{II.76})$$

Les résidus de prédiction standardisés $e_{i\text{std}}$ ou bien d_i , obtenus en divisant les e_i par l'écart type estimé(s).

- Les leviers, h_{ii} , permettent de juger de l'influence d'une observation i dans la détermination de l'équation de régression.

Les valeurs de h_{ii} , $i^{\text{ème}}$ terme diagonale de la matrice de projection (ou matrice chapeau) :

$H = X(X'X)^{-1}X'$ où X est la matrice des valeurs observées des variables explicatives et X' sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques.

Les observations avec des valeurs supérieures à $\frac{2(k+1)}{n}$ sont considérées comme potentiellement très influentes, où k est le nombre de prédicteurs et n est le nombre d'observations.

- Le résidu studentisé interne r_i , est le résidu d'une prédiction divisé par son écart type propre.

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \quad (\text{II.77})$$

- Les estimations $S_{(i)}^2$ de σ^2 calculées selon l'équation suivante :

$$S_{(i)}^2 = \frac{(n-p)\text{CME} - \frac{e_i^2}{1-h_{ii}}}{n-p-1} \quad (\text{II.78})$$

Pour $n-1$ observations, la $i^{\text{ème}}$ étant exclue ; CME est le carré moyen des écarts, et $k = (p - 1)$ le nombre de variables explicatives.

II.7.5. Diagnostics d'influence

D'autres mesures d'influence utiles dont l'étude complète la recherche des observations aberrantes [II.59].

- La statistique représentée par le symbole DFFITS ou DFITS permet de mesurer l'influence d'une observation i sur la valeur ajustée ou prédite ;

$$DFFITS = \frac{1}{p} \sqrt{\left(\frac{h_{ii}}{1-h_{ii}} \right)} t_i \tag{II.79}$$

Les DFFITS combinent le bras de levier h_{ii} et le résidu studentisé en une mesure globale, qui renseigne jusqu'ou une observation est inhabituelle.

Les valeurs absolues de $DFFITS_i$ supérieures à la valeur critique $2\sqrt{\frac{p}{n}}$ avec $(p=k+1)$ sont inhabituelles.

- La distance de Cook, D_i , permet d'étudier l'influence d'une observation i sur les coefficients de régression estimés par les moindres carrés. Les valeurs supérieures à $(4/n)$ sont considérées comme très influentes.

$$D_i = \frac{1}{p} \frac{h_{ii}}{1-h_{ii}} d_i^2 \tag{II.80}$$

Cette relation montre que la distance D_i est une fonction croissante du carré du résidu standardisé d_i et de h_{ii} . Pour une valeur fixée de p et, pour une régression avec terme indépendant, D_i sera d'autant plus grand que e_i est grand, en valeur absolue et que le vecteur X_i est éloigné du vecteur \bar{X}_i .

Les valeurs de D_i peuvent être comparées à une valeur $F_{1-\alpha}$ relative à la variable F de SNEDECOR à p et $(n-p)$ degrés de liberté, bien qu'il ne s'agisse pas d'un test statistique rigoureux [II.60].

Sur cette base, WEIBERG considère qu'une attention particulière doit être donnée aux observations ayant une valeur de D_i supérieure à l'unité.

En fait, il est suggéré de contrôler les observations pour lesquelles :

$$COOK_i > F(p;n-p) \cong 1$$

- La disposition des points dans l'espace des x (régresseurs) est importante pour la détermination des propriétés du modèle. En particulier, les observations éloignées ont, potentiellement, des influences disproportionnées sur les paramètres estimés, les valeurs prédites, et les statistiques élémentaires. La somme pondérée des carrés des distance du point i au centre des données $WSSD_i$ (Weighted of the Sum Squared Distance of the

center of data) peut être utilisée pour localiser les points éloignés dans l'espace des x [II.60] soit:

$$WSSD_i = \sum_{j=1}^n \left[\frac{\hat{\beta}_j (X_{ij} - \bar{X}_j)}{\sqrt{CME}} \right]^2, \quad i = 1, 2, \dots, n \quad (II.81)$$

II.7.6. Statistique DFBETAS_{j,i}

L'ampleur du changement, en unité d'écart-type, du $j^{\text{ème}}$ coefficient de régression $\hat{\beta}_j$ estimé, est indiquée par la statistique :

$$DFBETAS_{i,j} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}} \quad (II.82)$$

Où C_{jj} est le $j^{\text{ème}}$ élément diagonal de la matrice et $\hat{\beta}_{j(i)}$ le $j^{\text{ème}}$ coefficient de régression calculé sans utiliser l'observation i . L'examen de la colonne $DFBETAS_{j,i}$ en tenant compte de la valeur critique $\frac{2}{\sqrt{n}}$ fait ressortir les points influents, avec n le nombre d'individus de calibrage.

II.7.7. COVRATIO_i

Le $COVRATIO_i$, que l'on calcule pratiquement à partir de la relation (II.83) fait ressortir le rôle de la $i^{\text{ème}}$ observation sur la précision de l'estimation

$$COVRATIO_i = \frac{(S_{(i)})^p}{(CME)^p} \left(\frac{1}{1 - h_{ii}} \right), \quad i=1, 2, \dots, n \quad (II.83)$$

En résumé si $COVRATIO_i > 1$, la $i^{\text{ème}}$ observation améliore la précision de l'estimation, alors que si $COVRATIO_i < 1$, l'inclusion de la $i^{\text{ème}}$ observation détériore la précision.

II.7.8. Test de randomisation

Ce test permet de mettre en évidence des corrélations dues au hasard. Il consiste à générer un vecteur « propriété considérée » par permutation aléatoire des composantes du vecteur réel. On calcule alors sur le vecteur obtenu (considéré comme vecteur expérimental réel) un modèle QSAR, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 dans notre cas).

II.7.9. Le domaine d'application du modèle QSPR

Le domaine d'application (AD) [II.56-II.54] est une région théorique dans l'espace défini par les descripteurs du modèle et la réponse à modéliser, pour laquelle un QSPR donné devrait faire des prédictions fiables. Dans ce travail, le domaine d'application structurale a été vérifié par l'approche du levier (h_{ii}) [II.60].

Le levier critique est en général fixé à $3(k+1)/n$, où n est le nombre total des objets dans l'ensemble de calibrage et k est le nombre de descripteurs impliqués dans la corrélation.

La présence des valeurs aberrantes en réponse (Y outliers) et des composés structurellement influents (X outliers) a été, en plus, vérifiée par le diagramme de Williams [II.61], représentant les résidus standardisés en fonction des valeurs des leviers h_{ii} .

RÉFÉRENCES BIBLIOGRAPHIQUES

- [II.1] Pople, J. A., Beveridge, D. L., Approximate Molecular Theory, Mc Graw- Hill (196).
- [II.2] Hartree, D.R., (1928), "The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods". Proc. Cambridge. Phil. Soc.24, 328.
- [II.3] Fock, V., (1930), "Näherungsmethode zur lösung des quantenmechanischen mehrkörper problems Z. Physik.61, 126 .
- [II.4] Slater, J. C., Phys. Rev. The self consistent field and the structure of atoms, 1928. 32, 339 ; atomic shielding constants. 1930.35, 1210.
- [II.5] Löwdin, P. O., (1950), "On the Non Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals". J. Chem. Phys, 18, 365.
- [II.6] Pariser, R., Parr, R. G., (1953), "A Semi Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules .I", J. Chem. Phys. 21, 466.
- [II.7] Pariser, R., Parr, R. G., (1953), "A Semi Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules. II", J. Chem. Phys. 21, 767.
- [II.8] Pople, J. A., (1953), Electron interaction in unsaturated hydrocarbons, Trans. Faraday Soc. 49, 1375.
- [II.9] Pople, J. A., Santry, D. P., Segal, G. A., (1965), "Approximate Self Consistent Molecular Orbital Theory. I. Invariant Procedures", J. Chem. Phys., 43, S129-S135.
- [II.10] Pople, J.A., Beveridge, D. L., Dobosh, P. A., (1967), "Approximate Self Consistent Molecular Orbital Theory. V. Intermediate Neglect of Differential Overlap". J. Chem. Phys, 47, 2026.
- [II.11] Dewar, M. J. S., Thiel, W., (1977), " Ground states of molecules. 38. The MNDO method. Approximations and parameters", J. Am. Chem. Soc., 1977, 99, 4899- 4907, 4907-4914.
- [II.12] Thiel, W., Encyclopedia of Computational Chemistry, Vol. 3, P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Schreiner (Eds.), Wiley, Chichester, 1998, p. 1599.
- [II.13] Burstein, K. Y., Isaev, A.N., Theor. Chim. Acta 1984, 64, 397- 401.
- [II.14] Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., Stewart, J. J. P., J. Am. Chem. Soc. 1985, 107, 3902- 3909 ; AM1
- [II.15] Holder, A. J., AM1, in Encyclopedia of Computational Chemistry. Vol.1 (Eds.: P. V. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer, III, P. R. Schreiner), Wiley, Chichester, 1998, pp. 8-11.

- [II.16] Stewart, J. J. P., *J. Comput. Chem.*, 10, 1989, 209- 220 ; 221- 264; PM3 :
- [II.17] Stewart, J. J. P., *Encyclopedia of Computational Chemistry*; Vol. 3, Schleyer, P. V .R., Allinger, N. L., Clark .T., Gasteiger, J., Kollman, P. A., Schaefer III, H .F., Schreiner, P. R., (Eds.), Wiley, Chichester, 1998, pp. 2080-2086.
- [II.18] Jensen, F., *Introduction to Computational Chemistry*, Wiley, pp. 94- 96.
- [II.19] Stewart, J. J. P., *Application of Localized Molecular Orbitals to the Solution of Semiempirical Self-Consistent Field Equations*, *Int. J. Quantum. Chem.* 1996. 58 , 133.(page
- [II.20] Daniels,A. D., Millam,J. M., Scuseria,G. E., *J. Chem. Phys.*, 1997.107 , 425.
- [II.21] Ramachandran,K. I., Deepa,G.,Namboori, K.,"*Comptutional Chemistry and Molecular Modeling, Principles and Applications* ", Springer-Verlag Berlin Heidebberg, 2008.
- [II.22] Coulson, C. A., Longuet- Higgins, H. C., "*The Electronic Structure of Conjugated Systems*". *I. R Proc. Roy. Soc. (London) A* .1947.191, 39 .
- [II.23] Mulliken, R. S.,(1926), "*Criteria for the Construction of Good Self Consistent Field Molecular Orbital Wave Functions, and the Significance of LCAOMO Population Analysis*" , *J. Chem. Phys.*, 36, 3428.
- [II.24] Kutzelnigg, W., Del Re, G., Berthier, G., (1971), " *σ and π Electrons in Theoretical Organic Chemistry*", Springer Verlag, Berlin.
- [II.25] Pullman, B.,(1969), "*La Biochimie Electronique*", Collection Que sais-je ? PUF, n°1075, Deuxième édition, Paris.
- [II.26] Boyd, D. B., Lipkowitz, K. B., eds. *Reviews in Computational Chemistry, History of the Gordon Conferences on Computational Chemistry*, Wiley- VCH, New York, 2000.399-439.
- [II.27] Allinger, N. L., *Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms*, *J. Am. Chem. Soc.*, 1977, 99, 8127.
- [II.28] Burkert ,U., Allinger ,N. L., *Molecular Mechanics*, ACS Monograph No. 177, American Chemical Society, Washington, DC, 1982, 1986.
- [II.29] Allinger, N. L., Yuk ,Y. H., Lii, J. H., *Molecular Mxhanics. The MM3 Force Field for Hydrocarbon 3. 1*, *J. Am. Chem. Soc.*, 1989, 111, 8551.
- [II.30] Allinger, N. L., Chem,K., Katzenellbogen,J. A., Wilson,S. R., Anstead,G. M.,(1996), *Hyperconjugative Effects on Carbon-Carbon Bond Lengths in Molecular Mechanics (MM4)*, *J. Comput. Chem.*, 17, 747.
- [II.31] Mac Kerell, A. D., Jr., Bashford, D., Bellott, M., Dumbrack, R. L., Evaseck, J. D., Field, M. J., Fischer, S., Gao, J., Gao, H., He, S., Joseph- Mac Carthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michmick, S., Nego, T., Nguyen, D. T., Prodhom, B.,

- Reiher, W. E. III., Roux, B., Schlemkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., Karplus, M., *J. Phys. Chem. B*, 1998, 102, 3586-3616.
- [II.32] MacKerell, A. D. Jr., Wiorkiewicz-Kuczera, J., and Karplus, M., (1995), "An all-atom empirical energy function for the simulation of nucleic acids", *Journal of the American Chemical Society*, 117 (48), 11946-11975
- [II.33] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States David, J., Swaminathan, S., Karplus, M., CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations *J. Comput. Chem.* 1983.4, 187,
- [II.34] Momany, F.A., Rone, R., (1992), "Validation of the General Purpose QUANTA 3.2/CHARMm Force Field". *J. Comput. Chem.*, 13 (7):888–900.
- [II.35] Halgren, T. A., (1996d), "Merck Molecular Force Field: I. Basis, Form, Scope, Parameterization and Performance of MMFF94", *J. Comp. Chem.* 17, 490-519.
- [II.36] Halgren, T. A., (1996a), "Merck Molecular Force Field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions", *J. Comp. Chem.*, 17, 520-552.
- [II.37] Halgren, T. A. (1996b), " Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies", *J. Comp. Chem.* 17, 553-586.
- [II.38] Halgren, T. A., (1996c), "Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additional Computational Data and Empirical Rules", *J. Comp. Chem.* 17, 616-641.
- [II.39] Halgren, T. A., Nachbar, R. B., (1996). "Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94", *J. Comp. Chem.* 17, 587-615.
- [II.40] Hyperchem™. 2000. Release 6,03 for windows. Molecular Modeling System.
- [II.41]. Todeschini, R., Consonni, V., Mauri, A. and Pavan, M. (2006), "DRAGON Software for the calculation of molecular descriptors". Release 5.3 for Windows, Milano. Italy.
- [II.42] Donald, E. K., (1997), "The art of computer programming", Vol,2 (3rd ed), Boston: Addison Wesley.
- [II.43] Tourassi, G.A., Frederick, E. D., Markey, M. K., Floyd, C.E., (2001), "Application of the mutual information criterion for feature selection in computer-aided diagnosis", *Medical Physics*, Vol, 28(12), pp, 2394-2402.
- [II.44] Kennard, R., Stone, L. A., (1969), "Computer aided design of experiments *Technometrics*", Vol. 11 No. 1, pp. 137-148.

- [II.45] Todeshini, R., Ballabio, D., Consonni, V., Mauri, A., Pavan, M., (2009), MOBY DIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, Release 1,1 for windows, Milano.
- [II.46] Snedecor, G. W., Cochran, W.G., (1967), "Statistical methods", 6th Edition, Oxford and IBH Publishing Co., Bombay/ New Delhi.
- [II.47] Vapnik, V.,(1995), " The nature of Statistical Learning Theory" , *Springer-Verlag*, New York.
- [II.48] Vladimir, N., Vapnik, V., (1995), "The nature of statistical learning theory", Springer-Verlag, New York, USA.
- [II.49] Smola Alex, J., Bernhard Schölkop, A.,(2004), "Tutorial on support vector regression", *Statistics and Computing*, Vol 14(3), pp, 199-222.
- [II.50] Cristianini, N., Shawe-Taylor, J., "*An Introduction to Support Vector Machines and other Kernel Based Learning Methods*", Cambridge University, Press, 2000.
- [II.51] Bishop, C.M., (1995), "Neural Networks for Pattern Recognition", Oxford University Press.
- [II.52] Leardi, R., Boggia, R., Terrile, M., (1999), "Genetic algorithms as a strategy for feature selection", *J. Chemom*, Vol, 6, pp, 267- 281.
- [II.53] Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T. D., Mc Dowell, R. M., Gramatica, P., (2003), "Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification and Regression Based QSARs", *Health Perspectives*, Vol. 111 No. 10, pp.1361-1375.
- [II.54] Tropsha, A., Gramatica, P., Grombar, V. K., (2003), "The importance of being earnest validation in the absolute essential for successful application and interpretation of QSPR models", *QSAR and Combinatorial Science*, Vol 22 No.1, pp. 69-76.
- [II.55] Efron, B., (1994), "The jackknife, the Bootstrap and Other Resampling Planes", Society for Industrial and Applied Mathematics, Philadelphia, PA.
- [II.56] Shi, L.M., Fang, H., Tong, W., Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C.L., Sheehan, D.M., (2001), " QSAR models using a large diverse set of estrogens", *Journal of Chemical Information and Computer Science*, Vol. 41, pp. 186-195.
- [II.57] Golbraikh, A., Tropsha, A., (2002),"Beware of q^2 !", *Journal of Molecular Graphics and modelling*, Vol. 20 No 4, pp. 269-276.
- [II.58] MATLAB, Version 7.8.0.347 (R2009a), The Language of Technical Computing, The Math Works, February 12 (2009).

[II.59] Montgomery, D.C., Peck, E.A., (1992) Introduction to Linear Regression Analysis, Second Edition, Wiley-Interscience, New York, pp.527,

[II.60] Weisberg, S. (2005), "Applied Linear Regression", 3rd edn, John Wiley and sons, Inc, New Jersey.

[II.61] SCAN - Software for Chememometric Analysis, (1995), Version1.1 - for Windows, Minitab USA.



Chapitre III



**III: Résultats et
discussions**

III.1. Modélisation et prédiction des points d'éclair des hydrocarbures non saturés en utilisant l'approche hybride algorithme génétique / régression linéaire multiple.

III.1.1. Introduction

Le point (ou température) d'éclair (T_{ec}) est défini(e) comme la température la plus basse, corrigée à 101,325 kPa, à laquelle l'application d'une source d'inflammation provoque l'inflammation des vapeurs dans les conditions spécifiques du test [III.1,III.4].

Ce paramètre fournit l'explication des processus physiques et chimiques fondamentaux de la combustion. De plus, il est important pratiquement pour les conditions de sécurité lors du stockage, le traitement et la manipulation d'un composé donné. Et c'est l'une des principales caractéristiques d'inflammabilité utilisées pour évaluer les risques d'incendie et d'explosion des composés organiques [III.5].

Le point d'éclair de la plupart des composés peut être mesuré par deux méthodes expérimentales actuellement acceptées, qui sont le test de coupe fermée et le test de coupe ouverte [III.6]. Cependant, pour de nombreux autres composés, les valeurs du point d'éclair expérimental sont rares et trop coûteuses à obtenir. En outre, il est encore plus difficile de réaliser la détermination expérimentale du point d'éclair des composés toxiques, volatils, explosifs et radioactifs. Par conséquent, le développement de méthodes d'estimation pour prédire le point d'éclair, est nécessaire.

Parmi les méthodes principales de prédiction de cette propriété on peut citer la méthode de contribution de groupe (MCG), l'analyse par composantes principales (ACP) et la relation quantitative structure-propriété (QSPR).

Plusieurs modèles QSPR développés pour prédire le point d'éclair ont été publiés dans la littérature [III.7,III.9]. Vidal et al. [III.7] ont présenté les méthodes les plus importantes pour la prédiction du point d'éclair.

Une étude de prédiction des points d'éclair d'un grand ensemble de données de divers types d'hydrocarbures cycliques et acycliques [III.8]. Une corrélation simple a été proposée ; elle est

basée sur le nombre de carbones et d'atomes d'hydrogène et certaines fractions moléculaires spécifiques, qui peuvent être facilement utilisées pour tout type d'hydrocarbures.

Une autre méthode a été introduite pour la prédiction des points d'éclair de différentes classes d'hydrocarbures non saturés [III.9]. Une fonction centrale montrant que le nombre de carbones et d'atomes d'hydrogène peut être utilisé comme la fonction centrale et peut être révisée par une fonction de correction. La fonction de correction contient deux termes correcteurs déterminés sur la base de la structure moléculaire des hydrocarbures insaturés.

III.1.2 Méthodologies

III.1.2.1 La collecte des données

L'ensemble des composés étudiés, dont les caractéristiques ont été prélevées dans la littérature [III.9], est formé de différentes classes d'hydrocarbures insaturés (des alcènes, des alcynes et des composés aromatiques). Les valeurs des points d'éclair se distribuent dans la fourchette : 137 - 451 K. L'application de l'approche algorithme génétique a mené à la sélection d'un ensemble réduit de 269 descripteurs [III.10].

En première étape, l'ensemble de données a été éclaté de manière aléatoire en deux sous ensembles disjoints : un ensemble de calibrage formé de 139 composés pour la construction du modèle ; les 34 composés restants constituent l'ensemble de test externe destiné à l'évaluation du pouvoir prédictif du modèle calculé. Le tableau III.1 présente l'ensemble des composés étudiés (calibrage et test).

Tableau III.1: Nomenclature et valeurs des points d'éclair des composés étudiés

N°	Composés	Tec(K)	N°	Composés	Tec(K)
1	Benzène	262	6	Propylbenzène	303
2*	Toluène	280	7	Cumène	304
3*	Éthyl benzène	288	8	m-Éthyltoluène	311
4*	p-Xylène	300	9	1,2,3-Triméthylbenzène	324
5	o-Xylène	303	10*	1,2,4-Triméthylbenzène	321

Tableau III.1 (suite)

N°	Composés	Tec(K)	N°	Composés	Tec(K)
11	1,3,5-Triméthylbenzène	317	34	1,3-Diméthylnaphtalène	382
12	o-Éthyltoluène	312	35*	1,2-Diméthylnaphtalène	374
13*	p-Éthyltoluène	309	36	Hexylbenzène	356
14	Naphtalène	360	37	Hexaméthylbenzène	377
15	Butylbenzène	331	38	3,5-Diméthyl-tert-butylbenzène	357
16	1,2,4,5-Tétraméthylbenzène	346	39	1,2,4-Triméthylbenzène	349
17	2-Éthyl-p-xylène	329	40*	1,3,5-Triéthylbenzène	354
18*	3-Éthyl-o-xylène	338	41	1,4-Diisopropylbenzène	354
19*	4-Éthyl-m-xylène	330	42	m-Diisopropylbenzène	350
20*	Tert-Butylbenzène	307	43	n-Héptylbenzène	368
21	p-Cymène	320	44	1,2,3,4-Tétraéthylbenzène	367
22	o-Diéthylbenzène	322	45	2-Phényloctane	373
23	m-Diéthylbenzène	324	46	n-Octylbenzène	380
24	p-Diéthylbenzène	328	47*	n-Nonylbenzène	390
25	4-Éthyl-1,2-diméthylbenzène	331	48	1,3,5-Triisopropylbenzène	359
26	1-Méthylnaphtalène	355	49	Décylbenzène	380
27	n-Pentylbenzène	339	50	Pentaéthylbenzène	386
28	IsoPentylbenzène	335	51*	n-Undécylbenzène	409
29	Pentaméthylbenzène	364	52	Dodécylbenzène	418
30	p-tert-butyltoluène	321	53	1,2,4,5-Tétraisopropylbenzène	397
31	2-Phenyl-2méthylbutane	338	54	1,3,5-Tritert-butylbenzène	372
32	1-Éthylnaphtalène	380	55	Tridécylbenzène	385
33	2-Éthylnaphtalène	377	56*	1-Méthylantracène	430

Tableau III.1 (suite)

N°	Composés	Tec(K)	N°	Composés	Tec(K)
57	2-Méthylanthracène	431	80	3-Méthylbut-1-yne	221
58	9-Méthylanthracène	431	81	Hex-1-yne	252
59	1-méthylphenanthrène	431	82*	Hex-2-yne	263
60	7-isopropyl-1-méthylphénanthrène	451	83	Hex-3-yne	259
61	Phénylacétylène	303	84	3,3-Diméthylbut-1-yne	239
62	Styrène	304	85	4-Méthylpent-1-yne	249
63*	2-Vinyltoluène	320	86	Hepta-1,6-diyne	282
64	3-Vinyltoluène	324	87	Hept-1-yne	263
65	3-Phenylprop-1-ène	310	88	Hept-2-yne	275
66	beta-Méthylstyrène	333	89	Hept-3-yne	257
67*	Cis-1-Propenylbenzène	325	90*	3-Méthylhex-1-yne	268
68	Isopropenylbenzène	313	91	Octa-1,7-diyne	296
69*	Trans-1-phénylprop-1-ène	331	92	Octa-2,6-diène	307
70	m-Divinylbenzène	338	93	Oct-1-yne	289
71	p-Divinylbenzène	337	94	Oct-2-yne	301
72	1-Butynylbenzène	341	95*	Oct-4-yne	291
73	3-Éthylstyrène	333	96*	Nona-1,8-diyne	314
74	4-Éthylstyrène	335	97	Non-1-yne	306
75	2,4-Diméthyl-1-vinylbenzène	333	98	Déc-1-yne	323
76	Acétylène	155	99	Undéc-1-yne	338
77	Propyne	186	100	Undéc-4-yne	341
78	Pent-1-yne	230	101	Dodéc-1-yne	352
79	Pent-2-yne	253	102	Tridéc-1-yne	366

Tableau III.1 (suite)

N°	Composés	Tec(K)	N°	Composés	Tec(K)
103	Cyclobutène	202	126	Trans-pent-2-ène	225
104	Cyclopentène	244	127	Isopentène	211
105	Cyclohexène	256	128	cyclohexa-1,4-diène	248
106	4-Méthylcyclopentène	243	129*	Hexa-2,4-diène	264
107	Cycloheptène	267	130	Hexa-1,5-diène	246
108	4-Méthylcyclohexène	272	131	2,3-Diméthylbuta-1,3-diène	251
109*	3-Méthylcyclohexène	270	132	3-Méthylpenta-1,4-diène	239
110	4-Éthylcyclohexène	286	133	2-Méthylpenta-2,3-diène	255
111	Éthylène	137	134	Hex-1-ène	253
112	Propène	165	135	Cis-Hex-2-ène	252
113	Propadiène	177	136*	Cis-Hex-3-ène	261
114	Buta-1,2-diène	197	137	Trans-Hex-3-ène	261
115	Buta-1,3-diène	197	138*	Isohexène	241
116*	Butène	194	139	2,3-Diméthylbut-1-ène	255
117	Cis-but-2-ène	200	140	2,3-Diméthylbut-2-ène	256
118	Isobutylène	197	141	3,3-Diméthylbut-1-ène	244
119	Penta-1,2-diène	233	142	2-Méthylpent-1-ène	241
120	Penta-2,3-diène	235	143	2-Méthylpent-2-ène	246
121	Cis-penta-1,3-diène	232	144	4-Méthylpent-2-ène	241
122*	2-Méthylbutadiène	225	145	3-Méthylpent-1-ène	244
123*	Pentène	229	146	Trans-3-Méthylpent-2-ène	266
124	Pent-2-ène	253	147*	2-Éthylbut-1-ène	243
125	Cis-pent-2-ène	227	148	Hepta-1,6-diène	263

Tableau III.1 (suite et fin)

N°	Composés	Tec(K)	N°	Composés	Tec(K)
149	Hept-1-ène	264	162	Cis-Oct-4-ène	294
150	Cis-Hept-2-ène	265	163*	Nona-1,8-diène	299
151	Trans-Hept-2-ène	267	164	2-Éthylhex-1-ène	279
152	Trans-Hept-3-ène	266	165	Non-1-ène	298
153	2-Méthylhex-1-ène	267	166*	Undéc-1-ène	336
154*	4-Méthylhex-1-ène	258	167	Dodécène	351
155*	2-Éthylpent-1-ène	263	168	2-Méthylundéc-1-ène	345
156*	2,4-Diméthylpent-2-ène	264	169	Tridéc-1-ène	352
157	2,3,3-Triméthylbut-1-ène	256	170	Tétradéc-1-ène	383
158*	Cis-5-Méthylhex-2-ène	268	171	Pentadéc-1-ène	386
159	Trans-5-Méthylhex-2-ène	268	172	Hexadéc-1-ène	402
160	Trans-Oct-3-ène	282	173	Heptadéc-1-ène	408
161	Trans-Oct-4-ène	281	/	/	/

* Composés de l'ensemble de test

III.1.3 Résultats et discussions

III.1.3.1 Développement et validation de modèle

Une analyse par régression linéaire multiple (MLR) pour la sélection des variables a été effectuée après la génération des descripteurs moléculaires, en utilisant le logiciel Mobydigs [III.11] et en appliquant la méthode des moindres carrés ordinaire (OLS) et la sélection de sous ensembles de variables par algorithmes génétiques (GA-VSS) (Genetic Algorithms-Variable Subset Selection) [III.12].

Une attention particulière a été accordée à la colinéarité des descripteurs moléculaires sélectionnés en appliquant la règle QUIK (Q Under Influence of K) [III.13] une condition

nécessaire pour la validité du modèle. Les modèles acceptables sont uniquement ceux avec une corrélation globale du bloc [X + Y] (K_{xy}) supérieure à la corrélation globale de la variable X (K_{xx}), X représentant les variables explicatives et Y la variable à expliquer : $K_{xy} - K_{xx} \geq 0,05$.

La sélection par algorithme génétique conduit à un modèle MLR à quatre descripteurs qui décrit au mieux le point d'éclair. Le modèle retenu a pour équation :

$$\begin{aligned} \text{Tec} = & - 234,55(\pm 36,17) + 12,48(\pm 0,33) \text{ nsK} + 416,04(\pm 38,37) \text{ FDI} \\ & - 83,29(\pm 7,87) \text{ Mor26v} + 19,43(\pm 2,11) \text{ R5u} \end{aligned} \quad (\text{III.1})$$

Où,

- nsK est un descripteur constitutionnel appartenant au bloc numéro 1, représentant le nombre d'atomes autres que l'hydrogène [III.10]. De façon générale les descripteurs constitutionnels sont les descripteurs les plus simples et les plus couramment utilisés. Ils reflètent la composition moléculaire d'un composé sans donner aucune information sur sa géométrie moléculaire.

- FDI (Folding Degree Index en anglais) est l'indice de degré de repliement de la molécule. FDI est un descripteur géométrique calculé à l'aide du logiciel Dragon (bloc 12).

Les descripteurs géométriques sont définis de différentes manières, et sont toujours dérivés de la structure tridimensionnelle de la molécule. Généralement, les descripteurs géométriques sont calculés soit à partir de la géométrie moléculaire optimisée obtenue par les méthodes de la chimie computationnelle soit à partir des coordonnées cristallographiques.

Cet indice tend vers un (1) pour les molécules linéaires (de longueur infinie) et diminue en concordance avec le repliement de la molécule. Ainsi, il peut être considéré comme une mesure du degré de repliement de la molécule car il indique le degré de l'écart d'une molécule de la stricte linéarité [III.10].

- Mor26v est un descripteur Morse-3D du bloc 14; représentation tridimensionnelle des structures moléculaires, basée sur la diffraction des électrons. Ces descripteurs sont basés sur l'idée d'obtenir des informations à partir des coordonnées atomiques 3D par la transformée utilisée dans les études de diffraction d'électrons pour préparer des courbes de diffusion théoriques.

$$\text{Morsw} = \sum_{i=1}^{n\text{AT}-1} \sum_{j=i+1}^{n\text{AT}} W_i W_j \frac{\sin(s.r_{ij})}{s.r_{ij}} \quad (\text{III.2})$$

Où Morsw est l'intensité des électrons dispersés, w est une propriété atomique, les r_{ij} sont les distances interatomiques et $n\text{AT}$ le nombre d'atomes. Le terme s représente la dispersion dans diverses directions par une collection de n atomes.

Afin d'obtenir des descripteurs tout à fait uniformes, la répartition de l'intensité est rendue discrète, en calculant sa valeur comme une séquence de valeurs régulièrement distribuées.

Dans le logiciel DRAGON en particulier, ce descripteur est supposé prendre des valeurs entières de 0 à 31 (pour $s=0$ le rapport de dispersion est posé égal à 1) [III.10].

- R5u (R autocorrelation of lag 5 / unweighted) est l'autocorrélation R de décalage 5 / non pondéré ; c'est un descripteur GETAWAY du bloc 16. Les descripteurs GETAWAY ont été proposés comme descripteurs de structure chimique dérivés d'une nouvelle représentation de la structure moléculaire [III.10].

Ce descripteur est basé sur les formules d'auto-corrélation spatiale, pondérant les atomes de la molécule par des propriétés physico-chimiques w avec des informations 3D codées par les éléments de la matrice d'influence moléculaire H et de la matrice d'influence / distance R.

Les caractéristiques des descripteurs sélectionnées sont rassemblées dans le tableau III.2

Tableau III.2: Présentation des caractéristiques des descripteurs sélectionnées

Prédicteur	Coef	SE Coef	T	P	VIF
Constante	-234,55	36,17	-6,49	0,000	-
nsk	12,4845	0,3336	37,42	0,000	1,993
FDI	416,04	38,37	10,48	0,000	1,622
Mor26v	-83,290	7,873	-10,58	0,000	1,188
R3u	19,437	2,115	9,19	0,000	1,495

La valeur t d'un descripteur mesure la signification statistique de son coefficient de régression. Les valeurs absolues élevées de t indiquées dans le tableau III.2 expriment que les coefficients de régression des descripteurs impliqués dans le modèle RLM sont significativement plus grands

que l'écart-type. Remarquons ici que le descripteur représentant le nombre d'atomes autres que l'hydrogène (nSK) est le plus significatif comparativement aux autres descripteurs.

La probabilité de t d'un descripteur donne sa signification statistique lorsqu'il est combiné avec d'autres descripteurs dans un modèle QSPR global (les interactions entre descripteurs). Les descripteurs avec des valeurs de la probabilité de t inférieures à 0,05 (confiance de 95%) sont généralement considérés comme statistiquement significatifs dans un modèle particulier, ce qui montre que leur influence sur la variable réponse n'est pas due au hasard [III.14].

Des modèles ayant des descripteurs avec $VIF > 5$ ne seraient pas acceptables. Les valeurs des VIF suggèrent que ces descripteurs sont faiblement corrélés les uns avec les autres. Ainsi, le modèle peut être considéré comme une équation de régression optimale.

La matrice de corrélation établit que ces descripteurs sont faiblement corrélés 2 à 2. (Tableau III.3).

Tableau III.3: Matrice de corrélation entre les descripteurs et Tec.

	nSK	FDI	Mor26v	R5u	Tec
nSK	1,000				
FDI	0,503	1,000			
Mor26v	-0,234	-0,394	1,000		
R5u	0,525	0,061	-0,075	1,000	
Tec	0,950	0,634	-0,419	0,565	1,000

Les paramètres statistiques fournis par le logiciel MOBYDIGS [III.11] sont rapportés dans le tableau III.4 (page suivante).

Tableau III.4: Les paramètres statistiques du modèle développé.

n_{tr}	n_{ext}	Q_{LOO}^2 (%)	R^2 (%)	R_{adj}^2 (%)	Q_{ext}^2	Q_{boot}^2
139	34	97,11	97,41	97,34	97,71	96,96
F	SDEC	SDEP	$SDEP_{ext}$	Kxy	Kxx	s
1261,62	10,09	10,66	9,50	49,77	34,57	10,28

Les paramètres statistiques obtenus montrent que le modèle (d'équation III.1) établit une forte corrélation entre les variables sélectionnées et la propriété étudiée, caractérisée par un excellent coefficient de détermination. La valeur de $R^2 = 97,41\%$ indique que 97,41 de la variation totale est expliquée par le modèle.

Le coefficient de détermination ajusté R_{adj}^2 tient compte du nombre de variables et au contraire de R^2 , il n'augmente pas forcément lors de l'introduction de variables supplémentaires dans le modèle. La valeur élevée de R_{adj}^2 (%) = 97,34, très peu différente de celle de R^2 , indique un excellent accord entre la corrélation et la variation des données.

La grande valeur du F de Fisher ($F = 1261,62$), indique une excellente capacité prédictive du modèle, avec une erreur standard ($s = 10,28$).

La petite différence entre R^2 et Q_{LOO}^2 (0,30) informe sur la robustesse du modèle.

La valeur élevée de Q_{boot}^2 (%) = 96,96 confirme à la fois la prédictivité interne et la stabilité du modèle.

Une comparaison visuelle des résultats prédits avec les données expérimentales est également illustrée sur le graphe des valeurs prédites en fonction des valeurs expérimentales des points d'éclair (Figure III.1) pour les ensembles de calibrage et de test, confirme que le modèle linéaire a un très bon ajustement et peut être utilisé pour prédire la propriété étudiée.

Le graphe de la figure III.1, fait ressortir une faible dispersion autour de la première bissectrice et vérifie le bon ajustement du modèle obtenu.

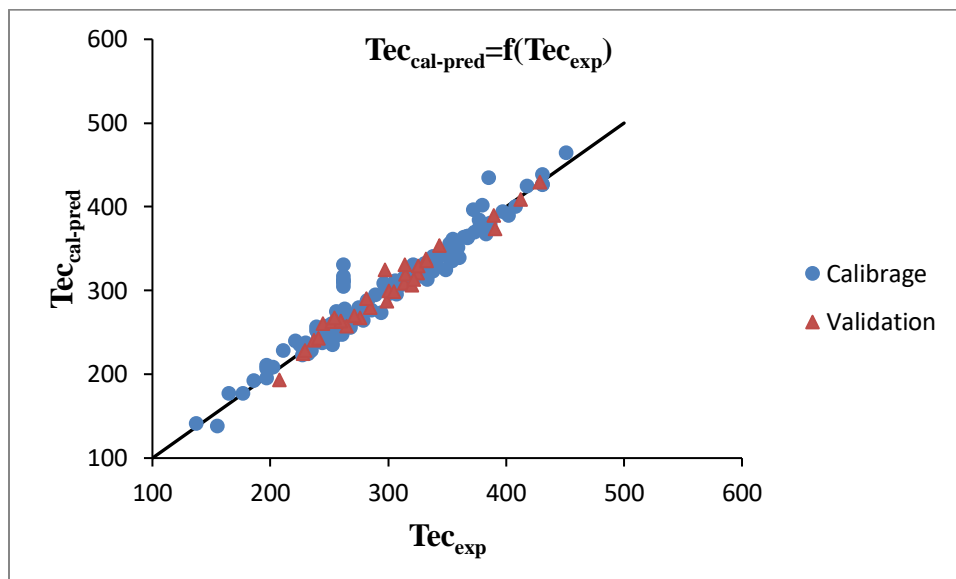


Figure III.1: Droite d'ajustement des Tec prédites en fonction des Tec expérimentales pour les ensembles de calibrage et de test.

III.1.3.2. Analyse des résidus

Les résultats obtenus sont condensés dans le tableau III.7, dont les colonnes sont numérotées de (1) à (9). La première rassemble les résidus ordinaires e_i , Notons également que 34 résidus ordinaires dépassent en valeur absolue l'erreur standard, $S = 10,28$, représentant un pourcentage de 24,46 % des composés de calibrage. Remarquons que 9 valeurs absolues des résidus ordinaires sont importantes. Ce sont dans l'ordre décroissant: e_{43} , e_{42} , e_{30} , e_{38} , e_9 , e_{130} , e_{126} , e_{64} et e_{52} . Les noms des composés correspondant à ces résidus sont respectivement : Tridécylbenzène, 1,2,3-tritert-butylbenzène, 1,2,3- triméthylbenzène, Décylbenzène, Naphtalène, Cis-oct-4-ène, 2,3,3- triméthylbut-1-ène et 3-méthylbut-1-ène. Notons aussi que le composé numéro 43 (Tridécylbenzène), a une valeur de résidu ordinaire, en valeur absolue, ($|e_i| = 50,1626$) supérieure à 3 fois l'erreur standard ($|e_i| < 3S$), soit $3 \times 10,28 = 30,84$.

Tous les résidus standardisés d_i de la colonne (2) sont compris entre les limites ± 3 , à l'exception du point 43(Tridécylbenzène) déjà signalé, qui est un point aberrant.

La colonne (3) rassemble les résidus studentisés internes r_i qui sont du même ordre de grandeur que les d_i correspondants. On a ici $p = 5$ et $n = 139$, et on constate que tous les r_i exceptés r_9 , r_{30} , r_{38} , r_{42} , r_{43} , r_{126} , et r_{130} sont inférieurs en valeur absolue à $t_{(0,025;n-p)} [= 1,960]$ qui est le 0,975 quantile d'une loi de Student avec $(n-p)$ degrés de liberté avec $(|r_{43}|=5,1145)$.

La colonne (4) donne les valeurs de h_{ii} , $i^{\text{ème}}$ terme diagonal de la matrice de projection (ou matrice chapeau) : $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ où \mathbf{X} est la matrice des valeurs observées des variables explicatives et \mathbf{X}' sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques.

Nous observons que les observations 28(Hexaméthylbenzène), 41(1,2,4,5-Tétraisopropylbenzène), 42(1,3,5-Tri-tert-butylbenzène), 44(2-Méthylanthracène), 45(9-Méthylanthracène), 47(7-Isopropyl-1-méthylphénanthrène), 60(Acétylène), 90(Éthylène) ont les plus grandes valeurs de h_{ii} ($h_{ii} > h^*$), avec $h^* = 3(k+1)/n_{tr} = 3(4+1)/139 = 0,1079$. Ces points sont des points influents.

La colonne (5) contient les résidus prédits, qui sont du même ordre de grandeur que les résidus ordinaires correspondants.

La colonne (6) montre le calcul de la somme des carrés des erreurs de prédiction (statistique PRESS) obtenue pour ce modèle.

La colonne (7) condense les estimations $S_{(i)}^2$ de σ^2 calculées selon l'équation (II.83).

$S_{(i)}^2$ intervient dans le calcul des résidus studentisés externes, rassemblés dans la colonne (8); tous les t_i sont du même ordre de grandeur que les r_i correspondants.

Comme les t_i sont inférieurs en valeur absolue à $t_{(0,025;n-p-1)} [=1,960]$, à l'exception encore une fois ceux des points 9, 30, 38, 42, 43, 126 et 130 dont les valeurs des résidus studentisés externes t_i supérieures en valeurs absolues à $t_{(\alpha;n-p-1)}$. L'analyse des résidus studentisés internes et externes permet de détecter l'observation 43 (Tridécylbenzène) comme observation aberrante.

La valeur du PRESS =15812,7944, et la valeur de SCE = 14518,6559 obtenue pour ce modèle

Enfin, la somme des carrés totale désignée par SCT(=547286,604) et la somme des carrés des résidus SCE conduit à une valeur élevée de coefficient de détermination R^2 égal à :

$$R^2 = 1 - \left(\frac{\text{SCE}}{\text{SCT}} \right) = 1 - \frac{14518,6559}{547286,604} = 97,34\%$$

et

$$R^2_{\text{préd}} = 1 - \left(\frac{\text{PRESS}}{\text{SCT}} \right) = 1 - \frac{15812,7944}{547286,604} = 97,11\%$$

Le modèle permet d'expliquer environ 99,34 % de la variabilité des nouvelles observations estimées.

Tableau III.5: Résidus caractéristiques et valeurs estimées des points d'éclair.

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Tec estimé
1	-1,8692	-0,1857	-0,1857	0,0407	-1,9486	3,7969	106,4107	-0,1850	263,8692
2	-9,5254	-0,9454	-0,9454	0,0391	-9,9131	98,2703	105,7281	-0,9450	312,5254
3	-6,5364	-0,6389	-0,6389	0,0094	-6,5983	43,5379	106,1138	-0,6375	309,5364
4	-1,4771	-0,1446	-0,1446	0,0120	-1,4951	2,2352	106,4215	-0,1441	305,4771
5	-1,2074	-0,1181	-0,1181	0,0102	-1,2199	1,4881	106,4270	-0,1176	312,2074
6	-7,1237	-0,7159	-0,7159	0,0627	-7,6006	57,7689	106,0310	-0,7146	331,1237
7	-0,1481	-0,0145	-0,0145	0,0193	-0,1510	0,0228	106,4379	-0,0145	317,1481
8	0,2038	0,0200	0,0200	0,0203	0,2080	0,0433	106,4378	0,0200	311,7962
9	20,3222	2,0240	2,0240	0,0457	21,2960	453,5182	103,1841	2,0480	339,6778
10	6,3554	0,6214	0,6214	0,0099	6,4189	41,2020	106,1314	0,6200	324,6446
11	3,6794	0,3649	0,3649	0,0375	3,8229	14,6142	106,3324	0,3637	342,3206
12	2,0313	0,1991	0,1991	0,0146	2,0613	4,2489	106,4066	0,1984	326,9687
13	-0,8620	-0,0843	-0,0843	0,0100	-0,8708	0,7583	106,4325	-0,0840	320,8620
14	-1,4369	-0,1416	-0,1416	0,0258	-1,4750	2,1755	106,4222	-0,1411	323,4369
15	0,5221	0,0511	0,0511	0,0104	0,5276	0,2783	106,4360	0,0509	323,4779

Tableau III.5 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Teb estimé
16	-0,1285	-0,0126	-0,0126	0,0144	-0,1304	0,0170	106,4380	-0,0125	328,1285
17	-1,8027	-0,1770	-0,1770	0,0187	-1,8370	3,3746	106,4132	-0,1764	332,8027
18	-6,5725	-0,6509	-0,6509	0,0349	-6,8100	46,3761	106,1016	-0,6495	361,4725
19	0,6179	0,0604	0,0604	0,0109	0,6247	0,3903	106,4352	0,0602	338,3821
20	4,2694	0,4178	0,4178	0,0116	4,3196	18,6588	106,2995	0,4165	330,7306
21	-0,0857	-0,0087	-0,0087	0,0919	-0,0943	0,0089	106,4381	-0,0087	364,0857
22	-10,1452	-0,9933	-0,9933	0,0126	-10,2748	105,5719	105,6544	-0,9933	331,1452
23	3,9523	0,3937	0,3937	0,0461	4,1434	17,1680	106,3150	0,3925	334,0477
24	1,3104	0,1303	0,1303	0,0423	1,3683	1,8724	106,4246	0,1298	378,6896
25	1,9186	0,1906	0,1906	0,0411	2,0009	4,0036	106,4092	0,1899	375,0814
26	8,1820	0,8132	0,8132	0,0417	8,5380	72,8978	105,9129	0,8121	373,8180
27	5,5959	0,5484	0,5484	0,0144	5,6778	32,2372	106,1992	0,5470	350,4041
28	-7,5665	-0,8013	-0,8013	0,1560	-8,9649	80,3694	105,9281	-0,8002	384,5665
29	10,8595	1,0656	1,0656	0,0169	11,0463	122,0218	105,5362	1,0661	346,1405

Tableau III.5 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Teb estimé
30	24,0694	2,3715	2,3715	0,0249	24,6840	609,2991	101,9710	2,4138	324,9306
31	18,2343	1,7994	1,7994	0,0280	18,7598	351,9283	103,8662	1,8148	335,7657
32	8,5118	0,8364	0,8364	0,0196	8,6820	75,3771	105,8825	0,8354	341,4882
33	2,4762	0,2433	0,2433	0,0192	2,5247	6,3742	106,3911	0,2424	365,5238
34	3,5022	0,3547	0,3547	0,0774	3,7961	14,4104	106,3382	0,3536	363,4978
35	2,7430	0,2711	0,2711	0,0308	2,8303	8,0107	106,3797	0,2701	370,2570
36	0,9525	0,0939	0,0939	0,0263	0,9783	0,9570	106,4311	0,0936	379,0475
37	7,3788	0,7495	0,7495	0,0825	8,0427	64,6847	105,9919	0,7483	351,6212
38	-22,0538	-2,1943	-2,1943	0,0439	-23,0652	532,0050	102,6135	-2,2265	402,0538
39	4,8036	0,4929	0,4929	0,1010	5,3432	28,5497	106,2451	0,4915	381,1964
40	-6,8102	-0,6873	-0,6873	0,0705	-7,3270	53,6851	106,0629	-0,6859	424,8102
41	2,0422	0,2160	0,2160	0,1540	2,4141	5,8280	106,4010	0,2153	394,9578
42	-25,1977	-2,6337	-2,6337	0,1335	-29,0815	845,7327	100,9284	-2,6945	397,1977
43	-50,1626	-5,1145	-5,1145	0,0894	-55,0897	3034,8747	85,6603	-5,6798	435,1626
44	4,1065	0,4357	0,4357	0,1590	4,8829	23,8423	106,2874	0,4343	426,8935

Tableau III.5 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Teb estimé
45	-7,6082	-0,7980	-0,7980	0,1396	-8,8424	78,1885	105,9323	-0,7969	438,6082
46	3,6447	0,3738	0,3738	0,1001	4,0501	16,4035	106,3271	0,3726	427,3553
47	-13,7288	-1,4152	-1,4152	0,1092	-15,4121	237,5318	104,8472	-1,4206	464,7288
48	0,8519	0,0843	0,0843	0,0341	0,8820	0,7778	106,4325	0,0840	302,1481
49	5,9917	0,5893	0,5893	0,0213	6,1221	37,4805	106,1623	0,5878	298,0083
50	8,4521	0,8290	0,8290	0,0161	8,5901	73,7892	105,8922	0,8280	315,5479
51	3,5028	0,3426	0,3426	0,0107	3,5405	12,5354	106,3449	0,3415	306,4972
52	19,0173	1,8706	1,8706	0,0217	19,4383	377,8473	103,6587	1,8884	313,9827
53	-2,3658	-0,2316	-0,2316	0,0123	-2,3951	5,7366	106,3955	-0,2308	315,3658
54	14,1771	1,3922	1,3922	0,0185	14,4439	208,6261	104,8985	1,3972	323,8229
55	8,0519	0,7903	0,7903	0,0174	8,1943	67,1472	105,9420	0,7892	328,9482
56	12,0977	1,2021	1,2021	0,0413	12,6188	159,2329	105,2903	1,2041	328,9023
57	8,2976	0,8114	0,8114	0,0101	8,3822	70,2607	105,9152	0,8104	324,7024
58	5,8406	0,5715	0,5715	0,0114	5,9081	34,9052	106,1787	0,5701	329,1594
59	3,4116	0,3338	0,3338	0,0115	3,4512	11,9105	106,3496	0,3327	329,5884

Tableau III.5 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Teb estimé
60	16,1885	1,7924	1,7924	0,2278	20,9646	439,5134	103,8864	1,8075	138,8115
61	-6,9008	-0,6870	-0,6870	0,0449	-7,2252	52,2041	106,0632	-0,6856	192,9008
62	-7,8567	-0,7732	-0,7732	0,0225	-8,0378	64,6064	105,9633	-0,7720	237,8567
63	6,7490	0,6640	0,6640	0,0220	6,9006	47,6188	106,0879	0,6626	246,2510
64	-19,2253	-1,8913	-1,8913	0,0219	-19,6559	386,3555	103,5968	-1,9099	240,2253
65	-9,1590	-0,8978	-0,8978	0,0148	-9,2969	86,4320	105,7979	-0,8971	261,1590
66	-0,4845	-0,0476	-0,0476	0,0173	-0,4930	0,2431	106,4363	-0,0474	259,4845
67	-18,3169	-1,8009	-1,8009	0,0208	-18,7055	349,8973	103,8620	-1,8163	257,3170
68	2,0408	0,2005	0,2005	0,0198	2,0820	4,3348	106,4062	0,1998	246,9592
69	-6,2793	-0,6200	-0,6200	0,0290	-6,4669	41,8209	106,1328	-0,6186	288,2793
70	-15,7050	-1,5376	-1,5376	0,0125	-15,9037	252,9265	104,5602	-1,5456	287,7050
71	-4,7456	-0,4655	-0,4655	0,0163	-4,8244	23,2752	106,2660	-0,4642	279,7456
72	-15,7791	-1,5504	-1,5504	0,0195	-16,0934	258,9960	104,5288	-1,5586	272,7791
73	-12,6626	-1,2487	-1,2487	0,0267	-13,0097	169,2523	105,1995	-1,2514	308,6626
74	11,2757	1,1063	1,1063	0,0166	11,4664	131,4789	105,4660	1,1072	295,7243

Tableau III.5 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Teb estimé
75	-6,6568	-0,6516	-0,6516	0,0122	-6,7392	45,4166	106,1008	-0,6502	295,6568
76	4,4493	0,4366	0,4366	0,0168	4,5254	20,4793	106,2867	0,4352	296,5507
77	-5,9339	-0,5813	-0,5813	0,0135	-6,0150	36,1806	106,1697	-0,5798	311,9339
78	-3,2373	-0,3175	-0,3175	0,0160	-3,2901	10,8246	106,3580	-0,3165	326,2374
79	-3,3829	-0,3321	-0,3321	0,0180	-3,4447	11,8662	106,3505	-0,3310	341,3829
80	11,0501	1,0829	1,0829	0,0143	11,2104	125,6728	105,5067	1,0836	329,9500
81	-0,9069	-0,0891	-0,0891	0,0200	-0,9254	0,8563	106,4318	-0,0888	352,9069
82	1,0361	0,1020	0,1020	0,0236	1,0612	1,1261	106,4298	0,1016	364,9639
83	-7,2842	-0,7276	-0,7276	0,0514	-7,6787	58,9628	106,0176	-0,7263	209,2842
84	5,6639	0,5645	0,5645	0,0472	5,9446	35,3380	106,1850	0,5631	238,3361
85	3,3629	0,3301	0,3301	0,0175	3,4227	11,7151	106,3516	0,3290	252,6371
86	-3,3158	-0,3288	-0,3288	0,0376	-3,4452	11,8693	106,3522	-0,3277	246,3158
87	6,6622	0,6561	0,6561	0,0241	6,8267	46,6035	106,0962	0,6547	260,3378
88	5,4497	0,5329	0,5329	0,0099	5,5043	30,2975	106,2126	0,5314	266,5503
89	9,1731	0,9031	0,9031	0,0235	9,3934	88,2365	105,7902	0,9025	276,8269

Tableau III.5 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Teb estimé
90	-5,0489	-0,5288	-0,5288	0,1372	-5,8516	34,2413	106,2160	-0,5274	142,0489
91	-12,2513	-1,2289	-1,2289	0,0592	-13,0223	169,5797	105,2386	-1,2313	177,2513
92	-0,5775	-0,0576	-0,0576	0,0498	-0,6078	0,3694	106,4355	-0,0574	177,5775
93	-14,3372	-1,4152	-1,4152	0,0284	-14,7566	217,7569	104,8474	-1,4205	211,3372
94	-10,3935	-1,0262	-1,0262	0,0290	-10,7035	114,5648	105,6017	-1,0264	207,3935
95	-8,5713	-0,8512	-0,8512	0,0402	-8,9302	79,7486	105,8626	-0,8503	208,5713
96	0,7741	0,0771	0,0771	0,0470	0,8122	0,6597	106,4334	0,0769	196,2259
97	1,8263	0,1799	0,1799	0,0245	1,8722	3,5050	106,4124	0,1793	231,1737
98	5,9481	0,5850	0,5850	0,0212	6,0772	36,9327	106,1663	0,5835	229,0519
99	7,1963	0,7074	0,7074	0,0205	7,3471	53,9802	106,0406	0,7061	224,8037
100	17,1479	1,6816	1,6816	0,0157	17,4216	303,5114	104,1919	1,6933	235,8521
101	3,3363	0,3293	0,3293	0,0285	3,4342	11,7934	106,3520	0,3282	223,6637
102	-10,9159	-1,0705	-1,0705	0,0157	-11,0898	122,9839	105,5279	-1,0710	235,9159
103	-17,8682	-1,7550	-1,7550	0,0188	-18,2106	331,6244	103,9916	-1,7689	228,8682
104	-9,4717	-0,9284	-0,9284	0,0147	-9,6128	92,4052	105,7535	-0,9279	257,4718

Tableau III.5 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Teb estimé
105	-5,3066	-0,5205	-0,5205	0,0160	-5,3926	29,0806	106,2230	-0,5190	251,3066
106	-8,1357	-0,8062	-0,8062	0,0360	-8,4397	71,2292	105,9219	-0,8051	259,1357
107	-13,5737	-1,3314	-1,3314	0,0161	-13,7962	190,3341	105,0301	-1,3353	252,5737
108	9,6576	0,9466	0,9466	0,0148	9,8025	96,0889	105,7263	0,9463	245,3424
109	1,5131	0,1482	0,1482	0,0137	1,5341	2,3534	106,4207	0,1477	251,4869
110	14,5113	1,4273	1,4273	0,0216	14,8314	219,9712	104,8199	1,4329	237,4887
111	13,1972	1,2925	1,2925	0,0131	13,3724	178,8212	105,1112	1,2958	247,8028
112	5,9782	0,5954	0,5954	0,0458	6,2650	39,2500	106,1565	0,5940	249,0218
113	-7,5749	-0,7641	-0,7641	0,0698	-8,1434	66,3153	105,9743	-0,7629	263,5749
114	3,4471	0,3388	0,3388	0,0200	3,5174	12,3718	106,3470	0,3377	240,5529
115	-5,9648	-0,5847	-0,5847	0,0148	-6,0546	36,6578	106,1666	-0,5832	246,9648
116	-3,1441	-0,3084	-0,3084	0,0160	-3,1954	10,2104	106,3626	-0,3073	249,1441
117	-9,8320	-0,9628	-0,9628	0,0129	-9,9601	99,2040	105,7018	-0,9625	250,8320
118	-6,0209	-0,5936	-0,5936	0,0262	-6,1828	38,2273	106,1582	-0,5922	250,0210
119	8,2320	0,8190	0,8190	0,0437	8,6081	74,1000	105,9053	0,8180	257,7680

Tableau III.5 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Teb estimé
120	-5,8272	-0,5718	-0,5718	0,0169	-5,9273	35,1325	106,1784	-0,5703	268,8272
121	-5,0238	-0,4924	-0,4924	0,0147	-5,0987	25,9966	106,2455	-0,4910	269,0238
122	4,9962	0,4893	0,4893	0,0129	5,0615	25,6191	106,2480	0,4879	260,0039
123	-3,1141	-0,3046	-0,3046	0,0104	-3,1469	9,9031	106,3644	-0,3035	270,1141
124	1,3996	0,1370	0,1370	0,0117	1,4162	2,0056	106,4232	0,1365	264,6004
125	4,9718	0,4866	0,4866	0,0119	5,0318	25,3195	106,2500	0,4852	262,0282
126	-19,5510	-1,9654	-1,9654	0,0633	-20,8720	435,6422	103,3699	-1,9869	275,5510
127	11,4414	1,1210	1,1210	0,0140	11,6039	134,6501	105,4399	1,1221	256,5586
128	-2,6118	-0,2553	-0,2553	0,0096	-2,6371	6,9545	106,3863	-0,2544	284,6118
129	1,3256	0,1299	0,1299	0,0138	1,3441	1,8067	106,4247	0,1294	279,6744
130	20,0961	1,9667	1,9667	0,0117	20,3332	413,4410	103,3658	1,9882	273,9040
131	13,9169	1,3735	1,3735	0,0281	14,3199	205,0597	104,9397	1,3781	265,0831
132	-4,0792	-0,4013	-0,4013	0,0218	-4,1700	17,3887	106,3102	-0,4000	302,0792
133	6,4498	0,6394	0,6394	0,0368	6,6964	44,8416	106,1134	0,6380	344,5502
134	2,0596	0,2032	0,2032	0,0275	2,1178	4,4851	106,4053	0,2025	342,9404

Tableau III.5 (Suite et fin)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	Teb estimé
135	-4,2852	-0,4261	-0,4261	0,0427	-4,4764	20,0383	106,2939	-0,4248	356,2852
136	15,3324	1,5318	1,5318	0,0517	16,1683	261,4143	104,5742	1,5397	367,6676
137	6,7779	0,6809	0,6809	0,0621	7,2269	52,2279	106,0698	0,6796	379,2221
138	12,0563	1,2215	1,2215	0,0778	13,0733	170,9105	105,2530	1,2237	389,9438
139	7,0598	0,7219	0,7219	0,0948	7,7995	60,8314	106,0241	0,7207	400,9402

III.1.3.3. Diagnostics d'influence

Le tableau III.8 condense d'autres mesures d'influence utiles, [III.15] dont l'étude complète la recherche des observations aberrantes.

Tableau III.6 : Diagnostics d'influence.

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
1	16,9951	0,0003	-0,0381	0,0194	0,0112	-0,0211	-0,0003	0,0162	1,0808
2	3,4154	0,0073	-0,1907	0,0825	0,1296	-0,0837	0,0285	-0,1576	1,0449
3	0,4117	0,0008	-0,0621	0,0193	0,0120	-0,0209	0,0129	-0,0017	1,0321
4	0,4073	0,0001	-0,0159	0,0042	0,0054	-0,0043	-0,0039	-0,0094	1,0500
5	0,4964	0,0000	-0,0120	0,0038	0,0034	-0,0041	0,0034	-0,0024	1,0484
6	4,6482	0,0069	-0,1849	0,0583	0,1065	-0,0572	0,0207	-0,1706	1,0866
7	2,2847	0,0000	-0,0020	0,0008	0,0004	-0,0009	0,0008	0,0004	1,0586
8	1,2308	0,0000	0,0029	-0,0002	-0,0012	0,0002	-0,0003	0,0023	1,0597
9	8,0330	0,0393	0,4483	-0,0313	0,0363	0,0397	-0,3187	-0,1668	0,9315
10	2,2076	0,0008	0,0620	-0,0074	0,0051	0,0082	-0,0220	-0,0045	1,0335
11	4,8887	0,0010	0,0718	-0,0209	-0,0332	0,0203	-0,0162	0,0603	1,0733
12	2,5506	0,0001	0,0241	0,0004	-0,0041	-0,0006	-0,0069	0,0145	1,0519
13	2,0043	0,0000	-0,0085	0,0012	0,0007	-0,0012	-0,0008	-0,0040	1,0484
14	3,7715	0,0001	-0,0230	-0,0101	0,0000	0,0102	0,0104	-0,0115	1,0648

Tableau III.6 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
15	2,0617	0,0000	0,0052	0,0009	0,0004	-0,0009	-0,0023	0,0010	1,0490
16	2,5685	0,0000	-0,0015	-0,0002	0,0001	0,0002	0,0008	-0,0005	1,0533
17	3,0759	0,0001	-0,0243	0,0021	0,0067	-0,0020	0,0093	-0,0157	1,0567
18	11,5349	0,0031	-0,1235	0,0243	0,0103	-0,0248	0,0888	-0,0112	1,0588
19	6,9316	0,0000	0,0063	-0,0012	0,0017	0,0012	-0,0010	-0,0009	1,0495
20	6,4010	0,0004	0,0451	0,0184	0,0230	-0,0186	-0,0115	-0,0057	1,0435
21	15,1763	0,0000	-0,0028	0,0005	0,0011	-0,0005	0,0003	-0,0026	1,1432
22	6,8301	0,0025	-0,1123	-0,0195	-0,0273	0,0213	-0,0115	-0,0396	1,0133
23	11,3162	0,0015	0,0863	0,0192	-0,0053	-0,0210	0,0039	0,0639	1,0821
24	20,0198	0,0001	0,0273	-0,0017	0,0003	0,0016	-0,0211	0,0040	1,0833
25	19,8971	0,0003	0,0393	-0,0020	0,0047	0,0020	-0,0300	-0,0030	1,0812
26	19,8776	0,0058	0,1694	-0,0055	0,0262	0,0057	-0,1289	-0,0238	1,0569
27	14,6288	0,0009	0,0662	-0,0126	0,0310	0,0119	0,0019	-0,0191	1,0416

Tableau III.6 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
28	29,8832	0,0237	-0,3440	0,0510	0,1293	-0,0433	0,0266	-0,3261	1,2008
29	14,1033	0,0039	0,1399	0,0639	0,0904	-0,0657	-0,0487	-0,0350	1,0120
30	1,8058	0,0287	0,3857	-0,1545	-0,2113	0,1569	-0,1160	0,2672	0,8592
31	16,1704	0,0187	0,3081	0,0575	0,0992	-0,0656	0,1429	0,1058	0,9451
32	14,9946	0,0028	0,1181	0,0027	0,0375	-0,0058	0,0532	0,0369	1,0316
33	25,5862	0,0002	0,0339	-0,0086	0,0175	0,0080	0,0047	-0,0089	1,0562
34	45,3292	0,0021	0,1024	0,0658	0,0461	-0,0689	-0,0087	0,0306	1,1200
35	39,3966	0,0005	0,0482	-0,0074	0,0234	0,0058	0,0272	0,0032	1,0683
36	39,5058	0,0000	0,0154	-0,0041	0,0088	0,0037	0,0036	-0,0044	1,0659
37	57,8372	0,0101	0,2244	0,0852	0,1776	-0,0904	0,1139	-0,0855	1,1080
38	75,3402	0,0442	-0,4768	0,0534	-0,3466	-0,0387	-0,1517	0,1651	0,9042
39	81,7583	0,0055	0,1647	0,1043	0,1027	-0,1099	0,0151	0,0210	1,1444
40	123,1815	0,0072	-0,1889	0,0034	-0,1507	0,0032	-0,0691	0,0701	1,0974
41	126,3788	0,0017	0,0919	0,0681	0,0801	-0,0707	-0,0002	-0,0298	1,2251

Tableau III.6 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
42	122,9788	0,2138	-1,0579	-0,6866	-0,9908	0,7142	-0,0523	0,5063	0,9185
43	151,8062	0,5139	-1,7801	-0,0006	-1,4316	0,0641	-0,7381	0,6468	0,3849
44	72,3418	0,0072	0,1889	0,0548	0,0724	-0,0553	-0,1611	-0,0654	1,2257
45	72,6413	0,0207	-0,3210	-0,0680	-0,0700	0,0708	0,2877	0,0113	1,1782
46	66,8494	0,0031	0,1243	0,0264	0,0412	-0,0276	-0,1047	-0,0195	1,1476
47	132,0974	0,0491	-0,4974	-0,1423	-0,2763	0,1535	0,3520	0,1062	1,0809
48	5,7006	0,0001	0,0158	-0,0053	-0,0020	0,0058	-0,0063	-0,0068	1,0745
49	3,5183	0,0015	0,0867	-0,0432	-0,0214	0,0464	-0,0144	-0,0251	1,0471
50	1,6784	0,0022	0,1058	-0,0409	-0,0193	0,0437	-0,0398	-0,0153	1,0283
51	0,5731	0,0003	0,0354	-0,0154	-0,0040	0,0164	0,0018	-0,0061	1,0448
52	2,5864	0,0155	0,2810	-0,0933	-0,0123	0,1014	-0,0992	-0,1098	0,9296
53	0,7372	0,0001	-0,0257	0,0105	0,0098	-0,0110	0,0071	-0,0084	1,0489
54	3,2896	0,0073	0,1917	-0,0008	0,0511	0,0047	-0,0889	-0,0990	0,9834
55	3,6246	0,0022	0,1050	-0,0254	0,0086	0,0274	-0,0446	-0,0337	1,0321

Tableau III.6 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
56	6,8627	0,0124	0,2499	-0,0248	0,0561	0,0303	-0,1184	-0,1543	1,0257
57	2,1325	0,0013	0,0818	0,0031	0,0068	-0,0023	-0,0376	0,0032	1,0232
58	2,4316	0,0008	0,0613	-0,0124	-0,0087	0,0127	-0,0229	0,0191	1,0374
59	2,4981	0,0003	0,0359	-0,0108	-0,0060	0,0110	-0,0112	0,0105	1,0459
60	103,2893	0,1896	0,9818	0,9107	0,3026	-0,8849	-0,3399	-0,4290	1,1908
61	57,4012	0,0044	-0,1487	-0,0638	0,0086	0,0552	0,0337	0,0888	1,0680
62	24,6273	0,0028	-0,1172	-0,0013	0,0313	-0,0058	0,0211	0,0556	1,0386
63	23,4888	0,0020	0,0993	-0,0478	-0,0678	0,0531	0,0113	0,0059	1,0441
64	24,4140	0,0160	-0,2858	-0,0559	0,0807	0,0388	0,1170	0,1031	0,9271
65	13,2492	0,0024	-0,1101	0,0193	0,0549	-0,0255	0,0274	0,0108	1,0225
66	13,6021	0,0000	-0,0063	0,0023	0,0029	-0,0026	0,0001	0,0013	1,0563
67	14,1134	0,0138	-0,2646	-0,0940	0,0425	0,0806	0,1563	0,0818	0,9380
68	14,4183	0,0002	0,0284	0,0111	0,0008	-0,0096	-0,0075	-0,0164	1,0576
69	8,5616	0,0023	-0,1069	0,0558	0,0431	-0,0601	0,0280	0,0201	1,0539

Tableau III.6 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
70	5,8274	0,0060	-0,1738	0,0692	0,0945	-0,0771	0,0169	-0,0160	0,9618
71	6,1612	0,0007	-0,0598	0,0363	0,0338	-0,0387	-0,0065	-0,0058	1,0469
72	7,0393	0,0096	-0,2200	0,0923	0,0513	-0,1023	-0,0241	0,0894	0,9672
73	4,3622	0,0085	-0,2072	0,1040	0,0839	-0,1104	0,0861	0,0027	1,0060
74	2,0225	0,0041	0,1440	-0,1016	-0,0748	0,1054	0,0376	0,0366	1,0084
75	1,7788	0,0011	-0,0723	0,0390	0,0336	-0,0415	0,0008	-0,0100	1,0345
76	2,1781	0,0007	0,0569	-0,0403	-0,0283	0,0419	0,0113	0,0101	1,0484
77	0,8706	0,0009	-0,0678	0,0443	0,0237	-0,0456	-0,0076	-0,0109	1,0392
78	2,9985	0,0003	-0,0404	0,0284	0,0077	-0,0287	-0,0109	-0,0046	1,0511
79	8,0128	0,0004	-0,0448	0,0304	0,0016	-0,0303	-0,0131	-0,0032	1,0528
80	6,8711	0,0034	0,1305	-0,0230	0,0585	0,0235	0,0241	-0,0621	1,0079
81	15,5958	0,0000	-0,0127	0,0074	-0,0022	-0,0072	-0,0049	0,0002	1,0590
82	26,1688	0,0001	0,0158	-0,0074	0,0054	0,0071	0,0068	-0,0012	1,0629
83	41,0898	0,0057	-0,1691	0,0570	0,0399	-0,0649	-0,0845	0,0756	1,0729

Tableau III.6 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
84	27,4502	0,0032	0,1253	-0,0577	-0,0330	0,0636	0,0302	-0,0604	1,0767
85	13,8613	0,0004	0,0439	-0,0109	-0,0131	0,0133	0,0065	-0,0168	1,0524
86	15,9147	0,0008	-0,0647	0,0317	0,0148	-0,0341	-0,0372	0,0219	1,0743
87	7,4007	0,0021	0,1029	-0,0318	-0,0340	0,0339	0,0776	0,0256	1,0468
88	5,9282	0,0006	0,0532	0,0029	-0,0180	-0,0007	0,0107	0,0075	1,0375
89	4,1635	0,0039	0,1399	0,0667	-0,0111	-0,0658	0,0118	0,0639	1,0311
90	93,4937	0,0089	-0,2103	-0,1804	-0,0545	0,1734	0,0334	0,1013	1,1907
91	61,0383	0,0190	-0,3089	-0,1942	-0,0398	0,1795	0,0191	0,1863	1,0427
92	59,7323	0,0000	-0,0132	-0,0041	0,0000	0,0034	-0,0040	0,0075	1,0925
93	38,5833	0,0117	-0,2429	0,0423	0,1097	-0,0544	-0,1279	0,0174	0,9910
94	39,5692	0,0063	-0,1773	-0,0574	0,0187	0,0470	-0,0078	0,0933	1,0278
95	41,9819	0,0061	-0,1740	-0,1197	0,0311	0,1134	0,0336	-0,0145	1,0527
96	42,5189	0,0001	0,0171	0,0085	0,0026	-0,0077	0,0005	-0,0117	1,0891
97	24,6677	0,0002	0,0284	-0,0091	-0,0115	0,0105	0,0141	-0,0052	1,0629
98	24,6239	0,0015	0,0860	-0,0138	-0,0373	0,0179	0,0447	-0,0012	1,0472

Tableau III.6 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
99	25,5283	0,0021	0,1022	0,0513	-0,0123	-0,0460	0,0008	-0,0274	1,0403
100	23,7038	0,0090	0,2139	-0,0009	-0,1165	0,0130	0,0268	0,0129	0,9480
101	27,0265	0,0006	0,0562	0,0394	-0,0044	-0,0372	-0,0082	-0,0035	1,0643
102	23,6991	0,0037	-0,1352	0,0002	0,0736	-0,0079	-0,0161	-0,0079	1,0104
103	25,0261	0,0118	-0,2449	-0,1135	0,0623	0,1010	0,0117	0,0189	0,9419
104	13,0688	0,0026	-0,1132	0,0437	0,0662	-0,0495	-0,0213	-0,0061	1,0202
105	13,4886	0,0009	-0,0661	0,0219	0,0292	-0,0252	-0,0266	0,0054	1,0444
106	15,4680	0,0049	-0,1556	-0,0167	0,0864	0,0143	0,0007	-0,1177	1,0511
107	13,7372	0,0058	-0,1709	-0,0213	0,0888	0,0151	-0,0176	-0,0754	0,9872
108	13,8786	0,0027	0,1159	-0,0047	-0,0354	0,0103	0,0501	-0,0131	1,0190
109	13,2856	0,0001	0,0174	-0,0037	-0,0086	0,0045	0,0059	0,0010	1,0517
110	15,9226	0,0090	0,2128	0,1381	0,0022	-0,1306	0,0115	-0,0227	0,9828
111	13,5876	0,0044	0,1493	0,0009	-0,0551	0,0066	0,0465	-0,0044	0,9880
112	18,1551	0,0034	0,1301	0,0307	-0,0524	-0,0294	0,0216	0,0892	1,0736
113	19,0298	0,0088	-0,2090	-0,0307	0,1058	0,0296	0,0148	-0,1730	1,0920

Tableau III.6 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
114	15,5567	0,0005	0,0482	0,0252	-0,0067	-0,0236	0,0066	0,0051	1,0548
115	14,1723	0,0010	-0,0715	-0,0289	0,0200	0,0259	-0,0013	-0,0084	1,0404
116	14,1356	0,0003	-0,0392	-0,0112	0,0157	0,0097	-0,0032	-0,0128	1,0514
117	13,5569	0,0024	-0,1099	-0,0355	0,0364	0,0299	0,0140	-0,0023	1,0158
118	15,2694	0,0019	-0,0971	-0,0148	0,0455	0,0128	-0,0212	-0,0581	1,0521
119	16,8134	0,0061	0,1749	0,0424	-0,0823	-0,0403	-0,0144	0,1274	1,0587
120	6,2425	0,0011	-0,0747	0,0385	0,0315	-0,0411	-0,0405	-0,0042	1,0432
121	6,0163	0,0007	-0,0600	0,0263	0,0281	-0,0284	-0,0307	-0,0112	1,0441
122	6,8401	0,0006	0,0558	0,0266	-0,0038	-0,0247	0,0056	0,0010	1,0424
123	5,7114	0,0002	-0,0311	0,0074	0,0144	-0,0088	-0,0080	-0,0048	1,0455
124	6,0877	0,0000	0,0149	-0,0027	-0,0039	0,0034	0,0063	-0,0018	1,0498
125	6,5202	0,0006	0,0533	0,0125	-0,0116	-0,0107	0,0149	0,0082	1,0415
126	11,9269	0,0522	-0,5165	-0,0932	0,2165	0,0931	0,0094	-0,4228	0,9575
127	7,0903	0,0036	0,1337	0,0702	0,0087	-0,0654	0,0193	-0,0248	1,0044
128	1,4929	0,0001	-0,0250	0,0080	0,0064	-0,0088	-0,0092	-0,0002	1,0457

Tableau III.6 (suite et fin)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	DFBETAS_{0,4}	DFBETAS_{0,5}	COVRATIO_i
129	1,9973	0,0000	0,0153	-0,0052	-0,0016	0,0057	0,0087	-0,0031	1,0520
131	5,6026	0,0109	0,2345	0,1348	0,0326	-0,1332	0,0702	0,0420	0,9951
132	1,2484	0,0007	-0,0597	0,0392	0,0159	-0,0398	-0,0395	-0,0118	1,0549
133	16,7010	0,0031	0,1247	-0,0738	0,0127	0,0722	0,0914	0,0076	1,0615
134	15,6617	0,0002	0,0340	-0,0163	0,0070	0,0158	0,0239	0,0009	1,0659
135	27,5204	0,0016	-0,0897	0,0451	-0,0211	-0,0434	-0,0667	0,0001	1,0772
136	41,4141	0,0256	0,3595	-0,1501	0,1233	0,1418	0,2683	-0,0195	1,0022
137	58,2384	0,0061	0,1749	-0,0597	0,0756	0,0551	0,1287	-0,0177	1,0879
138	78,3225	0,0252	0,3554	-0,0991	0,1753	0,0889	0,2610	-0,0483	1,0645
139	101,3290	0,0109	0,2333	-0,0525	0,1270	0,0454	0,1690	-0,0389	1,1248

Les résultats obtenus sont condensés dans le tableau III.6, dont les colonnes sont numérotées de (1) à (9). Les valeurs de la somme pondérée des carrés des distances du point i au barycentre des données ($WSSD_i$), données dans la colonne (1) du tableau III.6 montrent la présence de 7 observations parmi 139 composés qui sont les plus éloignées. Notons aussi que le composé le plus éloigné est l'observation 43 (Tridécylbenzène) ayant la valeur la plus élevée de $WSSD$, déjà signalé comme point aberrant.

La colonne (2) rassemble les valeurs de la distance de Cook, D_i . Les valeurs D_i des cinq observations (9, 38, 42, 43, 47, 60, 126) sont supérieures à la valeur critique ($4/139=0,0288$). Ces observations sont considérées comme très influentes. Une attention particulière doit être donnée à l'observation 43 dont la valeur de D_i est supérieure à la valeur critique ($D_{43}=0,5139 \gg 0,0288$). Le composé tridécylbenzène est un composé très influent.

Dans la colonne (3), huit observations (9, 30, 38, 42, 43, 47, 60, 126) ont des valeurs absolues de $DIFITS$, supérieures à la valeur critique $2\sqrt{5/139} = 0,3793$. Le composé 43 (tridécylbenzène) a une valeur très élevée de $DIFITS$ ($=1,7801 \gg 0,3793$). Ces observations ont une influence sur la valeur prédite.

Les colonnes (4), (5) et (6) rassemblent les valeurs de la statistique $DFBETAS_{j,i}$. L'examen des colonnes $DFBETAS_{j,i}$ en tenant compte de la valeur critique $\frac{2}{\sqrt{n}} = 2/\sqrt{139} = 0,1696$ fait ressortir l'observation 60 (Acétylène) comme point influent sur la régression dont toutes les valeurs absolues de $DFBETAS_{j,i}$ sont supérieures à la valeur critique.

La colonne (7) rassemble les valeurs du $COVRATIO$ pour 13 composés sur 139 composés ont des valeurs très proches de 1 à l'exception l'observation 43 qui a une valeur ($=0,3849$) inférieure à 1 et détériorent la précision de l'estimation.

III.1.3.4. Le test de randomisation

Le test de randomisation associé au modèle QSPR donné par la figure III.2, est la représentation graphique des coefficients statistiques permettant de comparer les résultats pour les modèles randomisés (signe +) au modèle de départ (triangle) qui est le modèle réel.

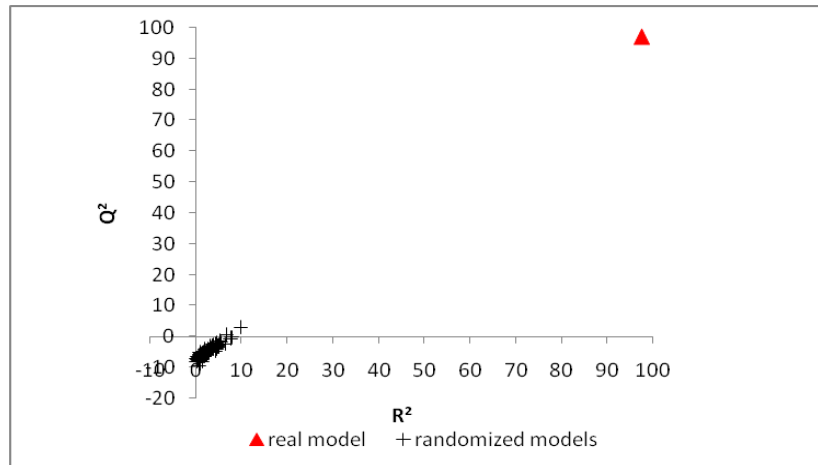


Figure III.2: Test de randomisation associé au modèle QSPR.

Il est clair que les statistiques des points d'éclair obtenues pour les vecteurs modifiés sont inférieures à celles du modèle QSPR réel, ce qui assure qu'une relation structure / propriété réelle (Tec) a été établie, et que le modèle proposé n'est pas aléatoire.

III.1.3.5. Étude de la contribution des descripteurs au modèle

En se basant sur une procédure précédente [II.16], la contribution relative des quatre descripteurs au modèle a été déterminée comme suit:

nsk (45,82%) > FDI (18,55%) > Mor26v (18,34) > R5u (17,29).

La figure III.3 est la représentation graphique de la contribution relative des descripteurs dans la détermination du modèle obtenu.

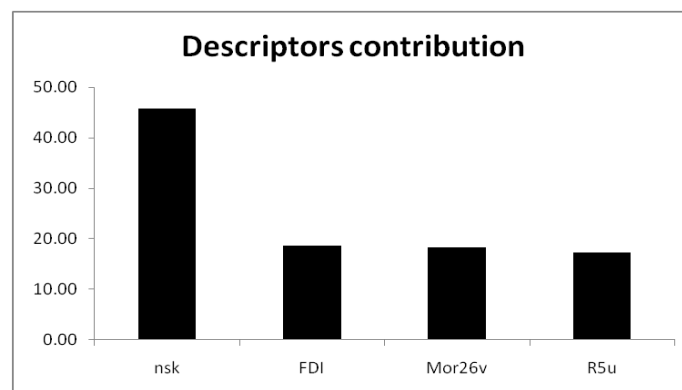


Figure III.3: Contributions relatives des descripteurs sélectionnés dans le modèle RLM.

Remarquons que, la contribution de nsk (le nombre d'atomes de Carbone) est supérieure aux contributions des descripteurs FDI, Mor26v et R5u, alors que la différence entre la contribution des trois derniers descripteurs n'est pas significative. Comme le fait ressortir la figure III.3, le descripteur nsk est le plus influent dans la génération du modèle prédictif et sa contribution est la plus élevée par comparaison aux autres descripteurs.

Les paramètres statistiques calculés selon Tropsha *et al.* [III.17] pour l'ensemble de test, vérifient les conditions générales pour le pouvoir prédictif du modèle réel, soit.

$$R^2_{CV_{ext}} = 0,9668 > 0,5 \quad ; r^2 = 0,968 > 0,6 \quad ; r_0^2 = 0,9996$$

$$r_0'^2 = 0,9993 \quad ; T1 = -0,0326 < 0,1, \quad T2 = -0,0324 < 0,1$$

$$0,85 < k = 1,0035 < 1,15 \quad ; 0,85 < k' = 0,9955 < 1,15$$

$$|r^2 - r_0'^2| = 0,0313 < 0,3$$

III.1.3.6. Domaine d'application

Le domaine d'application a été analysé à l'aide du diagramme de Williams, reproduit dans la figure III. 4. Les résidus de prédiction standardisés sont représentés en fonction des valeurs du levier de chaque composé utilisé pour évaluer le domaine d'applicabilité (AD) d'un modèle QSPR [III.18]. Le diagramme donne la possibilité de vérifier la présence des points aberrants qui sont les composés dont la valeur des résidus standardisées est supérieure à 3 unités de déviation et les composés influents dans la détermination du modèle ayant un levier supérieur au levier critique h^* , calculé selon :

$$h^* = 3(k+1)/n_{tr} = 0,1079 \quad (III.2)$$

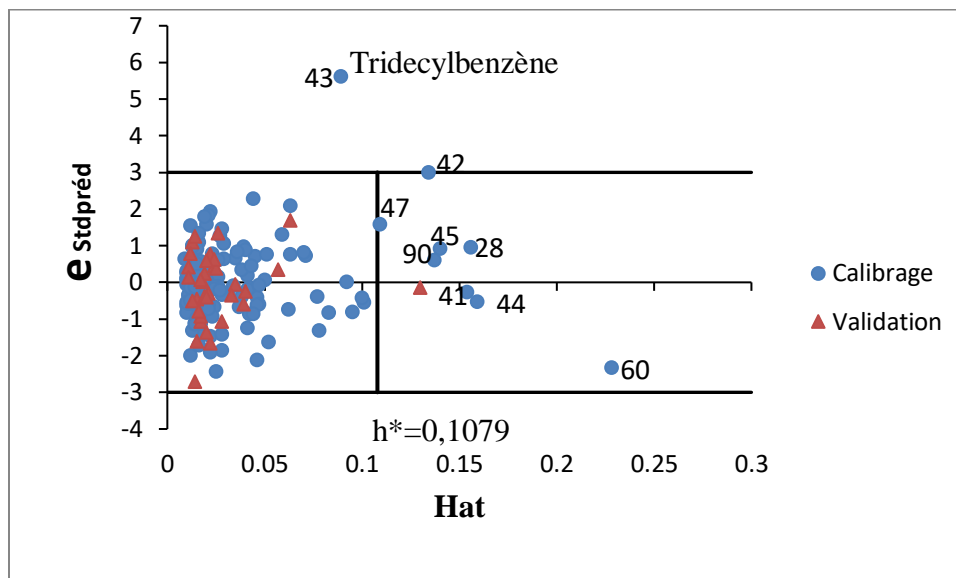


Figure III.4: Diagramme de Williams.

D'après la figure III.4 le seul point aberrant est le Tridécylobenzène appartenant à l'ensemble de calibrage, ayant une valeur de résidu standardisé supérieure à 3 unités d'écart type (3s).

Neuf composés, dont 8 de l'ensemble de calibrage (Hexaméthylbenzène, 1,2,4,5-Tétraisopropylbenzène, Éthylène, Acétylène, 2-Méthylantracène, 9-Méthylantracène, 7-Isopropyl-1-méthylphenanthrène, 1,3,5-Tri-tert-butylbenzène) et, un élément (1-Méthylantracène) de l'ensemble de test, ayant des valeurs de levier supérieures à la valeur critique ($h^* = 0,1079$), sont des points influents. Notons que ces 8 composés cités ont été détectés comme influents par l'étude des résidus.

III.1.4. Conclusion

Un modèle QSPR pour la prédiction des points d'éclair de 173 hydrocarbures insaturés a été établi en appliquant la technique de régression linéaire multiple.

Les résultats obtenus garantissent que les descripteurs moléculaires sélectionnés expliquent avec succès la propriété étudiée qui est le point d'éclair. Les grandes valeurs de R^2 et de Q^2 , ainsi que les valeurs proches des ($SDEP=10,66$ et $SPED_{ext}=9,50$) confirment le pouvoir prédictif du modèle calculé.

Le diagramme de Williams fait ressortir un seul point aberrant de l'ensemble de calibrage qui représente une erreur standardisée supérieure à 3 unités d'écart-type et neuf points influents dont un point appartenant à l'ensemble de test. Ces points ont un levier important ($h_i > h^*$).

Le modèle QSPR obtenu, peut être utilisé avec succès pour estimer les points d'éclair pour de nouveaux composés dont les valeurs expérimentales sont inconnues.

III.2. Modélisation du point d'éclair d'un ensemble de n-alcane

III.2.1. Introduction

Les alcanes sont des hydrocarbures saturés. Ils ne sont constitués que d'atomes de carbone (C) et d'hydrogène (H), liés par des liaisons simples, les atomes de carbone sont reliés à un nombre maximal d'atomes d'hydrogène. Ils peuvent être linéaires ou ramifiés et possèdent alors une formule brute C_nH_{2n+2} ou cycliques de formule générale C_nH_{2n} , où n est un nombre entier.

Les alcanes sont très répandus dans la nature. On les trouve sous forme de gisements de gaz naturel (composés principalement de méthane) et de pétrole (il existe une très grande variété de pétroles qui diffèrent selon les mélanges d'hydrocarbures qu'ils contiennent). La séparation de ces alcanes est réalisée par distillation dans la tour de distillation de la raffinerie.

Le risque d'incendie et d'explosion est l'un des risques majeurs lors de l'utilisation des solvants pétroliers. Les vapeurs de ces substances peuvent former avec l'air des mélanges explosifs. En effet, ils sont tous inflammables à l'exception des solvants de classe A3 ($Tec > 55$ °C) qui sont non inflammables à température ambiante.

Au cours des deux dernières décennies de nombreuses études QSPR ont été consacrées aux températures d'éclair des n-alcane. Nous citerons :

- Une méthode de contribution de groupes proposée par Wang en 1999 pour la prédiction du point d'éclair [III.19].
- Une prédiction de Tec des alcanes par la méthode de contribution de groupes en utilisant les réseaux de neurones artificiels [III.20].
- L'application d'une approche inductive dans les relations structure-propriétés pour le développement d'un modèle QSAR pour la prédiction des points d'éclair d'une série d'alcane [III.21].

Dans ce travail nous nous sommes intéressés à la modélisation des points d'éclair d'une série de 92 n-alcane. Nous avons développé un modèle QSPR en utilisant la technique de régression linéaire multiple en vue de prédire cette propriété à partir des descripteurs moléculaires théoriques calculés à l'aide du logiciel DRAGON [III.10].

III.2.2. Données et méthodes de recherches

Les valeurs expérimentales des points d'éclair des 92 alcanes étudiés reproduites dans le tableau III.5 varient entre 169 K (Propane) et 408 K (Hexadécane), ont été prélevées de la littérature [III.20].

Le nombre d'atomes de carbone dans les structures chimiques des alcanes étudiés varie de 3 à 16 atomes de carbone.

Après la génération de 1113 descripteurs moléculaires théoriques appartenant à diverses classes, une approche algorithme génétique / régression linéaire multiple (AG/RLM) a été appliquée pour modéliser les points d'éclair de l'ensemble de 92 alcanes, aléatoirement séparés en 2 sous ensembles : un ensemble de 74 composés pour le calcul du modèle et un ensemble de 18 composés pour sa validation statistique externe (tableau III.7).

Tableau III.7: Composés étudiés et valeurs des températures d'éclair expérimentales.

N°	Composés	Tec _{exp} (K)	N°	Composés	Tec _{exp} (K)
1	Propane	169	14	3-Éthylhexane	278
2	Butane	213	15	3-Méthylheptane	279
3	2,2-Diméthylpropane	208	16	2,3-Diméthylhexane	283
4	Hexane	250	17	2,4-Diméthylhexane	283
5	2,2-Diméthylbutane	225	18	2,5-Diméthylhexane	271
6	2,3-Diméthylbutane	244	19	3,4-Diméthylheptane	288
7	Heptane	269	20	3-Éthyl-2,3-diméthylpentane	288
8	3,3-Diméthylpentane	254	21	2,2,5-Triméthylhexane	286
9	2,4-Diméthylpentane	261	22	3-Méthyl-octane	297
10	2,2-Diméthylpentane	250	23	2,2,3,4-Tetraméthylpentane	284
11	2,2,4-Triméthylpentane	261	24	4-Éthylheptane	288
12	2,2,3-Triméthylpentane	270	25	3,3,4-Triméthylhexane	288
13	3-Éthyl-3-méthylpentane	276	26	2,3,5-Triméthylhexane	288

Tableau III.7 (suite)

N°	Composés	Tec _{exp} (K)	N°	Composés	Tec _{exp} (K)
27	2,2,3-Triméthylhexane	288	50	Hexadécane	408
28	3,5-Diméthylheptane	288	51	2,2,4,4,6,8,8-Heptaméthylnonane	368
29	Tetraéthylméthane	294	52	2-Méthylbutane	216
30	Décane	319	53	2-Méthylpentane	250
31	4-Éthyl-octane	314	54	2,3,3-Triméthylpentane	273
32	2,4,5-Triméthylheptane	304	55	2,2-Diméthylhexane	269
33	2,3-Diméthyl-octane	314	56	3-Éthyl-2,2-diméthylpentane	286
34	3,3-Diméthyl-octane	314	57	2,3,3-Triméthylhexane	288
35	3,5-Diméthyl-octane	314	58	2,4,6-Triméthylheptane	304
36	2,6-Diméthyl-octane	314	59	3-Éthyl-2,3,4-triméthylpentane	304
37	3,3,4,4-Tétraméthylhexane	304	60	2,3,4,4-Tétraméthylhexane	304
38	2,2,5,5-Tétraméthylhexane	304	61	3,4,5-Triméthylheptane	304
39	2,3,3,4-Tétraméthylhexane	304	62	3-Éthyl-5-méthylheptane	304
40	2,3,4,5-Tétraméthylhexane	304	63	2-Méthylpropane	186
41	2,2,4,4-Tétraméthylhexane	304	64	2,3-Diméthylpentane	258
42	3,3,5-Triméthylheptane	304	65	3-Méthylhexane	258
43	2,3,5-Triméthylheptane	304	66	2,2,3-Triméthylbutane	247
44	3-Éthyl-3-méthylheptane	314	67	Octane	286
45	4-Éthyl-3-méthylheptane	314	68	2,2,3,3-Tétraméthylbutane	273
46	2-Méthylnonane	314	69	2,3,4-Triméthylpentane	273
47	Undécane	333	70	3,4-Diméthylhexane	277
48	Dodécane	344	71	2,7-Diméthyl-octane	314
49	Tétradécane	372	72	Tridécane	352

Tableau III.7 (suite et fin)

N°	Composés	Tec _{exp} (K)	N°	Composés	Tec _{exp} (K)
73	Pentadécane	388	83	2,3,4-Triméthylhexane	288
74	Nonane	304	84	3-Éthyl-2,2-diméthylhexane	311
75	Pentane	224	85	2-Méthylheptane	277
76	3-Éthylpentane	255	86	2,4,4-Triméthylhexane	288
77	3,3-Diméthylhexane	272	87	4,4-Diméthylheptane	288
78	2,3-Diméthylheptane	288	88	5-Méthylnonane	312
79	2,6-Diméthylheptane	299	89	3-Méthylpentane	241
80	2,2,4,4-Tétraméthylpentane	276	90	3-Éthyl-4-méthylhexane	288
81	2,2-Diméthylheptane	297	91	2,2,4-Triméthylhexane	288
82	2,4-Diméthylheptane	288	92	2,2,3,4-Tétraméthylhexane	304

*Composés test

Certains modèles, parmi les 100 modèles obtenus par le logiciel MOBYDIGS [III.11], peuvent présenter des performances similaires. Dans ce cas nous avons opté pour les modèles avec les valeurs de ΔK les plus élevées tout en tenant compte des autres paramètres statistiques.

III.2.3. Résultats et discussions

III.2.3.1. Calcul du modèle

Après l'exécution du programme, le meilleur modèle MLR à deux dimensions obtenu, est donné par l'équation de régression suivante.

$$T_{ec} = 65,212(\pm 4,191) + 1,898(\pm 0,251) Pol + 63,993(\pm 1,631) DP01 \quad (III.5)$$

$$n_{cal} = 74 \quad n_{test} = 18 \quad R^2 = 98,26\% \quad R^2(ajst) = 98,21\% \quad Q^2 = 97,99\% \quad Q^2(ext) = 98,88\%$$

$$S = 5,39 \quad F = 2006,91 \quad k_x = 70,56 \quad k_{xy} = 82,57 \quad SDEC = 5,28 \quad SDEP = 5,67 \quad SDEP_{ext} = 4,24$$

les valeurs élevées et proches de R^2 et $R^2(\text{aj})$ confirment le bon ajustement du modèle, la petite différence en valeur absolue entre R^2 et $Q^2(=0,27)$ montre sa robustesse.

La grande valeur du F de Fisher est un gage de la bonne prédiction des n ($=74$) valeurs de Tec de l'ensemble de calibrage.

La bonne capacité prédictive externe du modèle est confirmée par la valeur élevée de $Q^2(\text{ext})$.

Ici on a: $K_{xy} - K_x = 82,57 - 70,56 = 12,01 > 5$ (une valeur positive). Cette valeur vérifie la condition de la règle QUICK (Q Under Influence K) [III.13], basée sur l'indice de corrélation multivariable K .

Cette règle est déduite de l'hypothèse, que la corrélation totale dans l'ensemble formé par les prédicteurs X du modèle plus la réponse Y (K_{xy}) doit toujours être plus grande que celle uniquement mesurée dans l'ensemble des prédicteurs (K_x). Le calcul de K_{xy} est réalisé en considérant la réponse Y comme une variable et en calculant la matrice de corrélation correspondante.

Pol et DP01 sont des descripteurs moléculaires calculés en utilisant le logiciel MOBYDIGS. La définition des descripteurs utilisés pour le calcul du modèle est donnée au tableau III.8 (page suivante).

Tableau III.8: Classe et définition des deux descripteurs moléculaires.

Descripteur	Classe	Définition
Pol	Descripteur topologique (bloc 2)	<p>Les descripteurs topologiques sont basés sur une représentation graphique de la molécule. Ce sont des quantificateurs numériques de la topologie moléculaire obtenus par l'application des opérateurs algébriques à des matrices représentant des graphes moléculaires et dont les valeurs sont indépendantes de la numérotation ou de l'étiquetage des sommets (atomes). Ils peuvent être sensibles à une ou plusieurs caractéristiques structurales de la molécule telle que la taille, la forme, la symétrie, la ramification et la cyclicité et peuvent également encoder des informations chimiques concernant le type d'atome et la multiplicité des liaisons. Pol est le nombre de polarité, il est calculé sur la matrice de distance comme le nombre de paires de sommets à une distance topologique égale à trois [III.22]. On suppose généralement que le nombre de polarité tient compte de la flexibilité des structures acycliques, le nombre de polarité étant égal au nombre de liaisons autour desquelles les rotations libres peuvent avoir lieu. De plus, il concerne les propriétés stériques des molécules.</p>
DP01	Profils moléculaires de Randic (bloc 11)	<p>Les profils moléculaires sont des séquences de descripteurs moléculaires proposés par Randic et dérivés des distances géométriques interatomiques d'une molécule [III.23-III.24-III.25]. Le logiciel DRAGON fournit deux profils moléculaires dont l'un est beaucoup plus lié à la structure 3D moléculaire globale: (DP01, DP02, ..., DPk, ..., DP20) et l'autre lié à la forme moléculaire où DP01 est le profil moléculaire N° :01.</p>

Une comparaison graphique entre les valeurs expérimentales et prédites du point d'éclair des deux ensembles est présentée dans la figure III.5. On remarque que la faible dispersion des données autour de la droite d'ajustement vérifie le bon ajustement du modèle obtenu.

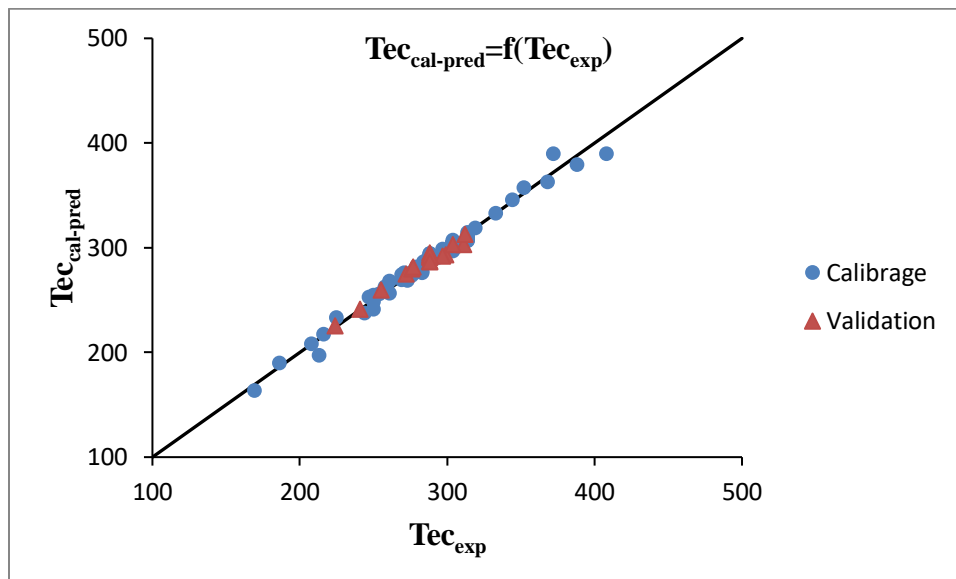


Figure III.5: Droite d'ajustement des Tec prédites en fonction des Tec expérimentales pour les ensembles de calibration et de test.

III.2.3.2. Analyse des résidus

Les résultats obtenus sont condensés dans le tableau III.9, dont les colonnes sont numérotées de (1) à (9). La première rassemble les résidus ordinaires e_i , Notons que 20 résidus ordinaires dépassent en valeur absolue l'erreur standard, $S = 5,39$. Remarquons que 3 valeurs absolues des résidus ordinaires sont importantes. Ce sont dans l'ordre décroissant: e_{50} , e_{49} et e_2 .

Notons que les observations 2, 49, et 50 désignent respectivement : le butane, le tétradécane et l'hexadécane.

Notons aussi que les observations e_{50} , e_{49} ont des résidu ordinaires, en valeurs absolues, supérieures à 3 fois l'erreur standard ($|e_i| > 3S$), soit $3 \times 5,39 = 16,17$.

Tous les résidus standardisés d_i de la colonne (2) sont compris entre les limites ± 3 , à l'exception des points déjà signalés 2, 49 et 50.

La colonne (3) rassemble les résidus studentisés internes r_i qui sont du même ordre de grandeur que les d_i correspondants. On a ici $p = 3$ et $n = 74$, et on constate que tous les t_i exceptés r_2 , r_{49} , et r_{50} sont inférieurs en valeur absolue à $t_{(0,025;n-p)} [= 1,9945]$ qui est le 0,975 quantile d'une loi de Student avec $(n-p)$ degrés de liberté.

La colonne (4) donne les valeurs de h_{ii} , $i^{\text{ème}}$ terme diagonal de la matrice de projection (ou matrice chapeau) : $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ où \mathbf{X} est la matrice des valeurs observées des variables explicatives et \mathbf{X}' sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques. Nous observons la plus grande valeur de h_{ii} pour le point (1) avec $h_{ii}=0,1475 > h^*$, avec $h^* = 3(k+1)/n_{tr} = 3(2+1)/74 = 0,1216$. Ce point est dit point influent. L'observation (1) désigne le propane.

La colonne (5) contient les résidus prédits, qui sont du même ordre de grandeur que les résidus ordinaires correspondants.

La colonne (6) montre le calcul de la somme des carrés des erreurs de prédiction (statistique PRESS) obtenue pour ce modèle. La valeur du PRESS = 2378,6162, et la valeur de SCE = 2059,6254, obtenue pour ce modèle

La colonne (7) condense les estimations $S_{(i)}^2$ de σ^2 calculées selon l'équation (II.83).

$S_{(i)}^2$ intervient dans le calcul des résidus studentisés externes, rassemblés dans la colonne (8); tous les t_i sont du même ordre de grandeur que les r_i correspondants.

Comme les t_i sont inférieurs en valeur absolue à $t_{(0,025;n-p-1)} [= 1,995]$, à l'exception encore une fois de ceux des points 2, 49 et 50 l'analyse des résidus studentisés internes et externes permet de détecter ces 3 observations comme aberrantes.

Enfin, en désignant par SCT la somme des carrées totale (= 11495,959), la statistique PRESS (=2378,6162) conduit à un R^2 de prédiction égal à :

$$R^2_{\text{pred}} = 1 - \frac{\text{PRESS}}{\text{SCT}} = 1 - \frac{2378,6162}{118495.959} = 97,99\%$$

Ainsi, le modèle permettrait d'expliquer environ 98 % de la variabilité de nouvelles observations estimées.

Tableau III.9: Résidus caractéristiques et valeurs estimées des points d'éclair.

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	t_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	Tec estimé
1	4,9836	1,0021	1,0021	1,0022	0,1475	5,8457	34,1723	29,0070	164,0164
2	15,2171	2,9534	2,9534	3,1311	0,0848	16,6278	276,4838	25,8086	197,7829
3	-0,7473	-0,1450	-0,1450	-0,1440	0,0840	-0,8158	0,6656	29,4145	208,7473
4	0,9381	0,1778	0,1778	0,1766	0,0404	0,9777	0,9558	29,4101	249,0619
5	-8,7036	-1,6516	-1,6516	-1,6724	0,0427	-9,0914	82,6541	28,2928	233,7036
6	6,4144	1,2130	1,2130	1,2171	0,0360	6,6538	44,2733	28,8135	237,5856
7	-0,6459	-0,1222	-0,1222	-0,1213	0,0362	-0,6702	0,4492	29,4170	269,6459
8	-1,9083	-0,3584	-0,3584	-0,3562	0,0228	-1,9527	3,8130	29,3700	255,9083
9	4,0246	0,7590	0,7590	0,7567	0,0306	4,1518	17,2373	29,1845	256,9754
10	-4,8636	-0,9171	-0,9171	-0,9160	0,0304	-5,0161	25,1608	29,0747	254,8636
11	-7,6644	-1,4403	-1,4403	-1,4515	0,0238	-7,8512	61,6410	28,5636	268,6644
12	-2,5671	-0,4813	-0,4813	-0,4787	0,0194	-2,6180	6,8537	29,3272	272,5671
13	1,7907	0,3371	0,3371	0,3350	0,0271	1,8405	3,3873	29,3761	274,2093
14	-0,0921	-0,0172	-0,0172	-0,0171	0,0145	-0,0935	0,0087	29,4231	278,0921
15	-2,7613	-0,5179	-0,5179	-0,5152	0,0201	-2,8180	7,9411	29,3121	281,7613

Tableau III.9 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	t_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	Tec estimé
16	5,5478	1,0377	1,0377	1,0382	0,0146	5,6300	31,6971	28,9770	277,4522
17	6,8061	1,2753	1,2753	1,2810	0,0182	6,9319	48,0512	28,7492	276,1940
18	-5,4076	-1,0180	-1,0180	-1,0183	0,0273	-5,5596	30,9088	28,9937	276,4076
19	-6,7510	-1,2629	-1,2629	-1,2683	0,0149	-6,8534	46,9689	28,7623	294,7510
20	-2,5906	-0,4953	-0,4953	-0,4926	0,0568	-2,7465	7,5433	29,3216	290,5906
21	-0,8807	-0,1655	-0,1655	-0,1643	0,0231	-0,9016	0,8129	29,4119	286,8807
22	-2,1457	-0,4030	-0,4030	-0,4006	0,0227	-2,1956	4,8206	29,3559	299,1457
23	-2,6663	-0,5019	-0,5019	-0,4992	0,0270	-2,7402	7,5085	29,3189	286,6663
24	-6,4526	-1,2067	-1,2067	-1,2106	0,0142	-6,5458	42,8477	28,8198	294,4526
25	-3,1242	-0,5911	-0,5911	-0,5883	0,0369	-3,2438	10,5220	29,2785	291,1242
26	-2,1651	-0,4048	-0,4048	-0,4024	0,0136	-2,1949	4,8176	29,3553	290,1651
27	-1,9515	-0,3653	-0,3653	-0,3630	0,0161	-1,9834	3,9340	29,3679	289,9515
28	-5,7487	-1,0749	-1,0749	-1,0761	0,0141	-5,8307	33,9976	28,9444	293,7487
29	3,2174	0,6150	0,6150	0,6123	0,0565	3,4100	11,6282	29,2665	290,7826

Tableau III.9 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	t_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	Tec estimé
30	-0,3674	-0,0700	-0,0700	-0,0695	0,0492	-0,3865	0,1494	29,4212	319,3674
31	4,3387	0,8129	0,8129	0,8109	0,0180	4,4182	19,5204	29,1494	309,6613
32	-1,6081	-0,3013	-0,3013	-0,2994	0,0182	-1,6379	2,6828	29,3856	305,6081
33	2,6109	0,4894	0,4894	0,4868	0,0190	2,6614	7,0832	29,3240	311,3891
34	4,5947	0,8608	0,8608	0,8592	0,0179	4,6782	21,8856	29,1162	309,4053
35	4,8506	0,9087	0,9087	0,9076	0,0177	4,9382	24,3854	29,0810	309,1494
36	3,5492	0,6670	0,6670	0,6644	0,0240	3,6365	13,2242	29,2388	310,4508
37	-1,9480	-0,3840	-0,3840	-0,3817	0,1127	-2,1954	4,8200	29,3621	305,9480
38	6,7101	1,2593	1,2593	1,2646	0,0212	6,8553	46,9952	28,7661	297,2899
39	-1,9696	-0,3818	-0,3818	-0,3795	0,0824	-2,1465	4,6075	29,3628	305,9696
40	0,0989	0,0188	0,0188	0,0186	0,0407	0,1031	0,0106	29,4231	303,9011
41	6,1134	1,1435	1,1435	1,1460	0,0148	6,2049	38,5011	28,8813	297,8866
42	-0,0083	-0,0016	-0,0016	-0,0016	0,0183	-0,0085	0,0001	29,4232	304,0083
43	-0,4563	-0,0855	-0,0855	-0,0849	0,0183	-0,4648	0,2160	29,4202	304,4563
44	7,1976	1,3537	1,3537	1,3618	0,0254	7,3851	54,5392	28,6639	306,8024
45	6,6217	1,2452	1,2452	1,2501	0,0252	6,7927	46,1401	28,7807	307,3783

Tableau III.9 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	t_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	Tec estimé
46	-1,0159	-0,1927	-0,1927	-0,1914	0,0420	-1,0605	1,1246	29,4078	315,0159
47	-0,1683	-0,0322	-0,0322	-0,0320	0,0576	-0,1785	0,0319	29,4228	333,1683
48	-1,9452	-0,3739	-0,3739	-0,3716	0,0670	-2,0849	4,3469	29,3653	345,9452
49	-17,8217	-3,5149	-3,5149	-3,8401	0,1138	-20,1099	404,4092	24,3033	389,8217
50	18,1783	3,5852	3,5852	3,9338	0,1138	20,5123	420,7523	24,0964	389,8217
51	4,6721	0,9067	0,9067	0,9056	0,0847	5,1044	26,0549	29,0825	363,3279
52	-1,9190	-0,3673	-0,3673	-0,3650	0,0590	-2,0393	4,1586	29,3673	217,9190
53	8,3613	1,5846	1,5846	1,6020	0,0402	8,7114	75,8886	28,3827	241,6387
54	-0,3134	-0,0590	-0,0590	-0,0586	0,0280	-0,3225	0,1040	29,4218	273,3134
55	-5,3598	-1,0084	-1,0084	-1,0085	0,0261	-5,5037	30,2903	29,0018	274,3598
56	-2,3301	-0,4383	-0,4383	-0,4358	0,0256	-2,3912	5,7177	29,3436	288,3301
57	-2,4419	-0,4589	-0,4589	-0,4563	0,0239	-2,5017	6,2586	29,3360	290,4419
58	-0,7555	-0,1416	-0,1416	-0,1407	0,0192	-0,7703	0,5933	29,4149	304,7555
59	-0,6682	-0,1319	-0,1319	-0,1310	0,1153	-0,7552	0,5704	29,4160	304,6682
60	-0,2634	-0,0505	-0,0505	-0,0501	0,0600	-0,2802	0,0785	29,4222	304,2634
61	-3,7407	-0,7079	-0,7079	-0,7054	0,0376	-3,8867	15,1060	29,2155	307,7407

Tableau III.9 (suite et fin)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	t_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	Tec estimé
62	-1,3522	-0,2534	-0,2534	-0,2517	0,0182	-1,3772	1,8968	29,3966	305,3522
63	-3,9974	-0,7822	-0,7822	-0,7800	0,0997	-4,4400	19,7135	29,1697	189,9974
64	-0,7879	-0,1479	-0,1479	-0,1468	0,0210	-0,8048	0,6477	29,4142	258,7879
65	-4,8411	-0,9093	-0,9093	-0,9082	0,0229	-4,9545	24,5466	29,0806	262,8411
66	-6,0926	-1,1455	-1,1455	-1,1481	0,0249	-6,2479	39,0360	28,8794	253,0926
67	-1,9262	-0,3646	-0,3646	-0,3623	0,0376	-2,0015	4,0060	29,3681	287,9262
68	3,7181	0,7019	0,7019	0,6994	0,0328	3,8441	14,7770	29,2190	269,2819
69	-1,2949	-0,2427	-0,2427	-0,2410	0,0182	-1,3190	1,7397	29,3988	274,2949
70	-0,8145	-0,1525	-0,1525	-0,1514	0,0162	-0,8280	0,6855	29,4136	277,8145
71	3,5276	0,6668	0,6668	0,6642	0,0353	3,6567	13,3713	29,2390	310,4724
72	-5,9542	-1,1510	-1,1510	-1,1537	0,0775	-6,4545	41,6607	28,8742	357,9542
73	8,2675	1,6187	1,6187	1,6377	0,1007	9,1932	84,5154	28,3374	379,7325
74	-0,3508	-0,0666	-0,0666	0,0423	-0,3663	0,1341	0,0180	29,4214	304,3508

III.2.3.2. Diagnostics d'influence

Le tableau III.10 condense d'autres mesures d'influence utiles, [III.15] dont l'étude complète la recherche des observations aberrantes.

Tableau III.10: Diagnostics d'influence.

	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5	Colonne 6	Colonne 7
<i>Observation i</i>	$WSSD_i$	D_i	DIFITS	$DFBETAS_{0,1}$	$DFBETAS_{0,2}$	$DFBETAS_{0,3}$	$Covratio_i$
1	418,3525	0,0579	0,4168	0,3726	-0,0053	-0,2777	1,1728
2	211,7425	0,2695	0,9534	0,7283	-0,2120	-0,4513	0,7695
3	150,7769	0,0006	-0,0436	-0,0220	0,0232	0,0066	1,1381
4	33,5899	0,0004	0,0363	0,0095	-0,0235	0,0038	1,0860
5	73,2523	0,0405	-0,3530	-0,2052	0,1349	0,0881	0,9691
6	66,1136	0,0183	0,2351	0,1621	-0,0483	-0,0930	1,0165
7	6,1622	0,0002	-0,0235	0,0013	0,0181	-0,0095	1,0820
8	28,7128	0,0010	-0,0544	-0,0381	-0,0020	0,0259	1,0620
9	21,3864	0,0061	0,1345	0,0349	-0,0797	0,0129	1,0505
10	24,9999	0,0088	-0,1622	-0,0509	0,0900	-0,0063	1,0384
11	7,7977	0,0169	-0,2266	-0,0321	0,1320	-0,0444	0,9779
12	8,5669	0,0015	-0,0674	-0,0426	-0,0270	0,0371	1,0538
13	8,9883	0,0011	0,0559	0,0369	0,0351	-0,0377	1,0673
14	2,5061	0,0000	-0,0021	-0,0006	0,0003	0,0002	1,0589
15	0,7290	0,0018	-0,0738	0,0042	0,0420	-0,0262	1,0529
16	2,8882	0,0053	0,1264	0,0405	-0,0140	-0,0124	1,0115
17	2,8620	0,0100	0,1742	0,0213	-0,0803	0,0311	0,9914
18	2,4056	0,0097	-0,1707	0,0127	0,1195	-0,0690	1,0265
19	0,8464	0,0081	-0,1562	-0,0233	-0,0442	0,0175	0,9895

Tableau III.10 (suite)

	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5	Colonne 6	Colonne 7
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	Covratio_i
20	3,0354	0,0049	-0,1208	-0,0608	-0,1051	0,0806	1,0948
21	0,6536	0,0002	-0,0253	0,0051	0,0163	-0,0124	1,0670
22	5,6794	0,0013	-0,0611	0,0230	0,0348	-0,0368	1,0604
23	1,5590	0,0023	-0,0831	-0,0413	-0,0580	0,0475	1,0610
24	1,2924	0,0070	-0,1455	0,0101	0,0211	-0,0326	0,9947
25	1,4589	0,0045	-0,1151	-0,0538	-0,0914	0,0687	1,0675
26	0,1165	0,0008	-0,0472	-0,0043	0,0012	-0,0026	1,0505
27	0,1442	0,0007	-0,0464	-0,0144	-0,0185	0,0133	1,0546
28	1,0124	0,0055	-0,1285	0,0056	0,0162	-0,0254	1,0075
29	2,9661	0,0076	0,1498	0,0751	0,1302	-0,0996	1,0884
30	37,4998	0,0001	-0,0158	0,0104	0,0109	-0,0133	1,0973
31	13,1586	0,0040	0,1098	-0,0388	-0,0124	0,0466	1,0332
32	6,7957	0,0006	-0,0407	0,0005	-0,0155	0,0013	1,0587
33	15,5763	0,0016	0,0677	-0,0270	-0,0103	0,0320	1,0529
34	12,8179	0,0045	0,1159	-0,0401	-0,0124	0,0483	1,0295
35	12,4818	0,0050	0,1219	-0,0413	-0,0123	0,0499	1,0257
36	16,8675	0,0037	0,1042	-0,0529	-0,0476	0,0689	1,0492
37	6,8504	0,0062	-0,1360	-0,0550	-0,1275	0,0858	1,1687
38	4,1715	0,0114	0,1861	-0,0621	-0,1018	0,1048	0,9962
39	5,9191	0,0044	-0,1137	-0,0427	-0,1036	0,0673	1,1302
40	4,2488	0,0000	0,0038	0,0012	0,0031	-0,0018	1,0878
41	2,1629	0,0065	0,1403	0,0051	0,0279	0,0014	1,0017
42	5,3970	0,0000	-0,0002	0,0000	-0,0001	0,0000	1,0629

Tableau III.10 (suite)

	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5	Colonne 6	Colonne 7
<i>Observation i</i>	$WSSD_i$	D_i	DIFITS	$DFBETAS_{0,1}$	$DFBETAS_{0,2}$	$DFBETAS_{0,3}$	$Covratio_i$
43	5,7709	0,0001	-0,0116	-0,0003	-0,0047	0,0008	1,0626
44	6,7964	0,0159	0,2198	0,0287	0,1382	-0,0556	0,9899
45	7,3151	0,0133	0,2009	0,0231	0,1244	-0,0476	1,0018
46	28,2717	0,0005	-0,0401	0,0248	0,0272	-0,0324	1,0875
47	69,3031	0,0000	-0,0079	0,0059	0,0048	-0,0069	1,1072
48	107,1532	0,0034	-0,0996	0,0797	0,0525	-0,0880	1,1118
49	294,9627	0,5288	-1,3760	1,1785	0,2908	-1,0970	0,6635
50	294,9627	0,5501	1,4095	-1,2073	-0,2978	1,1237	0,6467
51	131,4060	0,0254	0,2755	-0,0842	0,1665	0,0171	1,1009
52	124,4626	0,0028	-0,0914	-0,0609	0,0298	0,0318	1,1025
53	50,7295	0,0351	0,3278	0,1392	-0,1724	-0,0229	0,9759
54	10,0055	0,0000	-0,0100	-0,0067	-0,0063	0,0068	1,0734
55	3,4295	0,0091	-0,1652	0,0033	0,1113	-0,0584	1,0261
56	1,0280	0,0017	-0,0706	-0,0328	-0,0481	0,0379	1,0622
57	0,6289	0,0017	-0,0715	-0,0300	-0,0471	0,0351	1,0596
58	9,3000	0,0001	-0,0197	0,0076	0,0073	-0,0107	1,0630
59	6,5248	0,0008	-0,0473	-0,0197	-0,0444	0,0303	1,1786
60	4,6200	0,0001	-0,0127	-0,0045	-0,0111	0,0070	1,1099
61	6,8688	0,0065	-0,1394	-0,0320	-0,1078	0,0559	1,0614
62	6,5601	0,0004	-0,0343	0,0002	-0,0132	0,0014	1,0599
63	246,1438	0,0226	-0,2595	-0,1882	0,0754	0,1093	1,1293
64	23,3152	0,0002	-0,0215	-0,0138	0,0006	0,0087	1,0649
65	14,5763	0,0065	-0,1390	-0,0436	0,0647	-0,0024	1,0310

Tableau III.10 (suite et fin)

	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5	Colonne 6	Colonne 7
<i>Observation i</i>	WSSD_i	D_i	DIFITS	DFBETAS_{0,1}	DFBETAS_{0,2}	DFBETAS_{0,3}	Covratio_i
66	34,5432	0,0112	-0,1833	-0,1374	-0,0175	0,0992	1,0118
67	2,0327	0,0017	-0,0716	0,0233	0,0570	-0,0448	1,0782
68	15,2676	0,0056	0,1287	0,0927	0,0851	-0,0954	1,0565
69	6,7920	0,0004	-0,0329	-0,0196	-0,0122	0,0167	1,0602
70	3,8130	0,0001	-0,0194	-0,0099	-0,0058	0,0079	1,0596
71	20,0297	0,0054	0,1271	-0,0721	-0,0837	0,0977	1,0615
72	149,9057	0,0371	-0,3344	0,2788	0,1484	-0,2926	1,0690
73	244,1858	0,0978	0,5480	-0,4703	-0,1563	0,4542	1,0365
74	14,0004	0,0001	-0,0139	0,0073	0,0105	-0,0106	1,0894

Les résultats obtenus sont condensés dans le tableau III.10, dont les colonnes sont numérotées de (1) à (7). Les valeurs de la somme pondérée des carrés des distance du point i au barycentre des données [III.26] (WSSD _{i} : Weighted of Squared Distance of the center of data), données dans la colonne (1) du tableau III.10 montrent la présence de 6 observations parmi 92 composés sont les plus éloignées. L'observation (1) qui désigne le propane est le composé le plus éloigné avec une valeur de WSSD= 418,3524.

La colonne (2) rassemble les valeurs de la distance de Cook, D _{i} . Les composés indexés par les numéros (1, 2, 49, 50,73) dont les noms sont respectivement : propane, butane, tétradécane, hexadécane, pentadécane, ont des valeurs de D _{i} supérieures à la valeur critique ($4/74=0,0541$). Ces observations sont considérées comme très influentes.

La colonne (3) rassemble les valeurs de la statistique DIFITS, qui permet de mesurer l'influence d'une observation i sur la valeur ajustée ou prédite. Cinq observations (1, 2, 49, 50, 73) ont des valeurs absolues de DIFITS, supérieures à la valeur critique $2\sqrt{3/74} = 0,4027$. Ces observations ont une influence sur la valeur prédite.

Les colonnes (4), (5) et (6) rassemblent les valeurs de la statistique $DFBETAS_{j,i}$. L'examen des colonnes $DFBETAS_{j,i}$ en tenant compte de la valeur critique $\frac{2}{\sqrt{n}} = 2 / \sqrt{74} = 0,2325$ fait ressortir les points influents 1, 2, 49, 50, 72, 73.

La colonne (7) rassemble les valeurs du $COVRATIO_i$ où toutes les valeurs sont supérieures à 1. Ces observations améliorent la précision de l'estimation à l'exception des observations 2, 49, et 50 qui ont des valeurs inférieures à 1 et détériorent la précision de l'estimation.

III.2.3.3. Domaine d'applicabilité

Le domaine d'application a été vérifié à l'aide du diagramme de Williams. La figure III.6 fait ressortir les points aberrants et les points influents du modèle obtenu.

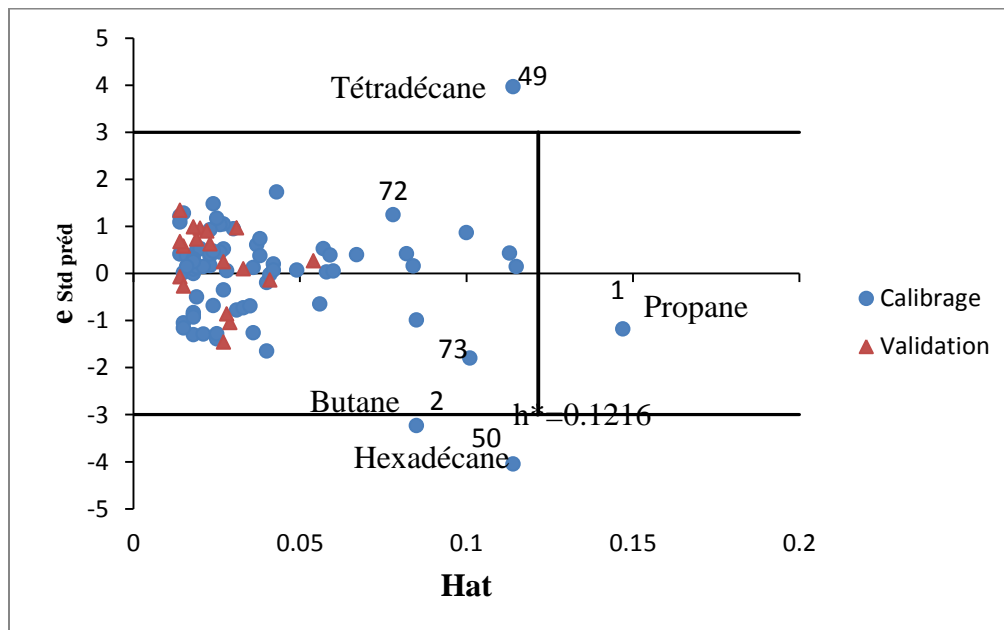


Figure III.6: Diagramme de Williams

On note 3 points aberrants (Butane, Tétradécane et Hexasadécane) appartenant à l'ensemble de calibrage avec des valeurs des résidus standardisés hors de l'intervalle $[-3, +3]$ unité et un seul point influent (Propane) de l'ensemble de calibrage avec une valeur de $h^* > h_i$, avec $h^* = 3(m+1)/n_{tr} = 3(2+1)/74 = 0,1216$ et $h_i_{\text{Propane}} > 0,1216$.

III.2.3.4. Test de randomisation

Le test de randomisation est représenté par la figure III.7 suivante:

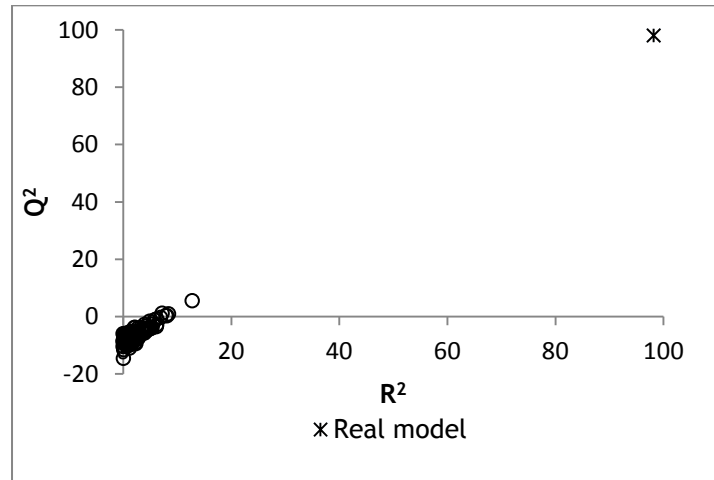


Figure III.7: Test de randomisation

Il est visiblement clair que les statistiques obtenues des 100 vecteurs modifiés du point d'éclair Tec sont plus petites que celles de modèle réel symbolisé par un astérisque (*). Les valeurs de Q^2 sont toutes inférieures à 5,49 % et la majorité des valeurs sont négatives.

En conclusion, une relation QSPR a été établie et le modèle obtenu n'est pas dû au hasard. On conclut que le modèle établi n'est pas dû au hasard.

III.2.3.5. Validation statistique externe

La valeur élevée du coefficient de prédiction externe ($Q^2_{\text{ext}} = 98,88 \%$), illustre la fiabilité du modèle obtenu. Ainsi que la valeur de la somme des erreurs de prédiction externe ($SDEP_{\text{ext}} = 4,24$) est proche de celle pour la prédiction de l'ensemble de calibration ($SDEP = 5,67$).

Les conditions générales pour un pouvoir prédictif du modèle réel sont vérifiées sur l'ensemble de test soient :

$$R^2_{CV_{\text{ext}}} = 0,9661 > 0,5 \quad ; r^2 = 0,9685 > 0,6 \quad ; r_0^2 = 0,9963$$

$$r_0'^2 = 0,9969 \quad ; T1 = -0,0286 < 0,1, \quad T2 = -0,0293 < 0,1$$

$$0,85 < k = 0,9954 < 1,15 ; 0,85 < k' = 1,0044 < 1,15$$

$$|r^2 - r_0'^2| = 0,0284 < 0,3$$

- Nous avons également modélisé l'ensemble des données, les paramètres caractéristiques sont condensés dans le tableau III.11.

Tableau III.11: Valeurs des paramètres statistiques pour tout l'ensemble étudié.

R2	Q2	R2adj	SDEC	SDEP	Kx	Kxy	F	SE
98,14	97,91	98,1	5,081	5,394	70,06	82,31	2352,16	5,1655

L'étude du modèle fait sortir cinq points influents dont deux points sont à la fois des points aberrants, ainsi que la présence d'un autre point aberrant.

- Un travail inverse a été réalisé en utilisant 18 composés pour le calcul et 74 composés pour sa validation statistique externe.

Les résultats représentés dans le tableau III.12 montrent une légère baisse des valeurs de R^2 , Q^2 par rapport à celles du modèle étudié et à celles du modèle calculé sur l'ensemble des données.

Tableau III.12: Valeurs des paramètres statistiques pour le modèle inverse.

R^2	Q^2	Q^2_{boot}	R^2	SDEC	SDEP
96,85	95,47	93,8	96,43	3,949	4,738
Kx	Kxy	Kxy-Kx		F	SE
66,19	80,18	13,99		230,87	4,3261

Remarquons que le nombre des points influents obtenus par la validation croisée est de 15. Cette valeur est très élevée en comparaison de celles des deux modèles précédents (le modèle obtenu au début et le modèle calculé sur la totalité des données).

III.2.4. Conclusion

La modélisation des points d'éclair d'un ensemble de n-alcanes, a conduit à un bon modèle dont les paramètres statistiques sont élevés. Notons aussi que la bonne qualité de l'ajustement, la robustesse, la bonne capacité prédictive et la fiabilité du modèle ont été vérifiées par les graphes et les tests réalisés. L'analyse des résidus appliquée sur l'ensemble de calibration fait ressortir un seul composé (le propane) de l'ensemble de calibration avec un bras de levier important ($h_i > h^*$), et 3 points (le butane, l'hexane et le tétradécane) de l'ensemble de calibration aussi avec des erreurs standardisées hors les limites ± 3 . Ces résultats sont confirmés graphiquement par diagramme de Williams.

III.3. Modélisation des températures d'ébullition d'un mélange de différentes classes de solvants

III.3.1. Données expérimentales et calcul des descripteurs

Le point d'ébullition d'un composé est prédéterminé par les interactions intermoléculaires dans le liquide et par la différence de la fonction de partition moléculaire interne dans la phase gazeuse et dans le liquide à la température d'ébullition. Il doit donc être directement lié à la structure chimique de la molécule. De nombreuses méthodes ont été développées pour estimer le point d'ébullition normale d'un composé à partir de sa structure [III.27].

D'autres propriétés physiques telles que les températures critiques et les points d'éclair peuvent être estimées à partir des points d'ébullition. Au début différentes règles et formules ont été proposées pour corrélérer les points d'ébullition des hydrocarbures homologues avec le nombre d'atomes de carbone ou le poids moléculaire [III.28]. Par la suite, d'autres méthodes tablaient sur des paramètres physiques tels que le parachor et la réfractivité molaire [III.29]. Un résumé plus détaillé des méthodes d'estimation a été rapporté par Horvath [III.30].

Pour des composés essentiellement non polaires tels que les alcanes, les forces intermoléculaires sont les forces de dispersion de London dues à des attractions dipolaires instantanées induisant un dipôle. Les forces de dispersion sont des forces à très courte portée qui augmentent avec le nombre d'électrons qui est proportionnel au poids moléculaire des alcanes. Le point d'ébullition d'un alcane doit dépendre du poids moléculaire (MW) et de la façon dont les molécules se regroupent, ce qui est lié à la géométrie de la molécule. La dépendance vis-à-vis de la géométrie est complexe mais le point d'ébullition devrait diminuer d'une manière générale car la compacité de la molécule augmente si le poids moléculaire reste le même. Balaban a noté que pour le même poids moléculaire, le point d'ébullition diminue avec l'augmentation de la ramification [III.31].

Les composés impliqués dans ce travail, extraits de la littérature et rassemblés dans le tableau III.14, sont formés de diverses classes de solvants : dérivés halogénés, alcools, cétones, esters, benzène et dérivés benzéniques, des composés nitros, composés phosphorés ...

Deux modèles quantitatifs structure-propriété (QSPR) seront développés pour prédire les températures d'ébullition des solvants appartenant à différentes classes en utilisant la régression linéaire multiple (RLM) et la régression par machine à support vecteur (SVM).

III.3.2. Résultats et discussions

III.3.2.1. Choix du modèle linéaire

L'approche hybride algorithme génétique / régression linéaire multiple (AG/RLM) a d'abord été utilisée pour modéliser la température d'ébullition.

L'ensemble a été séparé par algorithme CADEX, en deux ensembles de calibrage (89 composés) et de validation externe (les 22 composés restants). Le choix de la taille du modèle a été basé sur le calcul des modèles de tailles différentes de (2 à 10 descripteurs) en faisant ressortir les points influents et les points aberrants pour chaque modèle. Les paramètres statistiques sont rassemblés dans le tableau III.13 (page suivante).

Tableau III.13: Paramètres statistique des modèles obtenus.

Nombre de descripteurs	Descripteurs	R2	Q2	Q2 _{boot}	Q2 _{ext}	R2 _{adj}	kxy-kx	SDEP	SDEC	F	S	Points influents
2	ATS1e Hy	88,31	87,47	86,84	88,67	88,04	17,71	21,77	21,02	324,89	21,39	4 inf 1 abr
3	nBO FDI Hy	92,27	91,45	0,82	93,27	92	13,62	17,98	17,09	338,41	17,49	4 inf 0 abr
4	VEm1 Mor16p H-052 Hy	93,95	93,22	0,73	92,92	93,66	10,78	16,01	15,13	325,89	15,57	0 abr 2 inf
5	AAC VE _p 1 nHAcc H-052 Hy	95,69	95,06	0,63	94,49	95,43	8,25	13,67	12,77	368,57	13,22	1 inf
6	VRD1 VR _m 1 FDI G2e Q2 Hy	96,69	96,14	0,55	96,77	96,45	5,29	12,08	11,19	399,34	11,65	3 inf
7	AAC VEA1 RDF045u Mor09 pH-052 Hy AMR	97,17	96,61	0,56	93,36	96,92	5,97	11,33	10,35	396,75	10,85	3 inf
8	nBO MATS2m BEHe2 FDI Ds H-052 Hy MLOGP2	97,8	97,33	0,47	92,56	97,58	6,33	10,05	9,12	444,23	9,62	1 abr 2 inf
9	piID AAC GATS1m DP01 RDF025u Mor07v Mor19v G1e Hy	98,53	97,91	0,62	96,14	98,37	4,66	8,90	7,45	589,98	7,90	4 aber 2 inf
10	PJ12 TWC MATS2m GATS3m VE _e 1 Mor14m Mor18v R7e C-001 Hy	96,4	95,35	1,05	91,69	95,94	5,14	13,26	11,67	208,87	12,46	1 abr 2inf

Le graphe III.8 représente le nombre des points aberrants en fonction la taille du modèle (paramètres rassemblés dans le tableau III.13).

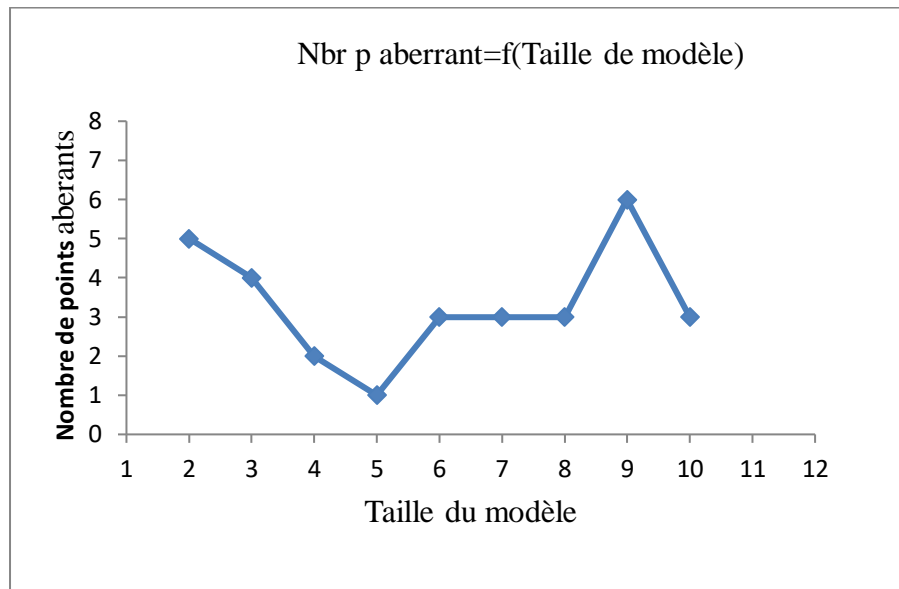


Figure III.8: Choix de la taille du modèle.

Un modèle de dimension cinq a été sélectionné parmi les dix modèles obtenus ; il est représenté par l'équation suivante :

$$\begin{aligned} \text{Teb} = & - 18,11(\pm 17,68) + 66,55(\pm 8,34) \text{ AAC} + 151,51(\pm 5,14) \text{ VE}p1 - 14,43(\pm 1,97) \text{ nHAcc} \\ & - 35,31(\pm 0,57) \text{ H-052} + 66,08(\pm 2,81) \text{ Hy} \end{aligned} \quad (\text{III.6})$$

$$n_{\text{cal}} = 89 \quad n_{\text{test}} = 22 \quad R^2 = 95,6\% \quad R^2(\text{aj}) = 98,21\% \quad Q^2 = 95,06\% \quad Q^2(\text{ext}) = 94,49\%$$

$$S = 13,22 \quad F = 368,67 \quad k_x = 29,55 \quad k_{xy} = 37,8 \quad \text{SDEC} = 12,766 \quad \text{SDEP} = 13,674 \quad \text{SDEP}_{\text{ext}} = 14,433$$

Où :

AAC, VE_{p1}, nHAcc, H-052 et Hy sont les descripteurs utilisés pour la modélisation de la température d'ébullition dont la définition précisée ci - après :

- AAC : indice d'information moyen sur la composition atomique ; appartenant aux indices d'information (bloc 5).

L'indice d'information moyen sur la composition atomique (AAC) est calculé comme contenu d'information total et moyen, respectivement, les relations d'équivalences étant basées sur les types d'atomes chimiques. Notons que même les hydrogènes sont pris en compte dans la dérivation de ces descripteurs [III.32].

- VEp1 : Somme des coefficients du vecteur propre de la matrice de distance pondérée par la polarisabilité appartenant aux indices à valeurs 1 du bloc 10 [III.10].
- H-052 : Descripteurs des fragments centrés sur l'atome (bloc 18), Ce sont des descripteurs moléculaires simples définis comme le nombre de types d'atomes spécifiques dans une molécule. Ils sont calculés en connaissant la composition moléculaire et les connectivités atomiques [III. 10].
- nHAcc: Le nombre d'atomes accepteurs des liaisons hydrogènes (nHAcc) est une mesure de la capacité de liaison hydrogène d'une molécule exprimée en termes de nombre d'accepteurs de liaisons hydrogène possibles. Spécifiquement, il est calculé en additionnant de l'azote, de l'oxygène et du fluor, à l'exclusion de N avec une charge formelle positive, des états d'oxydation plus élevés et de la forme pyrrolyle de l'azote [III. 10].
- Hy: Le facteur hydrophile Hy est un descripteur d'hydrophilie (bloc 20) défini par :

$$\text{Hy} = \frac{(1 + N_{\text{Hy}}) \cdot \text{Log}_2(1 + N_{\text{Hy}}) + nC \left(\frac{1}{nSK} \cdot \text{Log}_2 \frac{1}{nSK} \right) + \sqrt{\frac{N_{\text{Hy}}}{nSK^2}}}{\text{Log}_2(1 + nSK)} \quad (\text{III.7})$$

Où :

- N_{Hy} est le nombre de groupes hydrophiles (-OH, -SH, -NH) ;
- nC le nombre d'atomes de carbone ;
- nSK le nombre d'atomes autres que l'hydrogène [III.33].

Les composés étudiés et les valeurs des descripteurs sélectionnés sont réunis dans le tableau III.14 donné aux pages suivantes.

Tableau III.14: Les composés étudiés, les températures d'ébullitions et les descripteurs sélectionnés.

N°	Composés	Teb(K)	AAC	VEp1	nHAcc	H-052	Hy
1	Éthoxyéthane	307,6	1,159	2,213	1	6	-0,719
2	2-Isopropoxypropane	340,5	1,116	2,623	1	12	-0,802
3	Méthanol	337,7	1,252	1,402	1	0	1,262
4	Éthanol	351,4	1,224	1,724	1	3	0,638
5	Propanol	370,2	1,189	1,978	1	2	0,323
6	Propan-2-ol	355,4	1,189	1,978	1	6	0,323
7	Butan-2-one	352,6	1,239	2,205	1	0	-0,719
8	4-Méthyl pentan-2-one	391	1,167	2,609	1	0	-0,802
9	Heptan-4-one	416,7	1,143	2,779	1	0	-0,828
10	4-Hydroxy-4-méthylpentan-2-one	437	1,295	2,782	2	6	-0,039
11	Benzène	353,2	1	2,449	0	0	-0,921
12	Méthylbenzène	383,8	0,997	2,617	0	0	-0,936
13	Éthanoate de méthyle	329,9	1,435	2,197	2	0	-0,539
14	Éthanoate d'éthyle	350,2	1,379	2,424	2	3	-0,614
15	Éthanoate de propyle	374,6	1,333	2,623	2	2	-0,668
16	Éthanoate d'isopropyle	361,9	1,333	2,624	2	6	-0,668
17	Éthanoate de Butyle	399,2	1,295	2,801	2	2	-0,71
18	Éthanoate de 2-méthylpropyle	390,2	1,295	2,804	2	1	-0,71
19	Éthanoate de Sec-butyle	385,2	1,295	2,802	2	5	-0,71
20	Éthanoate de 3-méthylbutyle	415,1	1,265	2,97	2	2	-0,742
21	Trichloro-méthane	334,2	1,371	1,967	0	0	-0,215
22	Cyclohexane	354	0,918	2,449	0	0	-0,921
23	Nitrométhane	374,2	1,842	1,973	2	0	-0,215
24	Nitroéthane	387	1,761	2,208	2	3	-0,359
25	2-Nitropropane	393,3	1,669	2,417	2	6	-0,46
26	Dichlorométhane	313,6	1,522	1,709	0	0	-0,264
27	Perchlorométhane	349,8	0,722	2,201	0	0	-0,18
28	1,1,1,2,2-Pentachloroéthane	434,9	1,299	2,608	0	0	-0,267
29	(Z)-1,2-Dichloroéthène	333,3	1,585	1,959	0	0	-0,431
30	1,1,2-Trichloroéthène	359,9	1,459	2,189	0	0	-0,359
31	1,1,2,2-Tetrachloro-éthène	393,8	0,918	2,399	0	0	-0,307
32	1,2-Dichloro-propane	368,9	1,435	2,202	0	3	-0,539
33	1-Chlorobutane	351	1,198	2,204	0	2	-0,719
34	1,2-Dichloro-2-méthylbutane	408	1,333	2,601	0	5	-0,668
35	1-Chloro-3-méthyl-butane	372,1	1,166	2,413	0	2	-0,767
36	2,3-Dichloropentane	411	1,333	2,603	0	5	-0,668
37	1,2-Dichloro-pentane	416	1,333	2,602	0	2	-0,668

Tableau III.14 (suite)

N°	Composés	Teb(K)	AAC	VEp1	nHAcc	H-052	Hy
38	Chloro-benzène	405	1,325	2,622	0	0	-0,802
39	2-Chloroéthanol	401,6	1,658	1,978	1	0	0,538
40	1,3-Dichloropropan-2-ol	449	1,73	2,408	1	0	0,311
41	1,2-Dichloropropan-2-ol	455	1,73	2,409	1	0	0,311
42	2-(Chlorométhyl)oxirane	390	1,685	2,209	1	0	-0,539
43	1-Chloro-2-(2-chloroéthoxy)éthane	451	1,64	2,622	1	0	-0,535
44	(R)-Butan-2-ol	372,5	1,159	2,202	1	5	0,132
45	2-Méthylpropan-2-ol	355,5	1,159	2,205	1	9	0,132
46	2,2-Diméthylpropan-1-ol	385	1,135	2,404	1	0	0,004
47	(R)-3-Méthylbutan-2-ol	385	1,135	2,408	1	4	0,004
48	3-Méthylbutan-1-ol	404,2	1,135	2,405	1	2	0,004
49	Pentan-1-ol	411	1,135	2,405	1	2	0,004
50	2-Méthylbutan-1-ol	398	1,135	2,403	1	1	0,004
51	2-Éthylhexan-1-ol	457,6	1,086	2,936	1	1	-0,213
52	Benzène-méthanol	478	1,272	2,761	1	0	-0,158
53	Cyclohexanol	433	1,167	2,611	1	4	-0,088
54	Éthane-1,2-diol	470,2	1,371	1,979	2	0	1,769
55	(R)-Propane-1,2-diol	461	1,335	2,201	2	3	1,41
56	(R)-Butane-1,3-diol	480	1,299	2,403	2	3	1,164
57	3-Oxapentan-1,5-diol	517,5	1,383	2,607	3	0	1,118
58	Furan-2-ylméthanol	449	1,46	2,585	2	0	0,046
59	1-Butoxy butane	415	1,086	2,969	1	4	-0,848
60	2-Éthoxy éthanol	407,8	1,299	2,418	2	3	0,158
61	2-Butoxy éthanol	443,6	1,241	2,79	2	2	-0,039
62	2-(Éthoxyéthoxy) éthanol	474,9	1,325	2,954	3	3	-0,001
63	2-(2-Butoxyéthoxy)éthanol	504,2	1,278	3,271	3	2	-0,119
64	1,4-dioxane	374,3	1,379	2,448	2	0	-0,614
65	4-Méthylpent-3-ène-2-one	402	1,221	2,601	1	0	-0,802
66	Cyclohexanone	429,5	1,221	2,634	1	0	-0,802
67	Butanal	348	1,239	2,211	1	0	-0,719
68	1,1-Diéthoxyéthane	377,2	1,241	2,789	2	6	-0,71
69	Furan-2-carbaldéhyde	434,7	1,495	2,604	2	0	-0,668
70	Méthanoate de Butyle	379,8	1,333	2,614	2	2	-0,668
71	Méthanoate de 3-méthylButyle	397,2	1,295	2,789	2	2	-0,71
72	Éthane-1,2-diyl diacetate	463,2	1,485	3,13	4	0	-0,576
73	Butanoate d'éthyle	393	1,295	2,784	2	3	-0,71
74	Carbonate de diéthyle	400	1,415	2,789	3	6	-0,591

Tableau III.14 (suite et fin)

N°	Composés	Teb(K)	AAC	VEp1	nHAcc	H-052	Hy
75	2-Hydroxypropanoate de butyle	458	1,347	3,124	3	5	-0,065
76	Propane-1,2,3-triol	563	1,414	2,401	3	0	2,492
77	1,4-Dioxaspiro[4.5]déc-2-ylméthanol	523	1,333	3,416	3	0	-0,164
78	Phénoxybenzène	562	1,209	3,588	1	0	-0,897
79	Oxalate de diéthyle	458	1,485	3,102	4	6	-0,576
80	Oxalate de dibutyle	518	1,366	3,688	4	4	-0,696
81	Oxalate de dipentyle	541	1,324	3,946	4	4	-0,734
82	Diéthyl 2,3-dihydroxybutanedioate	553	1,493	3,655	6	6	0,686
83	2-hydroxy propane -1,3-diyl diacétate	532	1,49	3,416	5	0	-0,002
84	Éthanoate de 1,3-diacétyloxypropan-2-yle	531	1,501	3,811	6	0	-0,586
85	Méthyl cyclohexane	374,2	0,918	2,629	0	0	-0,936
86	1-Nitro propane	404,6	1,669	2,417	2	2	-0,46
87	2-Chloro-2-méthylbutane	358,7	1,166	2,41	0	8	-0,767
88	Méthanoate d'éthyle	327,1	1,435	2,217	2	3	-0,539
89	Phosphate de tributyle	562	1,382	4,039	4	6	-0,692
90*	Butan-1-ol	390,4	1,159	2,201	1	2	0,132
91*	2-Méthylpropan-1-ol	380,9	1,159	2,201	1	1	0,132
92*	Porpan-2-one	329,1	1,295	1,975	1	0	-0,646
93*	Méthanoate de 3-méthylbutyle	397,2	1,295	2,789	2	2	-0,71
94*	1,1-Dichloroéthane	313,6	1,5	1,969	0	0	-0,431
95*	1,2-Dichloroéthane	357,1	1,5	1,972	0	0	-0,431
96*	1,1,2,2-Tétrachloroéthane	419,2	1,5	2,413	0	0	-0,307
97*	(E)-1,2-Dichloroéthène	321,4	1,585	1,959	0	0	-0,431
98*	(S)-Butan-2-ol	372,5	1,159	2,202	1	5	0,132
99*	(S)-3-Méthylbutan-2-ol	385	1,135	2,408	1	4	0,004
100*	(R)-Pentan-2-ol	392	1,135	2,408	1	5	0,004
101*	(S)-Pentan-2-ol	392	1,135	2,408	1	5	0,004
102*	Pentan-3-ol	388,7	1,135	2,407	1	4	0,004
103*	2-Méthylbutan-2-ol	374,8	1,135	2,409	1	8	0,004
104*	3-Méthyl-butane-1-ol	404,2	1,135	2,405	1	2	0,004
105*	(S)-Propane-1,2-diol	461	1,335	2,201	2	3	1,41
106*	(S)-Butane-1,3-diol	480	1,299	2,403	2	3	1,164
107*	2-Méthoxy éthanol	397,5	1,335	2,197	2	0	0,312
108*	Éthanal	293,8	1,379	1,72	1	0	-0,528
109*	Diméthoxyméthane	315	1,335	2,197	2	0	-0,539
110*	Méthanoate de méthyle	304,5	1,5	1,972	2	0	-0,431
111*	1,2-Dichloro-3-méthylbutane	416	1,333	2,604	0	1	-0,668

Les paramètres statistiques reproduits dans le tableau III.15 permettent d'évaluer les descripteurs du modèle RLM :

Tableau III.15: Paramètres statistiques obtenus.

Prédicteur	Coéf	SE Coéf	t	P	VIF
Constante	-18,11	17,68	-1,02	0,309	
nHAcc	-14,432	1,969	-7,33	0.000	3,458
AAC	66,548	8,342	7,98	0.000	1,397
VEp1	151,516	5,138	29,49	0.000	3,124
H-052	-35,306	0,5708	-6,19	0.000	1,073
Hy	66,088	2,812	23,5	0.000	1,617

Les valeurs absolues de t des descripteurs dans le tableau III.15 montrent que le descripteur VEp1 et le descripteur Hy sont des descripteurs très significatifs.

La probabilité de t de chaque descripteur est inférieure à 0,05 (Confiance 95%). Ces descripteurs sont statistiquement significatifs dans le modèle obtenu, ce qui montre aussi que leur influence sur la variable réponse n'est pas due au hasard [III.14].

Pour le modèle obtenu les valeurs du VIF associé à chacun des descripteurs sont < 5 [III.34]. Ces descripteurs sont faiblement corrélés les uns avec les autres.

Les valeurs citées dans le tableau précédent sont vérifiées. Ainsi le modèle peut être considéré comme une équation de régression optimale.

III.3.2.1.A. Analyse des résidus et validation du modèle

Les résultats obtenus sont condensés dans le tableau III.16, dont les colonnes sont numérotées de (1) à (9). La première colonne rassemble les résidus ordinaires e_i , Notons également que 21 résidus ordinaires dépassent en valeur absolue l'erreur standard, $S = 13,22$. Remarquons aussi que 4 valeurs absolues des résidus ordinaires sont importantes. Ce sont dans l'ordre décroissant: e_{66} , e_{69} , e_{78} et e_{30} , associées aux composés: cyclohexanone, furan-2-carbaldehyde, phénoxybenzène et 1,1,2-trichloroéthène.

Tous les résidus ordinaires, sont inférieurs, en valeur absolue à 3 fois l'erreur standard ($|e_i| < 3S$), soit $3 \times 13,22 = 39,66$.

Tous les résidus standardisés d_i de la colonne (2) sont compris entre les limites ± 3 .

La colonne (3) rassemble les résidus studentisés internes r_i qui sont du même ordre de grandeur que les d_i correspondants. On a ici $p = 6$ et $n = 89$, et on constate que tous les t_i exceptés r_{30} , r_{66} , r_{69} et r_{78} sont supérieurs en valeur absolue à $t_{(0,025;n-p)} [= 1,9891]$ qui est le 0,975 quantile d'une loi de Student avec $(n-p)$ degrés de liberté.

La colonne (4) donne les valeurs de h_{ii} , $i^{\text{ème}}$ terme diagonale de la matrice de projection : $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ où \mathbf{X} est la matrice des valeurs observées des variables explicatives et \mathbf{X}' sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques.

Nous observons la plus grande valeur de h_{ii} pour le point (76) avec $h_{ii}=0,2463 > h^*$, avec $h^*=3(k+1)/n_{tr} = 3(5+1)/89=0,2022$. Ce point est dit point influent. Ce composé est le propane-1,2,3-triol.

La colonne (5) contient les résidus prédits, qui sont du même ordre de grandeur que les résidus ordinaires correspondants.

La colonne (6) montre le calcul de la somme des carrés des erreurs de prédiction (statistique PRESS) obtenue pour ce modèle. La valeur du PRESS = 16641,2911, et la valeur de SCE = 14504,8066 obtenue pour ce modèle et $R^2(\text{adj})=95,43\%$.

La colonne (7) condense les estimations $S_{(i)}^2$ de σ^2 calculées selon l'équation (II.83).

$S_{(i)}^2$ intervient dans le calcul des résidus studentisés externes, rassemblés dans la colonne (8); tous les t_i sont du même ordre de grandeur que les r_i correspondants.

Les valeurs de t_i sont inférieures en valeur absolue à $t_{(0,025;n-p-1)} [=1,9894]$, à l'exception encore une fois de ceux des points 30,66, 69,78.

Enfin, en désignant par SCT la somme des carrés totale, SCE la somme des carrés des résidus conduit à un coefficient de détermination multiple (R^2) égal à :

$$R^2 = 1 - \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{14504,8066}{336554,3560} = 95,6902\%$$

Ainsi, le modèle permettrait d'expliquer environ 95,69 % de la variabilité de nouvelles observations estimées.

Tableau III.16: Résidus caractéristiques et valeurs estimées des températures d'ébullition des solvants étudiés.

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	T_{eb} estimé
1	-3,5897	-0,2817	-0,2817	0,0710	-3,8640	14,9306	176,7188	-0,2802	311,1897
2	-3,2810	-0,2763	-0,2763	0,1933	-4,0670	16,5406	176,7252	-0,2748	343,7810
3	-8,9027	-0,7263	-0,7263	0,1402	-10,3541	107,2069	175,7638	-0,7242	346,6027
4	9,7030	0,7632	0,7632	0,0752	10,4918	110,0778	175,6464	0,7613	341,6970
5	9,6343	0,7457	0,7457	0,0449	10,0868	101,7429	175,7028	0,7437	360,5657
6	8,9565	0,7038	0,7038	0,0733	9,6645	93,4021	175,8323	0,7016	346,4435
7	16,1152	1,2526	1,2526	0,0528	17,0143	289,4858	173,5442	1,2570	336,4848
8	3,5795	0,2757	0,2757	0,0354	3,7110	13,7714	176,7259	0,2742	387,4205
9	6,8373	0,5275	0,5275	0,0385	7,1107	50,5616	176,2950	0,5251	409,8628
10	0,0392	0,0030	0,0030	0,0429	0,0410	0,0017	176,8879	0,0030	436,9608
11	-5,4317	-0,4239	-0,4239	0,0607	-5,7825	33,4371	176,5049	-0,4218	358,6317
12	0,9046	0,0707	0,0707	0,0622	0,9646	0,9304	176,8773	0,0702	382,8954
13	-15,8802	-1,2477	-1,2477	0,0731	-17,1325	293,5223	173,5700	-1,2520	345,7802
14	-10,6993	-0,8249	-0,8249	0,0373	-11,1142	123,5263	175,4377	-0,8233	360,8993
15	-13,3516	-1,0229	-1,0229	0,0251	-13,6950	187,5527	174,6580	-1,0232	387,9516

Tableau III.16 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	T_{eb} estimé
16	-12,0809	-0,9353	-0,9353	0,0453	-12,6545	160,1354	175,0236	-0,9346	373,9809
17	-10,4169	-0,7963	-0,7963	0,0208	-10,6378	113,1635	175,5365	-0,7945	409,6169
18	-23,4020	-1,7928	-1,7928	0,0250	-24,0028	576,1348	170,0377	-1,8175	413,6020
19	-13,9767	-1,0736	-1,0736	0,0301	-14,4111	207,6793	174,4316	-1,0746	399,1767
20	-16,0119	-1,2253	-1,2253	0,0228	-16,3860	268,4999	173,6883	-1,2291	431,1119
21	-22,7483	-1,7553	-1,7553	0,0389	-23,6695	560,2467	170,3215	-1,7780	356,9483
22	0,8252	0,0650	0,0650	0,0775	0,8946	0,8003	176,8789	0,0646	353,1748
23	13,8619	1,1215	1,1215	0,1258	15,8565	251,4284	174,2074	1,1233	360,3381
24	16,5544	1,3109	1,3109	0,0875	18,1414	329,1089	173,2255	1,3167	370,4456
25	14,5766	1,1543	1,1543	0,0875	15,9749	255,1984	174,0482	1,1567	378,7234
26	-11,0677	-0,8678	-0,8678	0,0692	-11,8903	141,3793	175,2831	-0,8665	324,6677
27	-1,7266	-0,1407	-0,1407	0,1383	-2,0035	4,0142	176,8457	-0,1399	351,5266
28	-10,9420	-0,8576	-0,8576	0,0684	-11,7450	137,9451	175,3207	-0,8562	445,8420
29	-22,4025	-1,7588	-1,7588	0,0716	-24,1299	582,2532	170,2956	-1,7817	355,7025
30	-27,0244	-2,0972	-2,0972	0,0499	-28,4424	808,9700	167,5143	-2,1421	386,9244

Tableau III.16 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	T_{eb} estimé
31	7,6231	0,5978	0,5978	0,0696	8,1935	67,1328	176,1262	0,5955	386,1769
32	4,0905	0,3173	0,3173	0,0488	4,3006	18,4947	176,6734	0,3156	364,8095
33	10,0246	0,7707	0,7707	0,0317	10,3532	107,1895	175,6222	0,7687	340,9754
34	5,1101	0,4028	0,4028	0,0791	5,5489	30,7902	176,5421	0,4008	402,8900
35	4,7595	0,3659	0,3659	0,0319	4,9164	24,1713	176,6025	0,3640	367,3405
36	7,8070	0,6155	0,6155	0,0793	8,4797	71,9054	176,0806	0,6132	403,1930
37	2,3668	0,1842	0,1842	0,0554	2,5057	6,2787	176,8156	0,1831	413,6332
38	-9,3365	-0,7276	-0,7276	0,0579	-9,9098	98,2035	175,7596	-0,7255	414,3365
39	-11,4467	-0,9014	-0,9014	0,0771	-12,4035	153,8476	175,1565	-0,9003	413,0467
40	-18,9880	-1,5268	-1,5268	0,1150	-21,4543	460,2871	171,9199	-1,5393	467,9880
41	-13,1396	-1,0566	-1,0566	0,1151	-14,8484	220,4735	174,5086	-1,0574	468,1396
42	11,3330	0,8908	0,8908	0,0739	12,2375	149,7559	175,1966	0,8897	378,6670
43	12,4872	0,9831	0,9831	0,0768	13,5258	182,9474	174,8282	0,9829	438,5128
44	3,2056	0,2482	0,2482	0,0451	3,3569	11,2685	176,7567	0,2467	369,2944
45	-0,1267	-0,0102	-0,0102	0,1171	-0,1435	0,0206	176,8877	-0,0101	355,6267

Tableau III.16 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	T_{eb} estimé
46	-22,4970	-1,7312	-1,7312	0,0337	-23,2819	542,0452	170,5004	-1,7527	407,4970
47	-8,9808	-0,6901	-0,6901	0,0308	-9,2666	85,8695	175,8730	-0,6879	393,9808
48	3,6126	0,2767	0,2767	0,0248	3,7045	13,7236	176,7247	0,2752	400,5874
49	10,4126	0,7976	0,7976	0,0248	10,6776	114,0107	175,5320	0,7959	400,5874
50	-5,8149	-0,4460	-0,4460	0,0274	-5,9788	35,7464	176,4639	-0,4439	403,8149
51	-9,3710	-0,7290	-0,7290	0,0545	-9,9114	98,2351	175,7552	-0,7270	466,9710
52	18,0010	1,3895	1,3895	0,0397	18,7446	351,3604	172,7730	1,3975	459,9990
53	12,2120	0,9386	0,9386	0,0313	12,6069	158,9345	175,0104	0,9379	420,7880
54	9,1786	0,7458	0,7458	0,1333	10,5898	112,1445	175,7026	0,7438	461,0214
55	3,0550	0,2430	0,2430	0,0955	3,3774	11,4071	176,7621	0,2416	457,9450
56	10,1021	0,7955	0,7955	0,0771	10,9459	119,8116	175,5394	0,7937	469,8979
57	17,9829	1,4187	1,4187	0,0805	19,5580	382,5165	172,5988	1,4275	499,5171
58	4,1066	0,3148	0,3148	0,0265	4,2183	17,7942	176,6767	0,3131	444,8934
59	-4,4135	-0,3416	-0,3416	0,0449	-4,6211	21,3543	176,6392	-0,3398	419,4135
60	2,1138	0,1618	0,1618	0,0229	2,1632	4,6794	176,8321	0,1608	405,6862

Tableau III.16 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	T_{eb} estimé
61	-5,1016	-0,3897	-0,3897	0,0191	-5,2012	27,0523	176,5643	-0,3877	448,7016
62	11,2107	0,8598	0,8598	0,0272	11,5239	132,7994	175,3124	0,8584	463,6893
63	-0,1243	-0,0096	-0,0096	0,0420	-0,1298	0,0168	176,8877	-0,0096	504,3243
64	-0,8274	-0,0641	-0,0641	0,0470	-0,8682	0,7538	176,8792	-0,0637	375,1274
65	12,1981	0,9375	0,9375	0,0312	12,5914	158,5436	175,0149	0,9368	389,8019
66	34,6980	2,6663	2,6663	0,0309	35,8048	1281,9835	161,7372	2,7715	394,8020
67	10,6061	0,8241	0,8241	0,0521	11,1892	125,1972	175,4407	0,8225	337,3939
68	-12,8829	-0,9958	-0,9958	0,0423	-13,4517	180,9473	174,7745	-0,9958	390,0829
69	31,7853	2,4532	2,4532	0,0394	33,0892	1094,8925	164,0617	2,5319	402,9147
70	-6,7880	-0,5202	-0,5202	0,0257	-6,9667	48,5347	176,3112	-0,5179	386,5880
71	-10,5987	-0,8103	-0,8103	0,0210	-10,8265	117,2122	175,4886	-0,8086	407,7987
72	4,0368	0,3188	0,3188	0,0826	4,4004	19,3635	176,6713	0,3171	459,1633
73	-10,5106	-0,8034	-0,8034	0,0205	-10,7309	115,1510	175,5125	-0,8016	403,5106
74	4,9050	0,3825	0,3825	0,0590	5,2124	27,1691	176,5761	0,3805	395,0950
75	-21,6203	-1,6707	-1,6707	0,0417	-22,5603	508,9666	170,9396	-1,6892	479,6203

Tableau III.16 (suite et fin)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>
<i>Observation i</i>	e_i	d_i	r_i	h_{ii}	$e_i/(1-h_{ii})$	$(e_i/(1-h_{ii}))^2$	$(s_i)^2$	t_i	T_{eb} estimé
76	1,8275	0,1592	0,1592	0,2463	2,4245	5,8783	176,8339	0,1583	561,1725
77	-11,0414	-0,8651	-0,8651	0,0678	-11,8439	140,2770	175,2931	-0,8637	534,0414
78	29,7288	2,4493	2,4493	0,1570	35,2633	1243,5033	164,1033	2,5275	532,2712
79	24,2626	1,9125	1,9125	0,0791	26,3462	694,1215	169,0925	1,9443	433,7374
80	4,2629	0,3351	0,3351	0,0742	4,6044	21,2004	176,6485	0,3333	513,7371
81	-6,5219	-0,5209	-0,5209	0,1029	-7,2698	52,8499	176,3097	-0,5186	547,5219
82	-19,5976	-1,6215	-1,6215	0,1641	-23,4459	549,7118	171,2845	-1,6379	572,5976
83	5,6677	0,4540	0,4540	0,1083	6,3558	40,3965	176,4486	0,4518	526,3323
84	-2,8861	-0,2420	-0,2420	0,1859	-3,5451	12,5677	176,7631	-0,2406	533,8861
85	-5,2563	-0,4137	-0,4137	0,0763	-5,6905	32,3816	176,5231	-0,4116	379,4563
86	11,7543	0,9147	0,9147	0,0550	12,4380	154,7045	175,1050	0,9138	392,8457
87	12,9975	1,0366	1,0366	0,1004	14,4485	208,7595	174,5977	1,0371	345,7025
88	-11,1188	-0,8668	-0,8668	0,0583	-11,8077	139,4212	175,2868	-0,8654	338,2188
89	0,8128	0,0662	0,0662	0,1370	0,9419	0,8871	176,8786	0,0658	561,1872

III.3.2.1.B. Diagnostics d'influence

Le Tableau III.17 les mesures d'influence pour compléter la recherche des observations aberrantes (voir la page suivante).

Les valeurs de la somme pondérée des carrés des distance du point i au centre des données ($WSSD_i$) [III.26], données dans la colonne (1) du tableau III.17 montrent la présence de quatre (4) observations parmi 89 composés sont les plus éloignées. Ces observations sont numérotées par 3, 81, 84, et 89 dont les noms des composés sont respectivement : Méthanol, Dipentyloxalate, propane-1,2,3-triyl triacetate (ou bien : Glycerol triacetate) et tributyl phosphate et ce dernier est le plus éloigné.

Les valeurs de la distance de Cook, D_i sont rassemblées dans la colonne (2). Les valeurs D_i des observations (78), (79) et (82) sont supérieures à la valeur critique ($4/89=0,0449$). Ces observations sont considérées comme très influentes.

La colonne (3) rassemble les valeurs de la statistique DFITS. Cinq observations (40, 78, 79, 82) ont des valeurs absolues de DFITS, supérieures à la valeur critique $2 \times \sqrt{6/89} = 0,5193$. Ces observations sont inhabituelles.

Les colonnes (4), (5),... et (9) rassemblent les valeurs de la statistique $DFBETAS_{j,i}$. L'examen des colonnes $DFBETAS_{j,i}$ en tenant compte de la valeur critique $\frac{2}{\sqrt{n}} = 2/\sqrt{89} = 0,2111$ fait ressortir le point influent 78 ayant toutes les valeurs de $DFBETAS_{j,i}$ inférieures au valeur critique .

Les valeurs du $COVRATIO_i$ rassemblées dans la colonne (10) montrent que toutes les observations ont des valeurs supérieures à 1. Ces observations améliorent la précision de l'estimation à l'exception les observations 18, 20, 21, 29, 30, 46, 52, 66, 69, 75, 78, 79 qui ont des valeurs inférieures à 1 et qui détériorent la précision de l'estimation.

Tableau III.17: Diagnostics d'influence.

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>	<i>Colonne10</i>
<i>Observation i</i>	WSSD _i	D _i	DFFITS	DFBETAS _{0,1}	DFBETAS _{0,2}	DFBETAS _{0,3}	DFBETAS _{0,4}	DFBETAS _{0,5}	DFBETAS _{0,6}	COVRATIO _i
1	24,7940	0,0010	-0,0775	-0,0450	0,0284	0,0463	-0,0297	-0,0419	0,0375	1,1510
2	15,4508	0,0031	-0,1345	-0,0021	0,0077	0,0017	0,0152	-0,1242	0,0080	1,3257
3	242,3905	0,0143	-0,2924	-0,1872	0,1032	0,1718	-0,1104	0,0570	-0,0666	1,2038
4	117,9089	0,0079	0,2171	0,1444	-0,0814	-0,1361	0,0784	0,0356	0,0308	1,1147
5	57,4591	0,0044	0,1612	0,1176	-0,0804	-0,0939	0,0579	-0,0032	0,0160	1,0814
6	58,4913	0,0065	0,1973	0,0901	-0,0563	-0,0826	0,0322	0,1228	0,0306	1,1195
7	24,4161	0,0146	0,2969	0,2100	-0,1175	-0,2047	0,1638	-0,1381	-0,1989	1,0126
8	8,3555	0,0005	0,0525	0,0217	-0,0224	-0,0081	0,0095	-0,0300	-0,0242	1,1088
9	14,2924	0,0019	0,1050	0,0203	-0,0381	0,0151	-0,0043	-0,0571	-0,0300	1,0961
10	7,9434	0,0000	0,0006	-0,0002	0,0001	0,0002	-0,0002	0,0005	0,0002	1,1236
11	18,5324	0,0019	-0,1072	-0,0508	0,0597	0,0116	-0,0011	0,0452	0,0366	1,1301
12	16,9451	0,0001	0,0181	0,0048	-0,0083	0,0026	-0,0034	-0,0073	-0,0036	1,1463
13	22,0928	0,0205	-0,3516	-0,2041	0,0496	0,2751	-0,2526	0,1427	0,2329	1,0356
14	6,5273	0,0044	-0,1621	-0,0810	0,0180	0,1161	-0,1062	-0,0145	0,1118	1,0633
15	4,3628	0,0045	-0,1641	-0,0716	0,0314	0,0842	-0,0972	0,0286	0,1132	1,0223

Tableau III.17 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>	<i>Colonne10</i>
<i>Observation i</i>	WSSD _i	D _i	DFFITs	DFBETAS _{0,1}	DFBETAS _{0,2}	DFBETAS _{0,3}	DFBETAS _{0,4}	DFBETAS _{0,5}	DFBETAS _{0,6}	COVRATIO _i
16	5,4062	0,0069	-0,2037	-0,0362	0,0017	0,0677	-0,0589	-0,1389	0,0843	1,0571
17	11,3541	0,0022	-0,1157	-0,0284	0,0238	0,0212	-0,0446	0,0229	0,0654	1,0489
18	11,6271	0,0138	-0,2912	-0,0776	0,0668	0,0511	-0,1153	0,1319	0,1582	0,8703
19	11,9722	0,0060	-0,1894	-0,0130	0,0090	0,0201	-0,0342	-0,1102	0,0715	1,0196
20	25,6427	0,0059	-0,1879	0,0005	0,0310	-0,0350	-0,0205	0,0356	0,0655	0,9864
21	52,9692	0,0208	-0,3578	-0,0319	-0,0919	0,0682	0,0894	0,1056	0,0110	0,8918
22	20,0093	0,0001	0,0187	0,0107	-0,0127	-0,0031	0,0019	-0,0076	-0,0062	1,1655
23	56,3015	0,0302	0,4261	-0,0121	0,2367	-0,1937	0,1145	-0,0773	-0,1437	1,1225
24	23,8087	0,0275	0,4077	-0,0745	0,2711	-0,1373	0,0568	0,0806	-0,1230	1,0395
25	8,8627	0,0213	0,3583	-0,1063	0,2265	-0,0549	-0,0077	0,2187	-0,0688	1,0695
26	104,2739	0,0093	-0,2362	-0,0238	-0,0956	0,0971	0,0108	0,0446	0,0539	1,0939
27	31,3465	0,0005	-0,0560	-0,0414	0,0487	0,0146	-0,0124	0,0180	0,0018	1,2462
28	3,3509	0,0090	-0,2319	0,1294	-0,0864	-0,1667	0,1945	0,0465	-0,1081	1,0943
29	56,5796	0,0398	-0,4947	0,1323	-0,3274	0,0310	0,1672	0,0772	0,0591	0,9223
30	24,1677	0,0385	-0,4907	0,1898	-0,2940	-0,1135	0,2852	0,1049	-0,0506	0,8164

Tableau III.17 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>	<i>Colonne10</i>
<i>Observation i</i>	WSSD _i	D _i	DFFITs	DFBETAS _{0,1}	DFBETAS _{0,2}	DFBETAS _{0,3}	DFBETAS _{0,4}	DFBETAS _{0,5}	DFBETAS _{0,6}	COVRATIO _i
31	11,6189	0,0045	0,1629	0,0750	-0,1081	0,0068	-0,0168	-0,0603	0,0165	1,1264
32	23,9739	0,0009	0,0715	-0,0277	0,0439	0,0117	-0,0418	0,0243	0,0003	1,1225
33	26,9054	0,0032	0,1392	0,0467	-0,0211	-0,0348	-0,0192	0,0021	-0,0519	1,0639
34	7,4231	0,0023	0,1174	-0,0688	0,0582	0,0651	-0,0924	0,0654	0,0270	1,1542
35	13,2738	0,0007	0,0661	0,0050	-0,0067	0,0085	-0,0272	0,0011	-0,0119	1,1002
36	7,4340	0,0054	0,1800	-0,1057	0,0893	0,1001	-0,1417	0,1001	0,0415	1,1365
37	6,8682	0,0003	0,0444	-0,0267	0,0223	0,0280	-0,0372	0,0051	0,0093	1,1358
38	10,4562	0,0054	-0,1798	0,0874	-0,0728	-0,0992	0,1268	0,0472	-0,0104	1,0985
39	67,9572	0,0113	-0,2603	0,0855	-0,1720	-0,0094	0,0785	0,0341	-0,0978	1,0986
40	17,5700	0,0505	-0,5548	0,4030	-0,4649	-0,2715	0,3502	0,0409	-0,2646	1,0242
41	17,5246	0,0242	-0,3813	0,2772	-0,3195	-0,1870	0,2409	0,0281	-0,1820	1,1205
42	24,1602	0,0106	0,2514	-0,0786	0,1824	-0,0262	-0,0358	-0,0518	-0,0639	1,0962
43	5,3359	0,0134	0,2835	-0,1988	0,2348	0,1248	-0,1565	-0,0524	0,0219	1,0858
44	24,5443	0,0005	0,0536	0,0232	-0,0195	-0,0151	0,0036	0,0309	0,0112	1,1212
45	26,9907	0,0000	-0,0037	-0,0007	0,0005	0,0005	0,0002	-0,0032	-0,0007	1,2180

Tableau III.17 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>	<i>Colonne10</i>
<i>Observation i</i>	WSSD _i	D _i	DFFITs	DFBETAS _{0,1}	DFBETAS _{0,2}	DFBETAS _{0,3}	DFBETAS _{0,4}	DFBETAS _{0,5}	DFBETAS _{0,6}	COVRATIO _i
46	7,5800	0,0174	-0,3274	-0,1593	0,1828	0,0277	-0,0253	0,1687	-0,0616	0,8926
47	7,2869	0,0025	-0,1227	-0,0399	0,0514	0,0027	0,0133	-0,0546	-0,0394	1,0720
48	7,1938	0,0003	0,0439	0,0205	-0,0246	-0,0028	-0,0006	-0,0023	0,0126	1,0967
49	7,1938	0,0027	0,1270	0,0593	-0,0711	-0,0081	-0,0017	-0,0067	0,0364	1,0531
50	7,3855	0,0009	-0,0745	-0,0369	0,0430	0,0060	-0,0029	0,0232	-0,0178	1,0899
51	18,4285	0,0051	-0,1746	0,0332	0,0418	-0,1159	0,0882	0,0374	-0,0823	1,0945
52	5,3443	0,0133	0,2840	-0,1163	0,0438	0,1933	-0,1738	-0,1091	0,1364	0,9724
53	2,1352	0,0048	0,1687	-0,0156	-0,0255	0,0678	-0,0778	0,0798	0,0823	1,0414
54	152,5357	0,0143	0,2917	0,0463	-0,0246	-0,0184	0,0109	-0,0496	0,2041	1,1917
55	90,0928	0,0010	0,0785	0,0035	-0,0031	0,0044	-0,0066	0,0151	0,0614	1,1839
56	56,0850	0,0088	0,2294	-0,0090	-0,0113	0,0468	-0,0400	0,0447	0,1916	1,1130
57	51,3955	0,0294	0,4225	0,0245	-0,0439	0,0350	0,0588	-0,1351	0,2681	1,0095
58	3,6593	0,0005	0,0517	-0,0100	0,0174	0,0058	-0,0008	-0,0264	0,0116	1,0968
59	29,9426	0,0009	-0,0737	0,0079	0,0180	-0,0334	0,0272	-0,0222	0,0006	1,1166
60	8,4294	0,0001	0,0246	0,0118	-0,0075	-0,0103	0,0104	0,0047	0,0035	1,0985

Tableau III.17 (suite)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>	<i>Colonne10</i>
<i>Observation i</i>	WSSD _i	D _i	DFFITs	DFBETAS _{0,1}	DFBETAS _{0,2}	DFBETAS _{0,3}	DFBETAS _{0,4}	DFBETAS _{0,5}	DFBETAS _{0,6}	COVRATIO _i
61	7,4686	0,0005	-0,0542	-0,0032	0,0165	-0,0158	0,0007	0,0067	-0,0209	1,0845
62	22,6275	0,0034	0,1435	0,0087	-0,0284	0,0107	0,0494	0,0104	0,0270	1,0477
63	65,6072	0,0000	-0,0020	0,0005	0,0003	-0,0010	0,0001	0,0003	-0,0007	1,1226
64	5,9076	0,0000	-0,0142	-0,0073	0,0026	0,0089	-0,0093	0,0073	0,0094	1,1282
65	7,9761	0,0047	0,1682	0,0498	-0,0449	-0,0183	0,0183	-0,0970	-0,0791	1,0414
66	8,2910	0,0378	0,4950	0,1176	-0,1192	-0,0180	0,0254	-0,2849	-0,2126	0,6485
67	23,8278	0,0062	0,1928	0,1357	-0,0761	-0,1319	0,1056	-0,0902	-0,1290	1,0800
68	11,8391	0,0073	-0,2092	-0,0369	0,0378	0,0344	-0,0455	-0,1407	0,0700	1,0448
69	5,3594	0,0412	0,5128	0,0178	0,1419	-0,1425	0,1643	-0,2520	-0,2775	0,7127
70	4,2746	0,0012	-0,0840	-0,0378	0,0166	0,0445	-0,0507	0,0146	0,0584	1,0823
71	10,6800	0,0024	-0,1185	-0,0321	0,0257	0,0254	-0,0485	0,0235	0,0688	1,0474
72	50,0226	0,0015	0,0952	0,0176	-0,0076	-0,0264	0,0631	-0,0461	-0,0437	1,1637
73	10,4543	0,0023	-0,1161	-0,0268	0,0203	0,0247	-0,0428	-0,0116	0,0649	1,0477
74	11,9522	0,0015	0,0953	0,0176	0,0004	-0,0379	0,0486	0,0517	-0,0421	1,1308
75	42,8472	0,0202	-0,3522	0,1133	-0,0314	-0,1392	0,0201	-0,1773	-0,1190	0,9140

Tableau III.17 (suite et fin)

	<i>Colonne1</i>	<i>Colonne2</i>	<i>Colonne3</i>	<i>Colonne4</i>	<i>Colonne5</i>	<i>Colonne6</i>	<i>Colonne7</i>	<i>Colonne8</i>	<i>Colonne9</i>	<i>Colonne10</i>
<i>Observation i</i>	WSSD _i	D _i	DFFITs	DFBETAS _{0,1}	DFBETAS _{0,2}	DFBETAS _{0,3}	DFBETAS _{0,4}	DFBETAS _{0,5}	DFBETAS _{0,6}	COVRATIO _i
76	198,4437	0,0014	0,0905	-0,0098	0,0013	0,0223	-0,0121	-0,0104	0,0781	1,4242
77	94,6672	0,0091	-0,2329	0,0936	-0,0119	-0,1484	0,0415	0,0988	-0,0779	1,0926
78	143,8847	0,1861	1,0906	-0,6859	0,2384	0,9626	-0,7500	-0,2290	0,3329	0,8133
79	46,7807	0,0524	0,5698	0,0322	0,0317	-0,1528	0,3066	0,2413	-0,2168	0,8911
80	172,7806	0,0015	0,0944	-0,0292	0,0028	0,0360	0,0130	0,0106	-0,0074	1,1522
81	257,3241	0,0052	-0,1756	0,0777	-0,0099	-0,1038	0,0136	-0,0171	-0,0138	1,1755
82	199,7969	0,0861	-0,7258	0,1039	0,0115	-0,1095	-0,2289	-0,1968	-0,2246	1,0606
83	108,6048	0,0042	0,1574	0,0085	-0,0162	-0,0070	0,0871	-0,0694	-0,0122	1,1881
84	225,8292	0,0022	-0,1150	-0,0122	0,0173	0,0135	-0,0728	0,0473	0,0373	1,3154
85	18,5058	0,0024	-0,1183	-0,0441	0,0698	-0,0094	0,0109	0,0465	0,0234	1,1499
86	7,8304	0,0081	0,2204	-0,0534	0,1492	-0,0521	0,0240	0,0074	-0,0730	1,0709
87	15,8118	0,0200	0,3465	-0,0348	0,0261	0,0405	-0,1318	0,2863	-0,0017	1,1056
88	19,8309	0,0078	-0,2154	-0,1136	0,0124	0,1750	-0,1464	-0,0181	0,1417	1,0814
89	291,8003	0,0001	0,0262	-0,0156	0,0060	0,0177	-0,0066	0,0090	0,0051	1,2458

III.3.2.1.C. Évaluation du modèle

Afin d'évaluer et vérifier la performance du modèle AGR-RML, des validations internes et externes ont été effectuées (Q^2_{LOO} , Q^2_{LMO}) ainsi qu'une vérification de la capacité de détermination (R^2).

La validation externe est effectuée sur l'ensemble de test pour vérifier la capacité prédictive du modèle, donnée par Q^2_{ext} .

Tableau III.18: Les paramètres statistiques pour modèle RLM.

n_{tr}	n_{ext}	Q^2_{LOO} (%)	R^2 (%)	R^2_{adj} (%)	Q^2_{ext}	Q^2_{boot}	F
89	22	95,06	95,69	95,43	94,49	94,66	368,67
SDEC	SDEP	SDEP _{ext}	K _{xy}	K _{xx}	s	ERAM	EAM
12,766	13,674	14,433	37,8	29,55	13,22	2,790	10,937

La valeur élevée de R^2_{adj} (%) = 95,43 indique un excellent accord entre la corrélation et la variation des données.

Les paramètres statistiques obtenus montrent que le modèle représenté par l'équation (III.4) a établi une forte corrélation entre les variables sélectionnées et la propriété étudiée, caractérisée par un excellent coefficient de détermination. La valeur de $R^2 = 95,69$ % indique que 95,69 de variation totale est expliquée par le modèle.

La grande valeur de F du Fisher (F = 368,57), indique une excellente capacité prédictive par le modèle, avec une erreur standard s = 13,22.

La petite différence entre R^2 et Q^2_{LOO} informe sur la robustesse du modèle. La valeur élevée de Q^2_{boot} (%) = 94,66 confirme à la fois la prévisibilité interne et la stabilité du modèle.

III.3.2.1.D. Vérification de la qualité d'ajustement

Le graphe des valeurs prédites et calculées en fonction des valeurs expérimentales des points d'éclair (Figure III.9) pour les ensembles de calibrage et de test, confirme que le modèle linéaire a un très bon ajustement et peut être utilisé pour prédire la propriété étudiée.

Le graphe de la figure III.9 vérifie le bon ajustement du modèle obtenu.

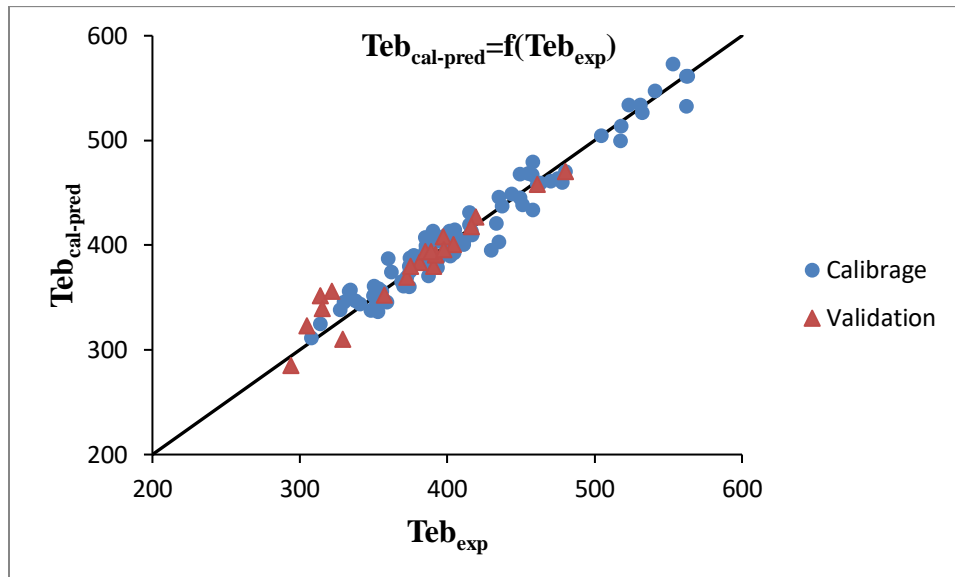


Figure III.9: Droite d'ajustement des Teb prédites en fonction des Teb expérimentales pour les ensembles de calibrage et de test.

III.3.2.1.E. Validation externe

La valeur élevée du coefficient statistique externe $Q^2_{\text{ext}} = 94,49\%$ atteste de la bonne validation externe du modèle, confirmée d'ailleurs par les valeurs proches du SDEP et $SDEP_{\text{ext}}$.

Les conditions générales acceptées pour le pouvoir prédictif du modèle réel sont données par les paramètres statistiques obtenus pour l'ensemble de test, calculés selon *Tropsha et al.* Ces paramètres vérifient ces conditions.

$$R^2_{CV_{\text{ext}}} = 0.9430 > 0.5 \quad ; \quad r^2 = 0.9161 > 0.6 \quad ; \quad r_0^2 = 0.9925$$

$$r_0'^2 = 0.9954 \quad ; \quad T1 = -0.0834 < 0.1, \quad T2 = -0.0865 < 0.1$$

$$0.85 < k = 0.9903 < 1.15 ; 0.85 < k' = 1.0085 < 1.15$$

$$|r^2 - r_0'^2| = 0.0029 < 0.3$$

III.3.2.1.F. Diagramme de Williams

Les résidus de prédiction standardisés sont représentés en fonction des valeurs du levier de chaque composé utilisé pour évaluer le domaine d'applicabilité (AD) Le diagramme de Williams donne la possibilité de vérifier la présence des points aberrants et les points influents dans la détermination du modèle ayant un levier supérieur à h^* au levier critique, calculé selon : $h^* = 3(k+1)/n_{tr} = 0,2022$.

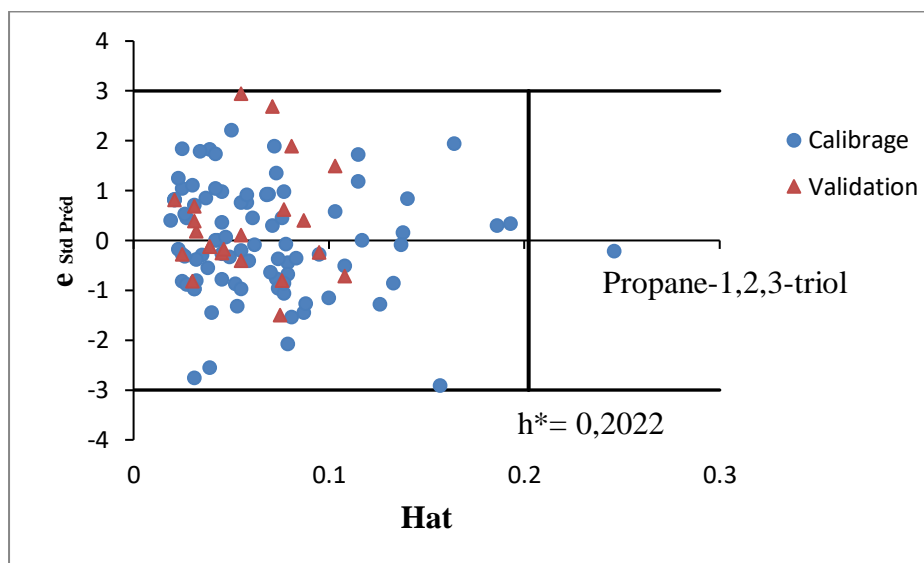


Figure III.10: Diagramme de Williams.

Sur la figure III.10, le seul point influent est le Propane-1,2,3-triol appartenant à l'ensemble de calibration, ayant une valeur de levier supérieure à la valeur critique ($h^* = 0,2022$). Ce résultat est déjà vérifié par l'étude des résidus.

Le propane-1,2,3-triol est un hydrocarbure oxygéné de l'ensemble de calibration, ayant la température d'ébullition la plus élevée dans l'ensemble des composés étudiés ($T_{eb} = 563$ K) et ayant la valeur de (H_y) la plus élevée dans l'ensemble étudié. Ce composé connu dans le commerce sous le nom « Glycérol », largement utilisé comme solvant; comme édulcorant; dans la fabrication de dynamite, de cosmétiques, de savons liquides, de bonbons, de liqueurs, d'encre

et de lubrifiants; pour garder les tissus flexibles; en tant que composant de mélanges antigel; comme source de nutriments pour les cultures de fermentation dans la production d'antibiotiques; et en médecine. Il a beaucoup d'autres utilisations.

Les composés indexés par les numéros 30, 66, 69 et 78 dont les noms sont respectivement : 1,1,2-trichloroéthène, cyclohexanone, furan-2-carbaldehyde et phénoxybenzène, ont des valeurs élevées des résidus standardisés mais ne dépassent pas ± 3 et qui ont déjà signalé dans l'étude des résidus.

III.3.2.1.G. Test de randomisation

La figure III.11 montre que les 100 vecteurs modifiés de la température d'ébullition T_{eb} (+) sont plus petites que celles de modèle réel symbolisé par un triangle (\blacktriangle).

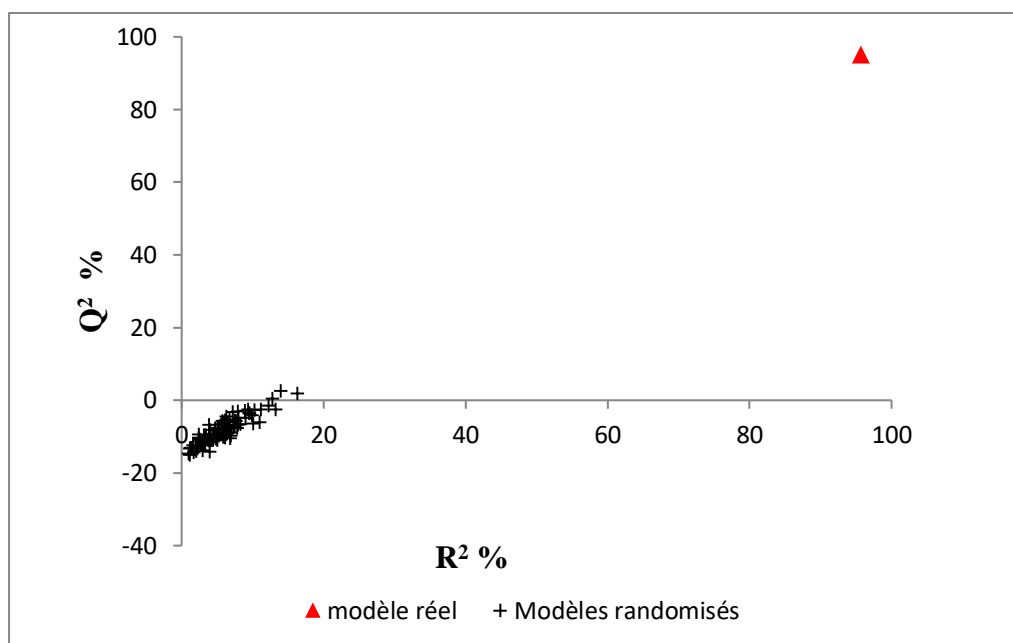


Figure III. 11: Test de randomisation.

Les modèles randomisés sont très loin de modèle réel. Le modèle obtenu n'est pas dû au hasard.

III.3.2.1.H. Autres analyses des erreurs

En plus des techniques de validation interne et externe mentionnées, une analyse des erreurs relatives moyennes (ERA) a été effectuée dans cette étude afin d'évaluer la performance prédictive des modèles développés.

La distribution de l'erreur relative est recommandée pour chaque composé (i) [III.35]. Pour l'erreur relative absolue (ERA) donnée selon Eq. (III.8), pour chaque individu, l'erreur relative absolue moyenne (ERAM) selon Eq. (III.9), et l'erreur absolue moyenne (EAM) selon les équations suivantes Eq. (III.10).

$$ERA = \left| \frac{T_{eb \text{ exp}} - T_{eb \text{ pred}}}{T_{eb \text{ exp}}} \right| \quad (III.8)$$

$$ERAM = 100 \times \frac{\sum_{i=1}^n ERA}{n} \quad (III.9)$$

$$EAM = \frac{\sum_{i=1}^n |T_{eb \text{ exp}} - T_{eb \text{ pred}}|}{n} \quad (III.10)$$

Où

$T_{eb \text{ exp}}$ et $T_{eb \text{ préd}}$ sont respectivement les températures d'ébullition expérimentales et prédites pour chaque composé i. Les valeurs des erreurs sont reproduites dans le tableau III.19.

Tableau III.19: Températures d'ébullition expérimentales et prédites et les valeurs des erreurs.

N°	Composés	T _{eb} exp	T _{eb} calc	EA	ERA
1	Éthoxy éthane	307.6	311.19	3.59	0.0117
2	2-isoPropoxypropane	340.5	343.78	3.28	0.0096
3	Méthanol	337.7	346.6	8.9	0.0264
4	Éthanol	351.4	341.7	9.7	0.0276
5	Propanol	370.2	360.57	9.63	0.0260
6	Propan-2-ol	355.4	346.44	8.96	0.0252
7	Butan-2-one	352.6	336.48	16.12	0.0457

Tableau III.19 (suite)

N°	Composés	Teb _{exp}	Teb _{Calc}	EA	ERA
8	4-Méthyl pentan-2-one	391	387.42	3.58	0.0092
9	Heptan-4-one	416.7	409.86	6.84	0.0164
10	4-Hydroxy-4-méthylpentan-2-one	437	436.96	0.04	0.0001
11	Benzène	353.2	358.63	5.43	0.0154
12	Méthylbenzène	383.8	382.9	0.9	0.0023
13	Éthanoate de méthyle	329.9	345.78	15.88	0.0481
14	Éthanoate d'éthyle	350.2	360.9	10.7	0.0306
15	Éthanoate de propyle	374.6	387.95	13.35	0.0356
16	Éthanoate d'isopropyle	361.9	373.98	12.08	0.0334
17	Éthanoate de Butyle	399.2	409.62	10.42	0.0261
18	Éthanoate de 2-méthylpropyle	390.2	413.6	23.4	0.0600
19	Éthanoate de Sec-butyle	385.2	399.18	13.98	0.0363
20	Éthanoate de 3-méthylbutyle	415.1	431.11	16.01	0.0386
21	Trichlorométhane	334.2	356.95	22.75	0.0681
22	Cyclohexane	354	353.17	0.83	0.0023
23	Nitrométhane	374.2	360.34	13.86	0.0370
24	Nitroéthane	387	370.45	16.55	0.0428
25	2-Nitropropane	393.3	378.72	14.58	0.0371
26	Dichlorométhane	313.6	324.67	11.07	0.0353
27	Perchlorométhane	349.8	351.53	1.73	0.0049
28	1,1,1,2,2-Pentachloroéthane	434.9	445.84	10.94	0.0252
29	(Z)-1,2-Dichloroéthane	333.3	355.7	22.4	0.0672
30	1,1,2-Trichloroéthane	359.9	386.92	27.02	0.0751
31	1,1,2,2-Tétrachloroéthane	393.8	386.18	7.62	0.0193
32	1,2-Dichloro-propane	368.9	364.81	4.09	0.0111
33	1-Chlorobutane	351	340.98	10.02	0.0285
34	1,2-Dichloro-2-méthylbutane	408	402.89	5.11	0.0125
35	1-Chloro-3-méthyl-butane	372.1	367.34	4.76	0.0128
36	2,3-Dichloropentane	411	403.19	7.81	0.0190
37	1,2-Dichloro-pentane	416	413.63	2.37	0.0057
38	Chloro-benzene	405	414.34	9.34	0.0231
39	2-Chloroéthanol	401.6	413.05	11.45	0.0285
40	1,3-Dichloropropan-2-ol	449	467.99	18.99	0.0423
41	1,2-Dichloropropan-2-ol	455	468.14	13.14	0.0289
42	2-(Chlorométhyl)oxirane	390	378.67	11.33	0.0291
43	1-Chloro-2-(2-chloroethoxy)éthane	451	438.51	12.49	0.0277
44	(R)-Butan-2-ol	372.5	369.29	3.21	0.0086

Tableau III.19 (suite)

N°	Composés	Teb _{exp}	Teb _{calc}	EA	ERA
45	2-Méthylpropan-2-ol	355.5	355.63	0.13	0.0004
46	2,2-Diméthylpropan-1-ol	385	407.5	22.5	0.0584
47	(R)-3-Méthylbutan-2-ol	385	393.98	8.98	0.0233
48	3-Méthylbutan-1-ol	404.2	400.59	3.61	0.0089
49	Pentan-1-ol	411	400.59	10.41	0.0253
50	2-Méthylbutan-1-ol	398	403.81	5.81	0.0146
51	2-Éthylhexan-1-ol	457.6	466.97	9.37	0.0205
52	Benzèneméthanol	478	460	18	0.0377
53	Cyclohexanol	433	420.79	12.21	0.0282
54	Éthane-1,2-diol	470.2	461.02	9.18	0.0195
55	(R)-Propane-1,2-diol	461	457.95	3.05	0.0066
56	(R)-Butane-1,3-diol	480	469.9	10.1	0.0210
57	Diéthylene glycol	517.5	499.52	17.98	0.0347
58	Furan-2-ylméthanol	449	444.89	4.11	0.0092
59	1-Butoxy butane	415	419.41	4.41	0.0106
60	2-Éthoxy éthanol	407.8	405.69	2.11	0.0052
61	2-Butoxy éthanol	443.6	448.7	5.1	0.0115
62	2-(Éthoxy éthoxy) éthanol	474.9	463.69	11.21	0.0236
63	2-(2-Butoxyéthoxy) éthanol	504.2	504.32	0.12	0.0002
64	1,4-Dioxane	374.3	375.13	0.83	0.0022
65	4-Méthylpent-3-ene-2-one	402	389.8	12.2	0.0303
66	Cyclohexanone	429.5	394.8	34.7	0.0808
67	Butanal	348	337.39	10.61	0.0305
68	1,1-Diethoxyéthane	377.2	390.08	12.88	0.0341
69	Furan-2-carbaldehyde	434.7	402.91	31.79	0.0731
70	Butyl formate	379.8	386.59	6.79	0.0179
71	Isopentyl formate	397.2	407.8	10.6	0.0267
72	Éthane-1,2-diyl diacetate	463.2	459.16	4.04	0.0087
73	Butanoate d'éthyle	393	403.51	10.51	0.0267
74	Diéthyl carbonate	400	395.1	4.9	0.0122
75	Butyl 2-hydroxypropanoate	458	479.62	21.62	0.0472
76	Propane-1,2,3-triol	563	561.17	1.83	0.0033
77	1,4-Dioxaspiro[4	523	534.04	11.04	0.0211
78	1,1'-Oxydibenzene	562	532.27	29.73	0.0529
79	Diéthyl oxalate	458	433.74	24.26	0.0530
80	Dibutyl oxalate	518	513.74	4.26	0.0082
81	Dipentyloxalate	541	547.52	6.52	0.0121
82	Diéthyl 2,3-dihydroxysuccinate	553	572.6	19.6	0.0354

Tableau III.19 (suite et fin)

N°	Composés	Teb _{exp}	Teb _{Calc}	EA	ERA
83	2-hydroxy propane -1,3	532	526.33	5.67	0.0107
84	Propane-1,2,3-triyl triacetate	531	533.89	2.89	0.0054
85	Méthyl cyclohexane	374.2	379.46	5.26	0.0141
86	1-Nitro propane	404.6	392.85	11.75	0.0290
87	2-Chloro-2-méthylbutane	358.7	345.7	13	0.0362
88	Éthyl formate	327.1	338.22	11.12	0.0340
89	Tributyl phosphate	562	561.19	0.81	0.0014
				EAM	ERAM
				10.38	2.5655

Selon [III.35], il convient de mentionner que le modèle est plus applicable et utile si le pourcentage de l'ERA% pour une classe chimique étudiée est inférieur à 10%. Deux fourchettes d'erreur ont été considérées pour illustrer les différences entre les 4 classes de solvant étudiées comme on peut le voir sur le tableau III .20 (page suivante).

Tableau III.20: Classes des solvants, leurs distributions dans chaque intervalle d'EAR% et les ERA % associées.

N°	Nom de classe	Nbr de composés par classe	ERA% ≤ 5	5 ≤ ERA% ≤ 10	ERA%
1	Solvants organiques oxygénés	60	60	0	2.5022
2	Hydrocarbures	4	4	0	1.4076
3	Dérivés halogénés	15	15	0	2.9604
4	Solvants organique phosphorés	1	1	0	0.1441
5	Composés nitro	4	4	0	3.6479
6	Solvants mixtes	5	5	0	3.1286

Toutes les classes étudiées ont un pourcentage inférieur à 10%. Les valeurs EAM % pour les composés de calibrage, sont inférieures à 5. Selon les données de Tableau III.20, nous pouvons mentionner que le modèle est applicable et utile pour la prédiction des points d'éclair de classes des solvants étudiées.

III.3.2.2. Modèle non-linéaire (machine à support vecteur)

Une régression SVM a été utilisée pour développer un modèle sur les composés de l'ensemble de calibrage utilisé dans le calcul du modèle RLM, en utilisant le même sous-ensemble de descripteurs sélectionné pour la construction du modèle linéaire.

La construction du modèle machine à support vecteur en exploitant la fonction de base radiale de RBF (fonction gaussienne radiale et sigmoïde) a inclus le choix des valeurs optimales des hyperparamètres C , γ , ε . Ces valeurs sont respectivement: 116 ; 0,075 ; 0,115.

Avec cette procédure de réglage, nous avons essayé d'obtenir la plus faible racine de l'erreur quadratique moyenne (RMSE) liée au meilleur paramètre de régression en utilisant le leave-one-out (LOO) en tenant compte du RMSE de l'ensemble de test.

Les valeurs optimales obtenues des paramètres SVM sont :

$$n_{\text{cal}}= 89 \quad n_{\text{test}} = 22 \quad R^2= 96,71\% \quad Q^2= 95,18\% \quad Q^2(\text{ext}) = 93,20\% \quad \text{SDEC} = 11,16$$

$$\text{SDEP}=13,51 \quad \text{SDEP}_{\text{ext}}= 12,48 \quad C= 116 \quad \gamma =0,075 \quad \varepsilon =0,115$$

Les valeurs des températures d'ébullition estimées et prédites pour l'ensemble de calibrage sont données dans le tableau suivant :

Tableau III.21 : Valeurs des températures d'ébullition estimées et prédites pour l'ensemble de calibrage

N°	Composés	Teb _{exp}	Teb _{cal}	ei
1	Éthoxyéthane	307.6	312.802	-5.202
2	2-IsoPropoxypropane	340.5	355.196	-14.696
3	Méthanol	337.7	339.9	-2.200
4	Éthanol	351.4	341.98	9.420
5	Propanol	370.2	355.492	14.708

Tableau III.21 (suite)

N°	Composés	Teb _{exp}	Teb _{cal}	ei
6	Propan-2-ol	355.4	354.737	0.663
7	Butan-2-one	352.6	333.662	18.938
8	4-Méthylpentan-2-one	391	389.118	1.882
9	Heptan-4-one	416.7	413.116	3.584
10	4-Hydroxy-4-méthylpentan-2-one	437	434.019	2.981
11	Benzène	353.2	366.911	-13.711
12	Méthylbenzène	383.8	391.827	-8.027
13	Éthanoate de méthyle	329.9	344.603	-14.703
14	Éthanoate d'éthyle	350.2	358.078	-7.878
15	Éthanoate de propyle	374.6	386.278	-11.678
16	Éthanoate d'isopropyle	361.9	370.971	-9.071
17	Éthanoate de Butyle	399.2	408.176	-8.976
18	Éthanoate de 2-méthylpropyle	390.2	414.29	-24.090
19	Éthanoate de Sec-butyle	385.2	394.288	-9.088
20	Éthanoate de 3-méthylbutyle	415.1	429.766	-14.666
21	Trichlorométhane	334.2	341.686	-7.486
22	Cyclohexane	354	362.915	-8.915
23	Nitrométhane	374.2	363.622	10.578
24	Nitroéthane	387	376.211	10.789
25	2-Nitropropane	393.3	386.153	7.147
26	Dichlorométhane	313.6	299.621	13.979
27	Perchlorométhane	349.8	361.113	-11.313
28	1,1,1,2,2-Pentachloroéthane	434.9	446.792	-11.892
29	(Z)-1,2-Dichloroéthene	333.3	334.856	-1.556
30	1,1,2-Trichloroéthene	359.9	374.627	-14.727
31	1,1,2,2-Tétrachloro-éthene	393.8	389.711	4.089
32	1,2-Dichloro-propane	368.9	354.187	14.713
33	1-Chlorobutane	351	339.442	11.558
34	1,2-Dichloro-2-méthylbutane	408	400.128	7.872
35	1-Chloro-3-méthylbutane	372.1	368.464	3.636
36	2,3-Dichloropentane	411	400.445	10.555
37	1,2-Dichloropentane	416	413.519	2.481
38	Chlorobenzene	405	419.666	-14.666
39	2-Chloroéthanol	401.6	396.554	5.046
40	1,3-Dichloropropan-2-ol	449	463.71	-14.710
41	1,2-Dichloropropan-2-ol	455	463.894	-8.894
42	2-(Chlorométhyl)oxirane	390	375.28	14.720

Tableau III.21 (suite et fin)

N°	Composés	Teb _{exp}	Teb _{cal}	ei
43	1-Chloro-2-(2-chloroéthoxy)éthane	451	446.827	4.173
44	(R)-Butan-2-ol	372.5	370.757	1.743
45	2-Méthylpropan-2-ol	355.5	370.245	-14.745
46	2,2-Diméthylpropan-1-ol	385	402.471	-17.471
47	(R)-3-méthylbutan-2-ol	385	391.885	-6.885
48	3-Méthylbutan-1-ol	404.2	396.344	7.856
49	Pentan-1-ol	411	396.344	14.656
50	2-Méthylbutan-1-ol	398	398.996	-0.996
51	2-Éthylhexan-1-ol	457.6	465.78	-8.180
52	Benzèneméthanol	478	463.263	14.737
53	Cyclohexanol	433	418.333	14.667
54	Éthane-1,2-diol	470.2	461.824	8.376
55	(R)-Propane-1,2-diol	461	466.411	-5.411
56	(R)-Butane-1,3-diol	480	475.729	4.271
57	3-Oxapentan-1,5-diol	517.5	502.834	14.666
58	Furan-2-ylméthanol	449	449.32	-0.320
59	1-Butoxy butane	415	414.191	0.809
60	2-Éthoxy éthanol	407.8	401.841	5.959
61	2-Butoxy éthanol	443.6	444.588	-0.988
62	2-(Éthoxyéthoxy) éthanol	474.9	460.151	14.749
63	2-(2-Butoxyéthoxy)éthanol	504.2	498.362	5.838
64	1,4-Dioxane	374.3	376.287	-1.987
65	4-Méthylpent-3-ène-2-one	402	392.515	9.485
66	Cyclohexanone	429.5	397.989	31.511
67	butanal	348	334.626	13.374
68	1,1-Diethoxyéthane	377.2	382.816	-5.616
69	Furan-2-carbaldehyde	434.7	412.809	21.891
70	Méthanoate de Butyle	379.8	384.812	-5.012
71	Méthanoate de 3- méthylbutyle	397.2	406.229	-9.029
72	Éthane-1,2-diyl diacetate	463.2	473.959	-10.759
73	Butanoate d'éthyle	393	400.296	-7.296
74	Carbonate de diéthyle	400	396.846	3.154
75	2-Hydroxypropanoate de Butyle	458	476.32	-18.320
76	Propane-1,2,3-triol	563	566.157	-3.157
77	1,4-Dioxaspiro[4,5]déc-2-ylméthanol	523	537.672	-14.672
78	Phénoxybenzène	562	547.337	14.663
79	Oxalate de diéthyle	458	443.324	14.676
80	Oxalate de dibutyle	518	506.902	11.098
81	Oxalatede dipentyle	541	531.867	9.133

Tableau III.21 (suite et fin)

N°	Composés	Teb _{exp}	Teb _{cal}	ei
82	Diéthyl 2,3-dihydroxybutanedioate	553	567.249	-14.249
83	2-Hydroxy propane -1,3-diyl diacétate	532	532.383	-0.383
84	Éthanoate de 1,3-diacétyloxypropan-2-yle	531	535.95	-4.950
85	Méthyl cyclohexane	374.2	388.89	-14.690
86	1-Nitro propane	404.6	400.713	3.887
87	2-Chloro-2-méthylbutane	358.7	352.766	5.934
88	Méthanoate d'éthyle	327.1	335.393	-8.293
89	Phosphate de tributyle	562	547.312	14.688
90*	Butan-1-ol	390.4	374.836	15.564
91*	2-Méthylpropan-1-ol	380.9	377.041	3.859
92*	Propan-2-one	329.1	304.327	24.773
93*	Méthanoate de 3-méthylbutyle	397.2	406.229	-9.029
94*	1,1-Dichloroéthane	313.6	333.789	-20.189
95*	1,2-Dichloroéthane	357.1	334.303	22.797
96*	1,1,2,2-Tétrachloroéthane	419.2	419.398	-0.198
97*	(E)-1,2-Dichloroéthène	321.4	334.856	-13.456
98*	(S)-Butan-2-ol	372.5	370.757	1.743
99*	(S)-3-Méthylbutan-2-ol	385	391.885	-6.885
100*	(R)-Pentan-2-ol	392	389.923	2.077
101*	(S)-Pentan-2-ol	392	389.923	2.077
102*	Pentan-3-ol	388.7	391.735	-3.035
103*	2-Méthylbutan-2-ol	374.8	386.015	-11.215
104*	3-Méthylbutan-1-ol	404.2	396.344	7.856
105*	(S)-Propane-1,2-diol	461	466.411	-5.411
106*	(S)-Butane-1,3-diol	480	475.729	4.271
107*	2-Méthoxy éthanol	397.5	388.979	8.521
108*	Éthanal	293.8	275.078	18.722
109*	Diméthoxyméthane	315	334.831	-19.831
110*	Méthanoate de méthyle	304.5	319.924	-15.424
111*	1,2-Dichloro-3-méthylbutane	416	418.988	-2.988

(*) Composés de validation

II.3.2.3. Comparaison entre les paramètres statistiques des deux modèles

Les paramètres statistiques des deux modèles sont présentés dans le tableau III.22.

Tableau III.22: Les paramètres statistiques pour les deux modèles (RLM et SVM).

Statistiques	Modèle RLM	Modèle SVM
Taille	5	5
R^2	95,69%	96,71%
Q^2	95,06%	95,18%
Q^2_{ext}	94,49%	93,20%
SDEC	12,766	11,16
SDEP	13,674	13,51
SDEP _{ext}	14,433	12,48
hyperparamètres		C= 116
		$\gamma = 0,075$
		$\varepsilon = 0,115$

Le modèle non linéaire (SVM) présente une légère amélioration des valeurs statistiques en comparant à ceux du modèle linéaire à l'exception d'une légère baisse de valeur de coefficient de prédiction externe.

III.3.2.4. Comparaison des droites d'ajustement

Une comparaison entre les qualités d'ajustement des deux modèles donnée par la figure suivante :

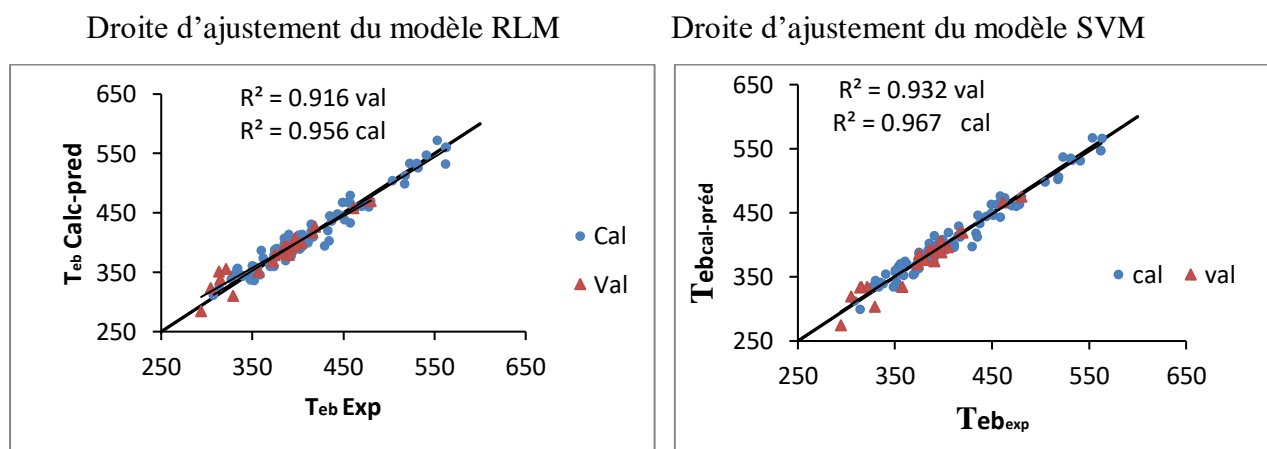


Figure III.12 : Droites d'ajustement de modèle RLM de modèle SVM

Remarquons ici que R^2 de la droite d'ajustement du modèle SVM est supérieure à celui du modèle RLM, ce qui montre que l'ajustement de la droite SVM est mieux que celui de la droite RLM.

III.3.2.5. Comparaison des distributions des erreurs

La distribution des résidus pour le modèle non linéaire est plus acceptable et plus satisfaisante par rapport à celle du modèle linéaire (figure III.13), puisque 1 résidu de l'ensemble de calibration pour la méthode SVM (soit 1,12 %) est supérieure à 2 fois l'erreur standard S ($\geq 2S = 2 \times 11,223$), mais deux résidus dans l'ensemble de validation (soit 9,09 %). Notons, que les plus importants résidus (calibrage et validation) valent respectivement 31,511 pour le composé (cyclohexanone), 24,773 et 22,797 pour les composés (porpan-2-one et 1,2-dichloroéthane).

Notons finalement aussi que le modèle RLM présente cinq résidus supérieures à l'erreur standard, 3 composés de l'ensemble de calibration et 2 de l'ensemble de validation.

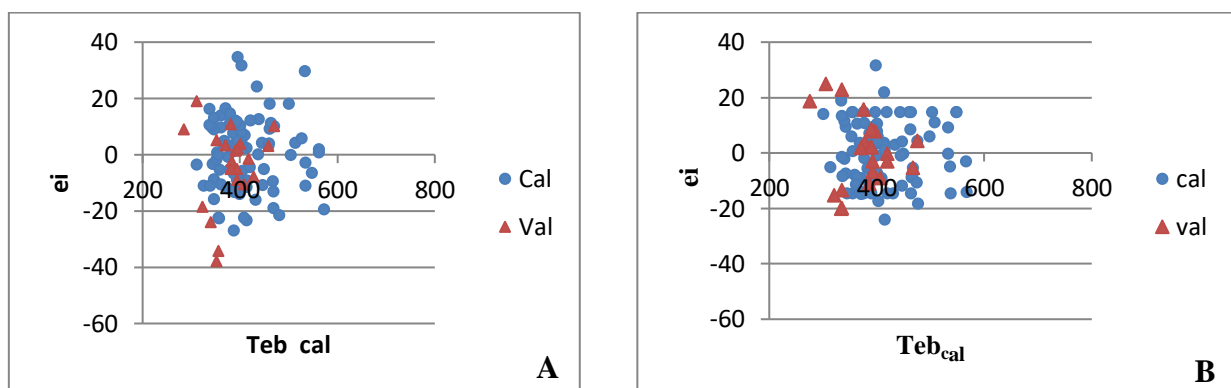


Figure III.13 : Distribution des résidus en fonction des valeurs prédites : (A)-MLR ; (B)-SVM

III.3.3. Conclusion

Une approche QSPR basée sur la méthode RLM et la méthode SVM pour prédire les températures d'ébullitions d'un mélange de différentes classes de solvant.

Les descripteurs sélectionnés portent sur les informations de la composition atomique et moléculaire, le nombre d'atomes inclus dans la formation des liaisons d'hydrogène et l'hydrophilie des molécules.

Une étude des erreurs a été appliquée sur le modèle RLM pour la détection des points aberrants et comparée avec la représentation graphique en utilisant le diagramme de Williams


Les deux modèles développés sont stables, robustes et d'une bonne capacité prédictive. Le peu de différence entre les résultats des deux modèles incite à opter pour le modèle MLR qui permet une interprétation facile et accepter.

RÉFÉRENCES BIBLIOGRAPHIQUES

- [III.1] Evlanov, S.F., Khim,Zh., Prik, I., (1991), J. Appl. Chem-USSR (Engl. Transl), 64: 747-752.
- [III.2] Tetteh, J., Takahiro, S., Metcalfe, E., Howells, S. J., (1999), Chem.Inf .Comp. Sci , 39: 491.
- [III.3] Katritzky, A.R., Petrukhin, R., Jain, R., Karelson, M., (2001). J. Chem.Inf .Comp. Sci, 41: 1521-1530.
- [III.4] ASTM International, General test Method. 2004;14.02, (ASTM, West. Conshohocken, PA.
- [III.5] Liaw, H.J., Lee, Y.H., Tang, C.L., Hsu, H.H., Liu, J.H., 2002. J. Loss Prevent .Proc. 15, 429-438.
- [III.6] Lyman, J., Reehl, W.F., Rosenblatt,D.H., Handbook of Chemical Property Estimation Methods. New York: Mc Graw Hill .1982: 751-752.
- [III.7] Vidal, M., Rogers, W.J., Holste, J.C., Mannan, M.S., (2004), A review of estimations method for flash points and flammability limits. Process .Saf. Prog. 23, 47-55.
- [III.8] Keshavarz, M.H., (2012), Estimation of the flash points of saturated and unsaturated hydrocarbons. Indian .J Eng Mater Sci. 19, 269-278.
- [III.9] Keshavarz, M.H., Ghanbarzadeh, M., 2011. Simple method for reliable predicting flash points of unsaturated hydrocarbons. J .Hazard .Mater . 193, 335- 341.
- [III.10] Todeschini, R., Consonni,V., Mauri. A., and Pavan, M., DRAGON Software for the Calculation of Molecular Descriptor. Version. 5.3 for Windows, Talete S. r. l., Milan. Italy. 2005.
- [III.11] Todeschini, R., Ballabio, D., Consonni, V., Mauri, A., and Pavan, M., MOBYDIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 1.1 for windows, Milano. 2009.
- [III.12] Leardi, R., Boggia, R.,Torrile, M., 1992. J.Chemometr. 6, 267-281.
- [III.13] Todeschini , R., Maiocchi, A., Consonni, V.,(1999), Chemom. Intell .Lab .Syst. 46,13-29.
- [III.14] Ramsey, F. L.,Schafer, D. W., 1997. The Statistical Sleuth: A Course in Methods of Data Analysis. Wadsworth Publishing Company, Belmont, CA.
- [III.15] Montgomery, D.C., Peck, E.A.,1992. Introduction to Linear Regression Analysis, Second Edition, Wiley-Interscience, New York, pp.527.
- [III.16] Zheng, F., Bayram, E., Sumithran, S.P., Ayers, J.T., Zhen, C.G., Schmitt, J.D., Dwoskim, L.P., Crooks, P.A., (2006), Bio.org .Med .Chem. 14, 3017-3037.

- [III.17] Golbraikh, A., Tropsha. A., (2002), Beware of q^2 !. J. Mol .Graph. Model. 20, 269-276.
- [III.18] Gramatica, P., (2007), Principles of QSAR models validation internal and external QSAR. Comb. Sci, 26: 694-701.
- [III.19] Wang, K. Q., (1999), A new method for predicting the densities of alkanes from the information of molecular structure-group bond contribution method, Chin. J. Org. Chem, 19: 304 - 308.
- [III.20] Pan, Y., Jiang, J., Wang, Z., (2007), Prediction of the flash points of alkanes by group bond contribution method using artificial neural networks, Front. Chem. Eng. China, 1(4), 390–394.
- [III.21] Mathieu, D., (2010), Inductive modeling of physico-chemical properties: Flash point of alkanes. Journal of Hazardous Materials, 179: 1161–1164.
- [III.22] Platt, J. R., (1947), influence of the neighbor bonds on additive bond properties in paraffins, J. Chem. Phys. 15, 419-420.
- [III.23] Randic, M., (1995), Molecular shape profile , J .Chem. Inf. Comput. Sci., 35, 373-382.
- [III.24] Randic, M., Razinger, M. (1995), On Characterization of Molecular Shapes J. Chem. Inf. Comput., Sci, 35, 594-606.
- [III.25] Randic, M., (1995), Molecular Profiles Novel Geometry-Dependent Molecular Descriptors, New. J. Chem., 19, 781-791.
- [III.26] Weisberg, S., (2005), Applied Linear Regression, 3rd edn, John Wiley and sons, Inc, New Jersey.
- [III.27] Handbook of Chemical Property Estimation Methods; Lyman, W. J., Reehl, W. F., Rosenblatt, D., Eds.; McGraw-Hill: New York, 1982; Chapter 12.
- [III.28] Walker, J., (1894), The boiling points of homologous compounds. Part I. Simple and mixed ethers, J .Chem. Soc .Trans, 65, 193–202.
- [III.29] Meissner, H. P., (1949), Critical Constants from Parachor and Molar Refraction, Chem. Eng. Prog, 45, 149.
- [III.30] Horvath, A. L., Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds; Elsevier: Amsterdam: 1992; Chapter 2.
- [III.31] Balaban, A. T., A Personal View About Topological Indices for QSAR / QSPR, QSPR / QSAR Studies by Molecular Descriptors, Nova Science Publishers, Huntington, NY, 2001, 1-30.

- [III.32] Dancoff, S.M., Quastler, H., *Essays on the Use of Information Theory in Biology*, University of Illinois, Urbana (ILL), 1953.
- [III.33] Todeschini, R., Vighi, M., Finizio, A., Gramatica, P., (1997), 3D-Modeling and Prediction by WIHIM Descriptors. Part 8. Toxicity and Physico-Chemical Properties of Environmental Priority Chemicals by 2D-TI and 3D-WIHIM Descriptors, SAR & QSAR. *Environ. Res.*, 7, 173-193.
- [III.34] Bagheri, M., Golbraikh, A., (2011), Rank-based ant system method for non-linear QSPR analysis: QSPR studies of the solubility parameter, SAR QSAR. *Environ. Res.* 23, 59–86.
- [III.35] Borhani, T.N. G., Afzali, A., Bagheri, M., (2016), QSPR estimation of the auto-ignition temperature for pure hydrocarbons, *Process Safety and Environmental Protection*, 103, 115–125.



*Conclusion
générale*

Conclusion générale

Les produits chimiques tout au long de leur production, manipulation, transport et utilisation représentent un véritable danger pour la santé et l'environnement. Les gens de tous âges, des plus jeunes au plus âgés, parlant différentes langues et utilisant différents alphabets, de conditions sociales très différentes, éventuellement illettrés, sont chaque jour confrontés aux produits dangereux.

Face à ce danger, et étant donné l'importance du commerce mondial des produits chimiques et la nécessité de mettre au point des programmes nationaux pour assurer l'utilisation, le transport et l'élimination de ces produits en toute sécurité.

Le SGH (Système Générale Harmonisé) décrit la classification des produits chimiques par types de danger et propose des éléments de communication correspondant à ces dangers, y compris des étiquettes et des fiches de données de sécurité.

Les dangers les plus évidents sont les dangers d'incendie ou d'explosion. Plusieurs critères techniques et méthodes d'essai spécifiques pour l'identification des liquides inflammables et combustibles.

Dans le cas des liquides inflammables, la classification s'effectue en fonction de la valeur du point d'éclair et du point d'ébullition. Ces deux propriétés ont fait le sujet de notre travail de recherche.

Le point d'éclair d'un liquide est la température la plus basse à laquelle ce liquide libère assez de vapeur pour s'enflammer (commencer à brûler) à la surface de ce liquide. On trouve parfois plusieurs valeurs de point d'éclair pour un produit chimique donné selon la méthode de mesure utilisée.

Nous avons utilisé la méthodologie QSPR pour relier ces deux propriétés pour trois ensemble de composés chimiques, à des descripteurs moléculaires théoriques, calculés à l'aide de logiciels spécialisés du commerce.

Nous avons développé deux modèles linéaires pour les deux séries d'hydrocarbures saturés et les n -alcane pour prédire les points d'éclair en utilisant une régression linéaire multiple. Enfin, nous avons réalisé une étude comparative entre deux modèles ; un modèle linéaire et un autre non linéaire en utilisant les Machines à Vecteurs Supports.

Conclusion générale

Pour les points d'éclair des deux premiers ensembles, les deux modèles RLM obtenus montrent qu'une forte linéarité de la relation entre ces propriétés et les descripteurs moléculaires a été établie. La qualité de l'ajustement, robustesse interne et externe, capacité prédictive ont été vérifiés. La détection des points aberrants a été réalisée en procédant à une étude des erreurs et graphiquement en donnant le domaine d'applicabilité de chaque ensemble étudié.

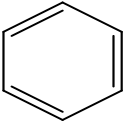
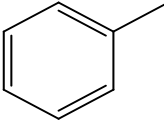
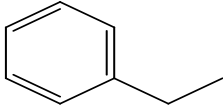
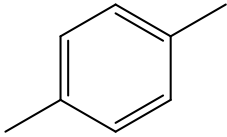
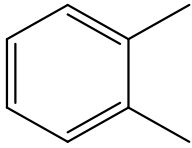
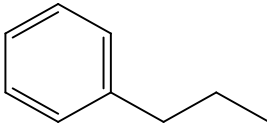
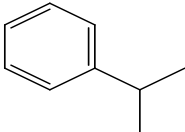
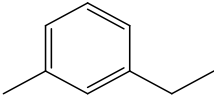
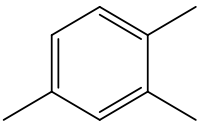
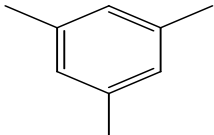
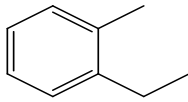
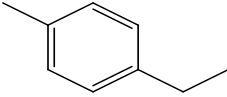
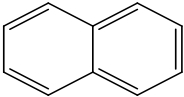
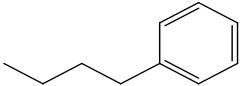
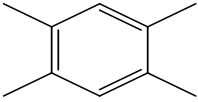
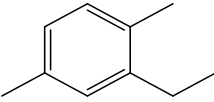
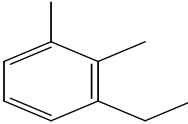
La comparaison de la qualité des modèles RLM et SVM pour le troisième ensemble formé de différentes classes de solvants, montre qu'il a une différence légère supériorité et une meilleure prédiction du modèle SVM mais ce peu de différence entre les résultats des deux modèles incite à opter pour le modèle MLR qui permet une interprétation facile et acceptable.

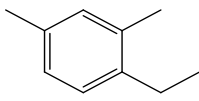
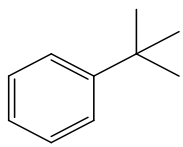
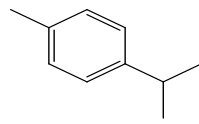
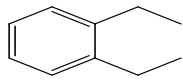
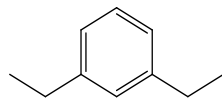
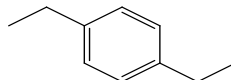
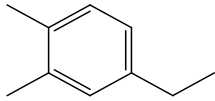
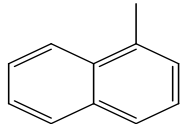
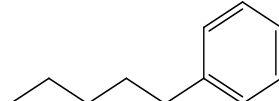
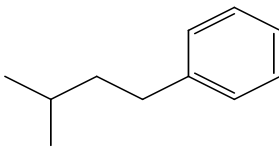
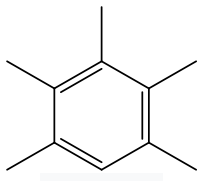
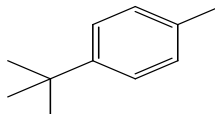
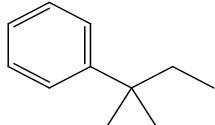
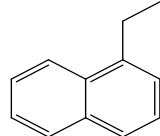
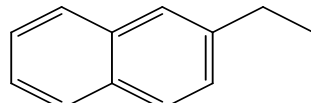
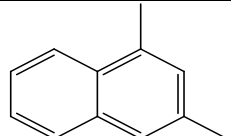
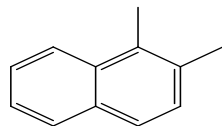
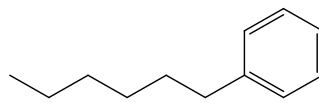
Annexe I

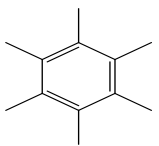
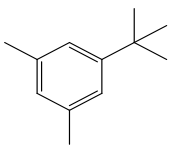
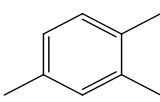
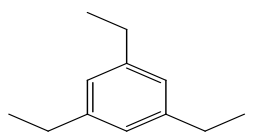
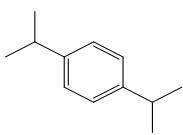
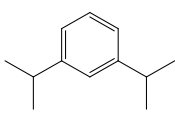
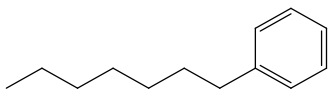
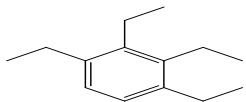
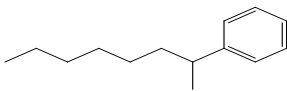
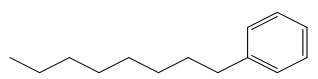
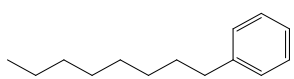
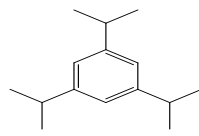
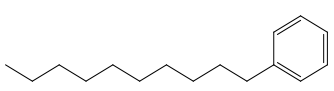
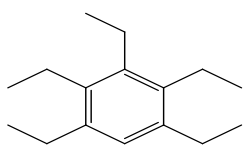
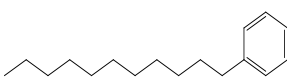
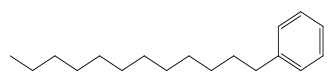
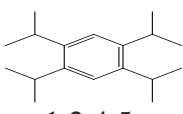
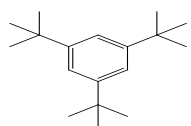
Présentations des données

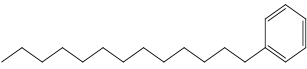
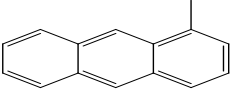
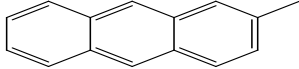
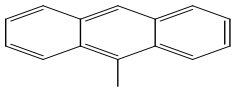
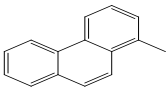
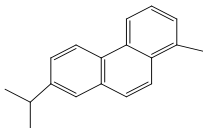
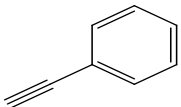
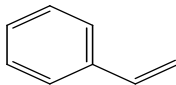
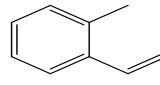
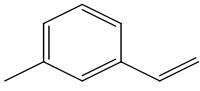
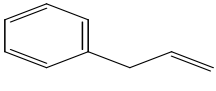
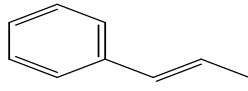
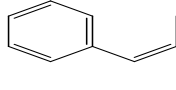
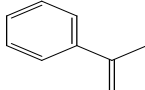
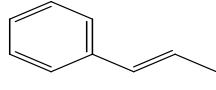
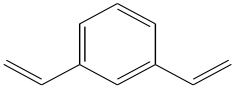
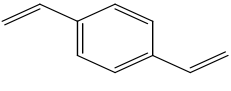
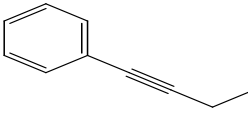
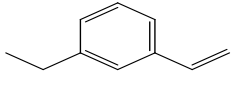
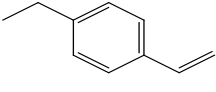
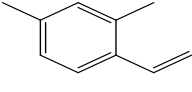
- **Tableau 1 : Les hydrocarbures non-saturés.**
- **Tableau 2 : Les n- alcanes.**
- **Tableau 3 : Les différentes classes de solvants.**


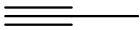
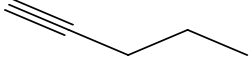
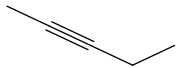
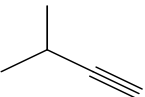
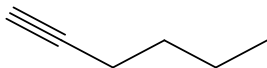
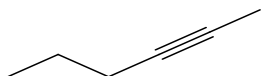

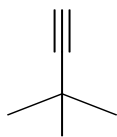
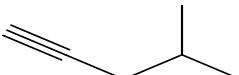
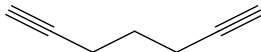
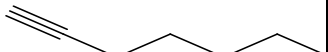
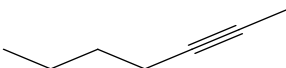

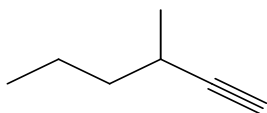
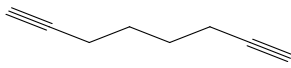
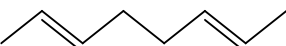
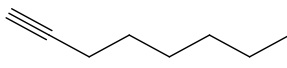
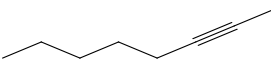

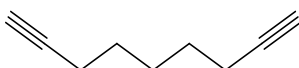
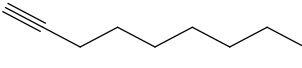
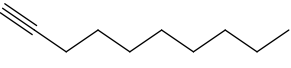
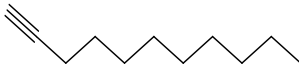
Tableau 1 : Les hydrocarbures non-saturés

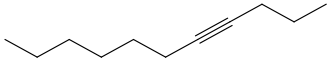
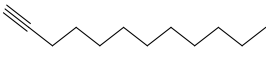
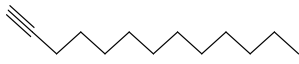

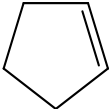
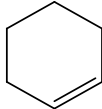
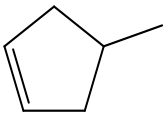
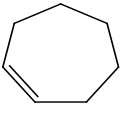
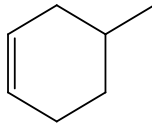
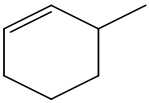
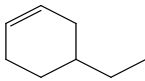

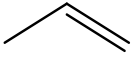
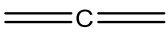
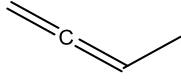
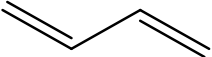
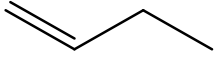
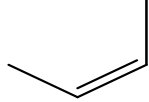
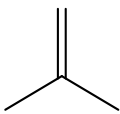
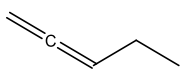
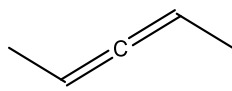
N°	Structure/Nom IUPAC/ N° de CAS	N°	Structure/Nom IUPAC/ N° de CAS	N°	Structure/Nom IUPAC/ N° de CAS
1	 Benzène 71-43-2	2	 Toluène 108-88-3	3	 Éthylbenzène 100-41-4
4	 1,4-Xylène 106-42-3	5	 1,2-Xylène 95-47-6	6	 Propylbenzène 103-65-1
7	 Isopropylbenzène 98-82-8	8	 1-Éthyl-3-méthylbenzène 620-14-4	9	 1,2,3-Triméthylbenzène 526-73-8
10	 1,2,4-Triméthylbenzène 95-63-6	11	 1,3,5-Triméthylbenzène 108-67-8	12	 1-Éthyl-2-méthylbenzène 611-14-3
13	 1-Éthyl-4-méthylbenzène 622-96-8	14	 Naphthalène 91-20-3	15	 Butylbenzène 104-51-8
16	 1,2,4,5-Tétraméthylbenzène 95-93-2	17	 2-Éthyl-1,4-diméthylbenzène 1758-88-9	18	 1-Éthyl-2,3-diméthylbenzène 933-98-2

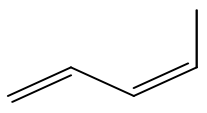
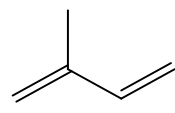
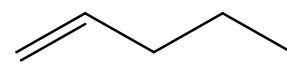
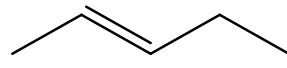
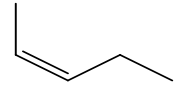
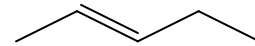
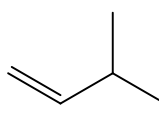
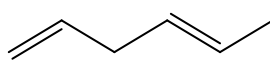
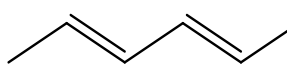
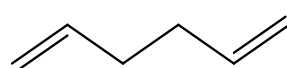
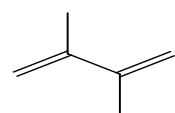
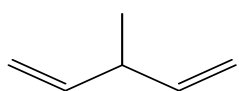
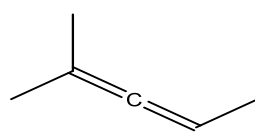
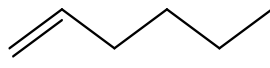
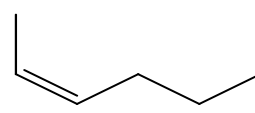
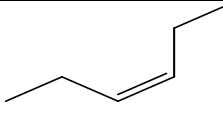
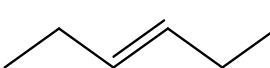
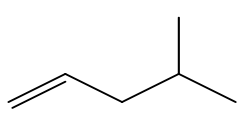
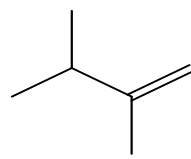
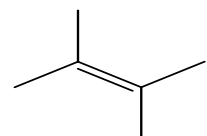
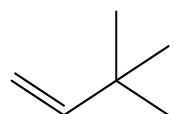
19		20		21	
	1-Éthyl-2,4-diméthylbenzène 874-41-9		tert-Butylbenzène 98-06-6		1-Méthyl-4-propan-2-ylbenzène 99-87-6
22		23		24	
	1,2-Diéthylbenzène 135-01-3		1,3-Diéthylbenzène 141-93-5		1,4-Diéthylbenzène 105-05-5
25		26		27	
	4-Ethyl-1,2-diméthylbenzène 934-80-5		1-Méthylnaphtalène 90-12-0		Pentylbenzène 538-68-1
28		29		30	
	(3-Méthylbutyl)benzène 2049-94-7		1,2,3,4,5-Pentaméthylbenzène 700-12-9		1-Méthyl-4-(2-méthyl-2-propanyl)benzène 98-51-1
31		32		33	
	2-Méthylbutan-2-ylbenzène 2049-95-8		1-Éthylnaphtalène 1127-76-0		2-Éthylnaphtalène 939-27-5
34		35		36	
	1,3-Diméthylnaphtalène 575-41-7		1,2-Diméthylnaphtalène 573-98-8		Hexylbenzène 1077-16-3

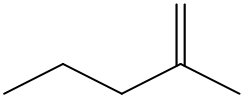
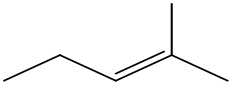
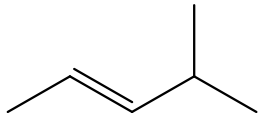
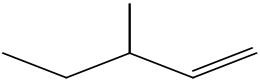
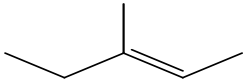
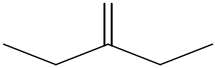
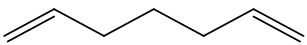
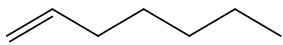
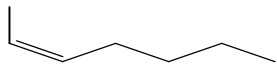
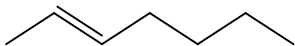
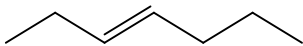
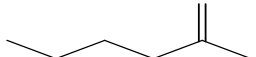
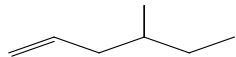
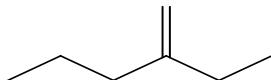
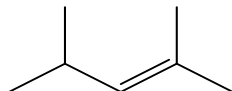
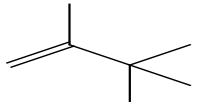
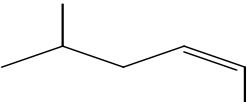
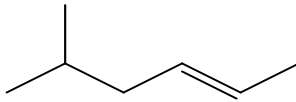
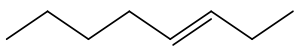
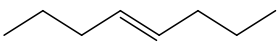
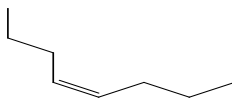
37		38		39	
	1,2,3,4,5,6- Hexaméthylbenzène 87-85-4		1,3-Diméthyl-5-(2- méthyl-2-propanyl) benzène 98-19-1		1,2,4-Triméthylbenzène 95-63-6
40		41		42	
	1,3,5-Triéthylbenzène 102-25-0		1,4-Di(propan-2- yl)benzène 100-18-5		1,3-Di(propan-2- yl)benzène 99-62-7
43		44		45	
	Heptylbenzène 1078-71-3		1,2,3,4- Tétraéthylbenzène 642-32-0		2-Octanylbenzène 777-22-0
46		47		48	
	Octylbenzène 2189-60-8		Nonylbenzène 1081-77-2		1,3,5- Triisopropylbenzène 717-74-8
49		50		51	
	Décylbenzène 104-72-3		1,2,3,4,5- Pentaéthylbenzène 605-01-6		Undecylbenzène 6742-54-7
52		53		54	
	Dodécylbenzène 123-01-3		1,2,4,5- Tétraisopropylbenzèn e 635-11-0		1,3,5-Tri(2-méthyl-2- propanyl)benzène 1460-02-2

55		56		57	
	Tridécylbenzène 123-02-4		1-Méthylantracène 610-48-0		2-Méthylantracène 613-12-7
58		59		60	
	9-Méthylantracène 779-02-2		1-Méthylphenanthrène 832-69-9		1-Méthyl-7-propan-2-ylphenanthrène 483-65-8
61		62		63	
	Éthynylbenzène 536-74-3		Éthenylbenzène 100-42-5		1-Éthenyl-2-méthylbenzène 611-15-4
64		65		66	
	1-Éthenyl-3-méthylbenzène 100-80-1		Prop-2-enylbenzène 300-57-2		[(E)-Prop-1-enyl]benzène 637-50-3
67		68		69	
	[(Z)-Prop-1-enyl]benzène 766-90-5		Isopropénylbenzène 98-83-9		Trans-1-phényl-1-propène 873-66-5
70		71		72	
	1,3-bis(Éthenyl)benzène 108-57-6		1,4-bis(Éthenyl)benzène 105-06-6		1-Butyn-1-ylbenzène 622-76-4
73		74		75	
	1-Éthenyl-3-éthylbenzène 7525-62-4		1-Éthenyl-4-éthylbenzène 3454-07-7		1-Éthenyl-2,4-diméthylbenzène 2234-20-0

76	 Éthyne 74-86-2	77	 Propyne 74-99-7	78	 Pent-1-yne 627-19-0
79	 Pent-2-yne 627-21-4	80	 3-Méthylbut-1-yne 598-23-2	81	 Hex-1-yne 693-02-7
82	 Hex-2-yne 764-35-2	83	 Hex-3-yne 928-49-4	84	 3,3-Diméthylbut-1-yne 917-92-0
85	 4-Méthylpent-1-yne 7154-75-8	86	 Hepta-1,6-diyne 2396-63-6	87	 Hept-1-yne 628-71-7
88	 Hept-2-yne 1119-65-9	89	 Hept-3-yne 2586-89-2	90	 3-Méthylhex-1-yne 40276-93-5
91	 Octa-1,7-diyne 871-84-1	92	 2,6-Octadiene 18152-31-3	93	 Oct-1-yne 629-05-0
94	 Oct-2-yne 2809-67-8	95	 Oct-4-yne 1942-45-6	96	 Nona-1,8-diyne 2396-65-8
97	 Non-1-yne 3452-09-3	98	 Dec-1-yne 764-93-2	99	 Undec-1-yne 2243-98-3

100		101		102	
	Undec-4-yne 60212-31-9		Dodec-1-yne 765-03-7		Tridec-1-yne 26186-02-7
103		104		105	
	Cyclobutène 822-35-5		Cyclopentène 142-29-0		Cyclohexène 110-83-8
106		107		108	
	4-Méthylcyclopentène 1759-81-5		Cycloheptène 628-92-2		3-Méthylcyclohexène 591-48-0
109		110		111	
	3-Méthylcyclohexène 591-48-0		4-Éthylcyclohexène 3742-42-5		Éthène 74-85-1
112		113		114	
	Prop-1-ène 115-07-1		Propa-1,2-diène 463-49-0		Buta-1,2-diène 590-19-2
115		116		117	
	Buta-1,3-diène 106-99-0		But-1-ène 106-98-9		(Z)-but-2-ène 590-18-1
118		119		120	
	2-Méthylpropène 115-11-7		Penta-1,2-diène 591-95-7		Penta-2,3-diène 591-96-8

121	 (3Z)-penta-1,3-diène 1574-41-0	122	 2-Méthylbuta-1,3-diène 78-79-5	123	 Pent-1-ène 109-67-1
124	 Pent-2-ene 627-20-3	125	 (Z)-Pent-2-ene 627-20-3	126	 (E)-Pent-2-ene 646-04-8
127	 3-Méthylbut-1-ène 563-45-1	128	 (4E)-Hexa-1,4-diène 7319-00-8	129	 Hexa-2,4-diène 592-46-1
130	 Hexa-1,5-diène 592-42-7	131	 2,3-Diméthyl-1,3-butadiène 513-81-5	132	 3-Méthyl-1,4-pentadiène 1115-08-8
133	 2-Méthyl-2,3-pentadiène 3043-33-2	134	 Hex-1-ène 592-41-6	135	 (2Z)-Hex-2-ène 7688-21-3
136	 (3Z)-Hex-3-ène 7642-09-3	137	 (3E)-Hex-3-ène 13269-52-8	138	 4-Méthylpent-1-ène 691-37-2
139	 2,3-Diméthylbut-1-ène 563-78-0	140	 2,3-Diméthylbut-2-ène 563-79-1	141	 3,3-Diméthylbut-1-ène 558-37-2

142		143		144	
	2-Méthylpent-1-ène 763-29-1		2-Méthylpent-2-ène 625-27-4		4-Méthylpent-2-ène 4461-48-7
145		146		147	
	3-Méthylpent-1-ène 760-20-3		(E)-3-méthylpent-2-ène 616-12-6		2-Éthylbut-1-ène 760-21-4
148		149		150	
	Hepta-1,6-diène 3070-53-9		Hept-1-ène 592-76-7		(Z)-Hept-2-ène 6443-92-1
151		152		153	
	(E)-Hept-2-ène 14686-13-6		(E)-Hept-3-ène 14686-14-7		2-Méthylhex-1-ène 6094-02-6
154		155		156	
	4-Méthylhex-1-ène 3769-23-1		3-Méthylidenehexane 3404-71-5		2,4-Diméthylpent-2-ène 625-65-0
157		158		159	
	2,3,3-Triméthylbut-1-ène 594-56-9		(Z)-5-Méthylhex-2-ène 13151-17-2		(E)-5-Méthylhex-2-ène 7385-82-2
160		161		162	
	(E)-Oct-3-ène 14919-01-8		(E)-Oct-4-ène 14850-23-8		(Z)-Oct-4-ène 7642-15-1


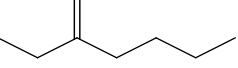
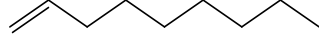
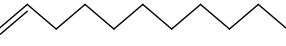

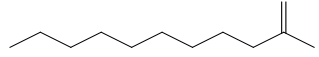

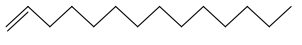



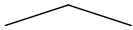

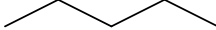


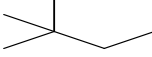
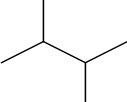
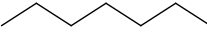

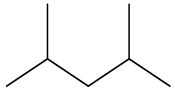
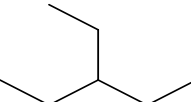
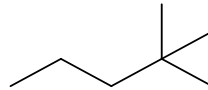
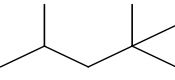
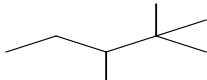
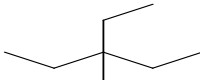
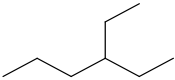
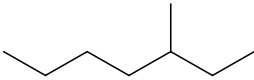
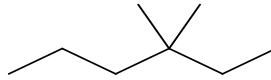
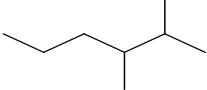
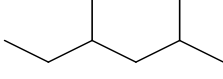
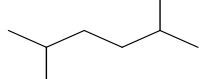
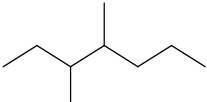
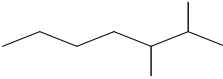
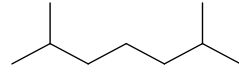
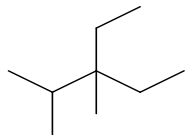
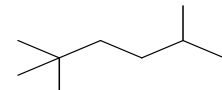
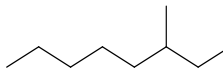
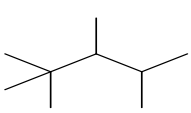
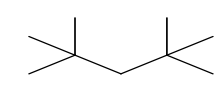
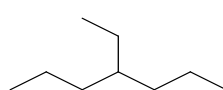
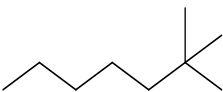
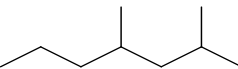
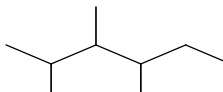
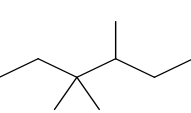
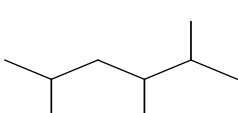
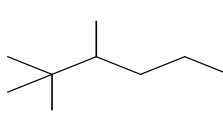
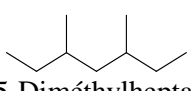
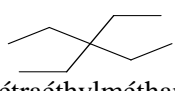
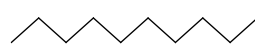
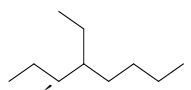
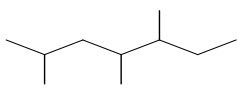
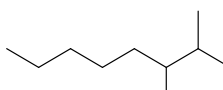
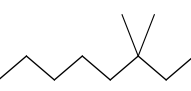
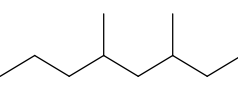
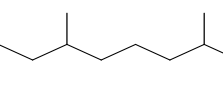
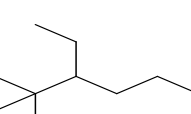
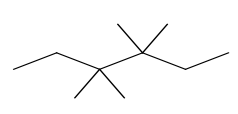
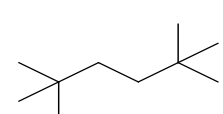
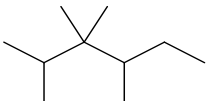
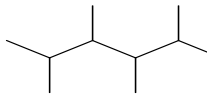
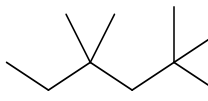
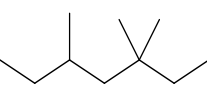
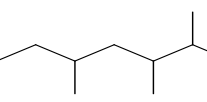
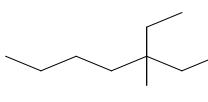
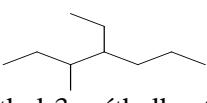
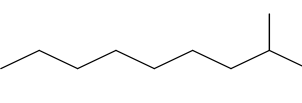


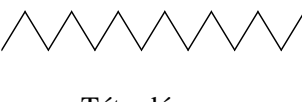
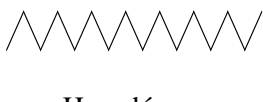
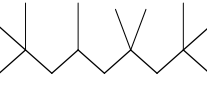
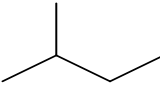
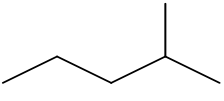
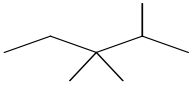
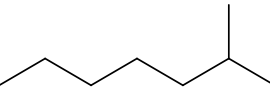
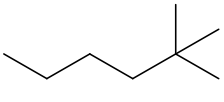
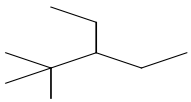
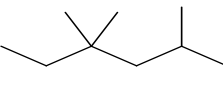
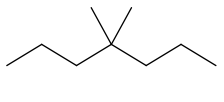
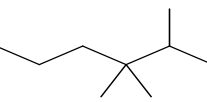
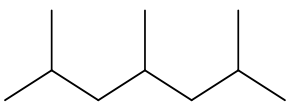
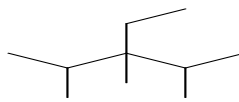
163	 Nona-1,8-diene 4900-30-5	164	 2-Éthylhex-1-ène 1632-16-2	165	 Non-1-ène 124-11-8
166	 Undec-1-ène 821-95-4	167	 Dodec-1-ène 112-41-4	168	 2-Méthylundec-1-ène 18516-37-5
169	 Tridec-1-ène 2437-56-1	170	 Tétradec-1-ène 1120-36-1	171	 Pentadec-1-ène 13360-61-7
172	 Hexadec-1-ène 629-73-2	173	 Heptadec-1-ène 6765-39-5		/

Tableau 2 : Les composés n-alcanes

N°	Structure/Nom IUPAC/ N° de CAS	N°	Structure/Nom IUPAC/ N° de CAS	N°	Structure/Nom IUPAC/ N° de CAS
1	 Propane 74-98-6	2	 Butane 106-97-8	3	 Pentane 109-66-0
4	 2,2-Diméthylpropane 463-82-1	5	 Hexane 110-54-3	6	 2,2-Diméthylbutane 75-83-2
7	 2,3-Diméthylbutane 79-29-8	8	 Heptane 142-82-5	9	 3,3-Diméthylpentane 562-49-2
10	 2,4-Diméthylpentane 108-08-7	11	 3-Éthylpentane 617-78-7	12	 2,2-Diméthylpentane 590-35-2
13	 2,2,4-Triméthylpentane 540-84-1	14	 2,2,3-Triméthylpentane 564-02-3	15	 3-Éthyl-3-méthylpentane 1067-08-9
16	 3-Éthylhexane 619-99-8	17	 3-Méthylheptane 589-81-1	18	 3,3-Diméthylhexane 563-16-6
19	 2,3-Diméthylhexane 584-94-1	20	 2,4-Diméthylhexane 589-43-5	21	 2,5-Diméthylhexane 592-13-2
22	 3,4-Diméthylheptane 922-28-1	23	 2,3-Diméthylheptane 3074-71-3	24	 2,6-Diméthylheptane 1072-05-5

25	 3-Éthyl-2,3-diméthylpentane 16747-33-4	26	 2,2,5-Triméthylhexane 3522-94-9	27	 3-Méthyl-octane 2216-33-3
28	 2,2,3,4-Tétraméthylpentane 1186-53-4	29	 2,2,4,4-Tétraméthylpentane 1070-87-7	30	 4-Éthylheptane 2216-32-2
31	 2,2-Diméthylheptane 1071-26-7	32	 2,4-Diméthylheptane 2213-23-2	33	 2,3,4-Triméthylhexane 921-47-1
34	 3,3,4-Triméthylhexane 16747-31-2	35	 2,3,5-Triméthylhexane 1069-53-0	36	 2,2,3-Triméthylhexane 16747-25-4
37	 3,5-Diméthylheptane 926-82-9	38	 Tétraméthylméthane 1067-20-5	39	 Décane 124-18-5
40	 4-Éthyl-octane 15869-86-0	41	 2,4,5-Triméthylheptane 20278-84-6	42	 2,3-Diméthyl-octane 7146-60-3
43	 3,3-Diméthyl-octane 4110-44-5	44	 3,5-Diméthyl-octane 15869-93-9	45	 2,6-Diméthyl-octane 2051-30-1
46	 3-Éthyl-2,2-diméthylhexane 20291-91-2	47	 3,3,4,4-Tétraméthylhexane 5171-84-6	48	 2,2,5,5-Tétraméthylhexane 1071-81-4

49	 2,3,3,4-Tétraméthylhexane 52897-10-6	50	 2,3,4,5-Tétraméthylhexane 52897-15-1	51	 2,2,4,4-Tétraméthylhexane 51750-65-3
52	 3,3,5-Triméthylheptane 7154-80-5	53	 2,3,5-Triméthylheptane 20278-85-7	54	 3-Éthyl-3-méthylheptane 17302-01-1
55	 4-Éthyl-3-méthylheptane 52896-89-6	56	 2-Méthylnonane 63335-88-6	57	 Undécane 1120-21-4
58	 Dodécane 112-40-3	59	 Tétradécane 629-59-4	60	 Hexadécane 544-76-3
61	 2,2,4,4,6,8,8-Heptaméthylnonane 4390-04-9	62	 2-Méthylbutane 78-78-4	63	 2-Méthylpentane 107-83-5
64	 2,3,3-Triméthylpentane 560-21-4	65	 2-Méthylheptane 592-27-8	66	 2,2-Diméthylhexane 590-73-8
67	 3-Éthyl-2,2-diméthylpentane 16747-32-3	68	 2,4,4-Triméthylhexane 16747-30-1	69	 4,4-Diméthylheptane 1068-19-5
70	 2,3,3-Triméthylhexane 16747-28-7	71	 2,4,6-Triméthylheptane 2613-61-8	72	 3-Éthyl-2,3,4-triméthylpentane 52897-19-5

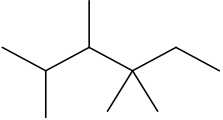
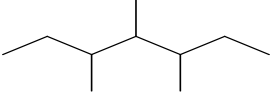
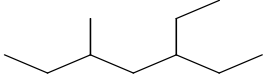
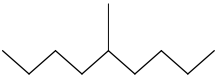
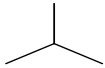
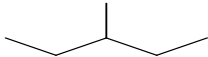
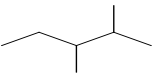
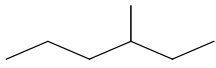
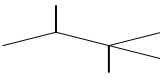
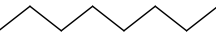
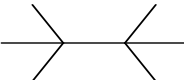
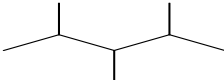
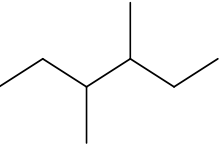
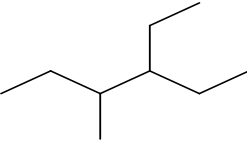
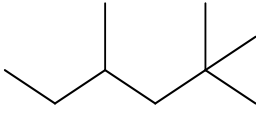
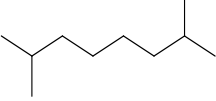
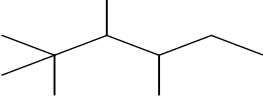



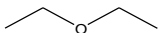
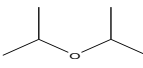

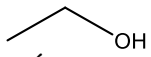
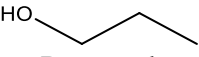
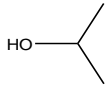
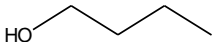
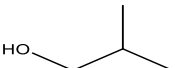
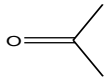
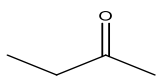
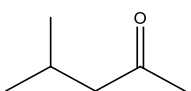
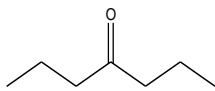
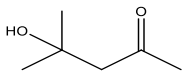

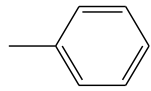
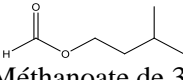
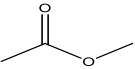
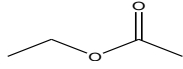
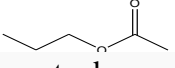
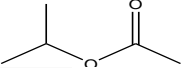
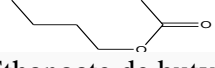
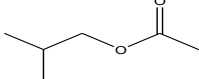
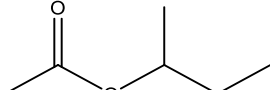
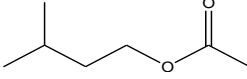

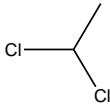
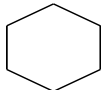
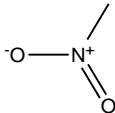
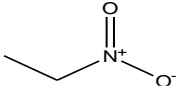
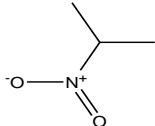
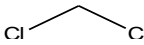

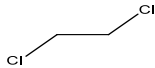
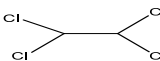


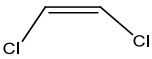
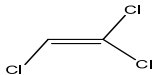
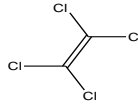
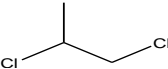
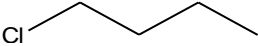
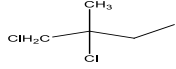
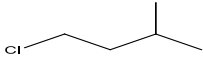
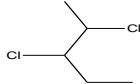
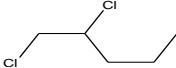
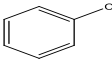
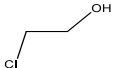
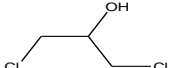
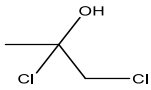
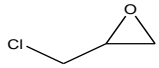
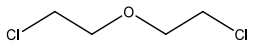
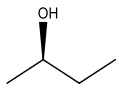
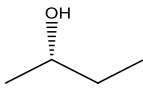
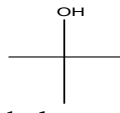
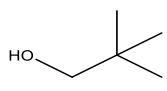
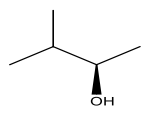
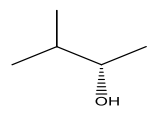
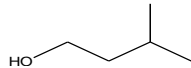
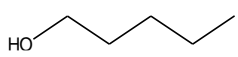
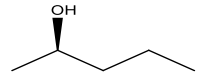
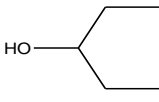
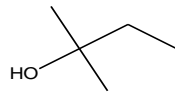
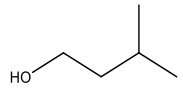
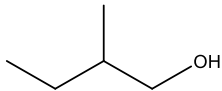
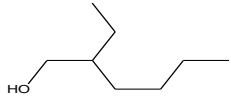
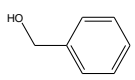
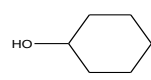
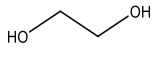
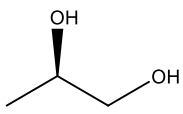
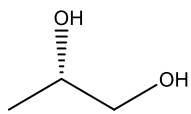
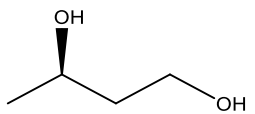
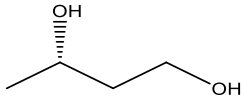
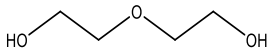
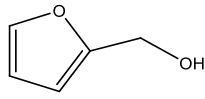
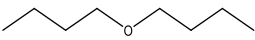
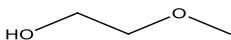
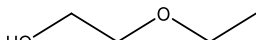
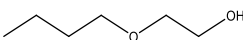

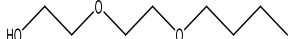
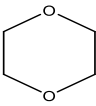
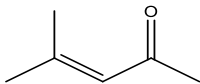
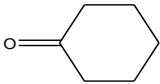
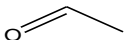
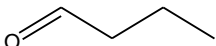
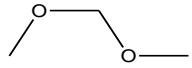
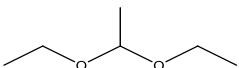
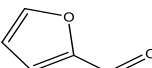
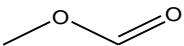
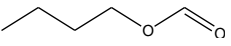
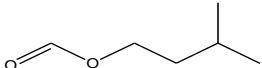
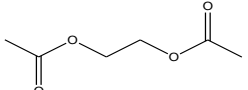
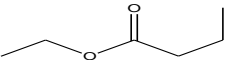
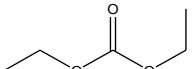
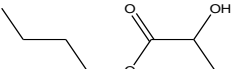
73		74		75	
	2,3,4,4-Tétraméthylhexane 52897-12-8		3,4,5-Triméthylheptane 20278-89-1		3-Éthyl-5-méthylheptane 52896-90-9
76		77		78	
	5-Méthylnonane 15869-85-9		2-Méthylpropane 75-28-5		3-Méthylpentane 96-14-0
79		80		81	
	2,3-Diméthylpentane 565-59-3		3-Méthylhexane 589-34-4		2,2,3-Triméthylbutane 464-06-2
82		83		84	
	Octane 111-65-9		2,2,3,3- Tétraméthylbutane 594-82-1		2,3,4-Triméthylpentane 565-75-3
85		86		87	
	3,4-Diméthylhexane 583-48-2		3-Éthyl-4-méthylhexane 3074-77-9		2,2,4-Triméthylhexane 16747-26-5
88		89		90	
	2,7-Diméthyloctane 1072-16-8		2,2,3,4- Tétraméthylhexane 52897-08-2		Tridecane 29-50-5
91		92		/	/
	Nonane 111-84-2		Pentadécane 629-62-9		

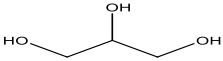
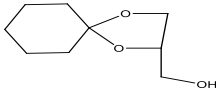
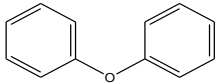
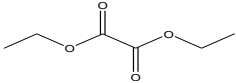
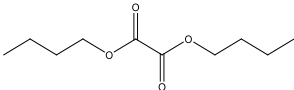
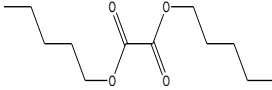
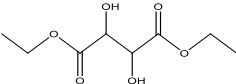
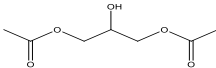
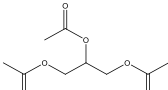
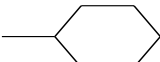
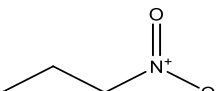
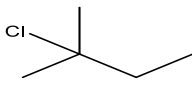
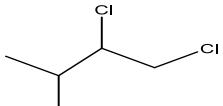
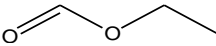
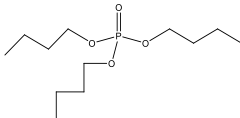
Tableau 3 : Les différentes classes de solvants

N°	Structure/Nom IUPAC/ N° de CAS	N°	Structure/Nom IUPAC/ N° de CAS	N°	Structure/Nom IUPAC/ N° de CAS
1	 Éthoxyéthane 60-29-7	2	 2-Isopropoxypropane 108-20-3	3	 Méthanol 67-56-1
4	 Éthanol 64-17-5	5	 Propanol 71-23-8	6	 Propan-2-ol 67-63-0
7	 Butan-1-ol 71-36-3	8	 2-Méthylpropan-1-ol 78-83-1	9	 Propan-2-one 67-64-1
10	 Butan-2-one 78-93-3	11	 4-Méthylpentan-2-one 108-10-1	12	 Heptan-4-one 123-19-3
13	 4-Hydroxy-4- méthylpentan-2-one 123-42-2	14	 Benzène 71-43-2	15	 Méthylbenzène 108-88-3
16	 Méthanoate de 3- méthylbutyle 110-45-2	17	 Éthanoate de méthyle 79-20-9	18	 Éthanoate d'éthyle 141-78-6
19	 Éthanoate de propyle 109-60-4	20	 Éthanoate d'isopropyle 108-21-4	21	 Éthanoate de butyle 123-86-4
22	 Éthanoate de 2- méthylpropyle 110-19-0	23	 Éthanoate de sec-butyle 105-46-4	24	 Éthanoate de 3- méthylbutyle 123-92-2

25		26		27	
	Trichlorométhane 67-66-3		1,1-Dichloroéthane 75-34-3		Cyclohexane 110-82-7
28		29		30	
	Nitrométhane 75-52-5		Nitroéthane 79-24-3		2-Nitropropane 79-46-9
31		32		33	
	Dichlorométhane 75-09-2		Tétrachlorométhane 56-23-5		1,2-Dichloroéthane 107-06-2
34		35		36	
	1,1,2,2-Tétrachloroéthane 79-34-5		1,1,1,2,2-Pentachloroéthane 76-01-7		(E)-1,2-Dichloroéthène 156-60-5
37		38		39	
	(Z)-1,2-Dichloroéthène 156-59-2		1,1,2-Trichloroéthène 79-01-6		1,1,2,2-Tétrachloroéthène 127-18-4
40		41		42	
	1,2-Dichloropropane 78-87-5		1-Chlorobutane 109-69-3		1,2-Dichloro-2-méthylbutane 23010-04-0
43		44		45	
	1-Chloro-3-méthylbutane 107-84-6		2,3-Dichloropentane 600-11-3		1,2-Dichloropentane 1674-33-5
46		47		48	
	Chlorobenzène 108-90-7		2-Chloroéthanol 107-07-3		1,3-Dichloropropan-2-ol 96-23-1

49		50		51	
	1,2-Dichloropropan-2-ol 52515-75-0		2-(Chlorométhyl)oxirane 106-89-8		1-Chloro-2-(2-chloroéthoxy)éthane 111-44-4
52		53		54	
	(R)-Butan-2-ol 17898-79-4		(S)-Butan-2-ol 4221-99-2		2-Méthylpropan-2-ol 75-65-0
55		56		57	
	2,2-Diméthylpropan-1-ol 75-84-3		(R)-3-Méthylbutan-2-ol 1572-93-6		(S)-3-Méthylbutan-2-ol 1517-66-4
58		59		60	
	3-Méthylbutan-1-ol 123-51-3		Pentan-1-ol 71-41-0		(R)-Pentan-2-ol 31087-44-2
61		62		63	
	(S)-Pentan-2-ol 26184-62-3		Pentan-3-ol 584-02-1		2-Méthylbutan-2-ol 75-85-4
64		65		66	
	3-Méthylbutan-1-ol 123-51-3		2-Méthylbutan-1-ol 137-32-6		2-Éthylhexan-1-ol 104-76-7
67		68		69	
	Phénylméthanol 100-51-6		Cyclohexanol 108-93-0		Éthane-1,2-diol 107-21-1
70		71		72	
	(R)-Propane-1,2-diol 4254-14-2		(S)-Propane-1,2-diol 4254-15-3		(R)-Butane-1,3-diol 6290-03-5

73		74		75	
	(S)-butane-1,3-diol 24621-61-2		3-Oxapentane-1,5-diol 111-46-6		Furan-2-ylmethanol 98-00-0
76		77		78	
	1-Butoxybutane 142-96-1		2-Méthoxyéthanol 109-86-4		2-Éthoxyéthanol 110-80-5
79		80		81	
	2-Butoxyéthanol 111-76-2		2-(Éthoxyéthoxy)éthanol 111-90-0		2-(2-Butoxyéthoxy)éthanol 112-34-5
82		83		84	
	1,4-Dioxane 123-91-1		4-Méthylpent-3-ène-2-one 141-79-7		Cyclohexanone 108-94-1
85		86		87	
	Éthanal 75-07-0		Butanal 123-72-8		Diméthoxyméthane 109-87-5
88		89		90	
	1,1-Diéthoxyéthane 105-57-7		Furan-2-carbaldéhyde 98-01-1		Méthanoate de méthyle 107-31-3
91		92		93	
	Méthanoate de butyle 592-84-7		Méthanoate de 3-méthylbutyl 110-45-2		Éthane-1,2-diyldiacétate 111-55-7
94		95		96	
	Butanoate d'éthyle 105-54-4		Carbonate de diéthyle 105-58-8		2-Hydroxypropanoate de butyle 138-22-7

97	 Propane-1,2,3-triol 56-81-5	98	 1,4-Dioxaspiro[4.5]dec-2-ylméthanol 4167-35-5	99	 Phénoxybenzène 101-84-8
100	 Oxalate de diéthyle 95-92-1	101	 Oxalate de dibutyle 2050-60-4	102	 Oxalate de dipentyle 20602-86-2
103	 Diéthyl 2,3-dihydroxybutanedioate 87-91-2	104	 (3-Acetyloxy-2-hydroxypropyl)acetate 105-70-4	105	 Acétate de 1,3-diacétyloxypropan-2-yle 102-76-1
106	 Méthylcyclohexane 108-87-2	107	 1-Nitro-propane 108-03-2	108	 2-Chloro-2-méthylbutane 594-36-5
109	 1,2-Dichloro-3-méthylbutane 600-10-2	110	 Méthanoate d'éthyle 109-94-4	111	 Phosphate de tributyle 126-73-8



Annexe II



Article publié

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Modeling and Prediction of Flash Point of Unsaturated Hydrocarbons Using Hybrid Genetic Algorithm/Multiple Linear Regression Approach.

Mabrouka DIDI, Hamza HADDAG, Youcef DRIOUCHE, and Djelloul MESSADI*.

Environmental and Food Safety Laboratory, Badji Mokhtar Universit , Annaba 23000,Algeria

ABSTRACT

A quantitative structure property relationship (QSPR) study is developed using Genetic Algorithm (GA) / Multiple Linear Regression (MLR) for modeling the flash points of 173 unsaturated hydrocarbons, using theoretical molecular descriptors derived from DRAGON software. The studied dataset was randomly separated into two independent subsets: a training set of 139 compounds to build the model and a test set of the removed 34 compounds to validate its predictive ability. The selection of a minimum set of meaningful descriptors was carried out using Genetic Algorithm in the MOBYDIGS Todeschini software. An MLR model of 4 descriptors with a high predictive power was developed for the prediction of the flash points of unsaturated hydrocarbons. The predictive ability of the obtained model was verified using a set of criteria according to Golbraikh and Tropsha and its applicability domain was studied using Willians plot.

Keywords: Flash point; Unsaturated hydrocarbons; Multiple linear regression; Quantitative structure-property relationship; Model prediction.

**Corresponding author*

INTRODUCTION

The flash point (FP) is defined as the lowest temperature, corrected to 101.3 k Pa, at which an application of an ignition source causes the vapors of the specimen to ignite under specific conditions of a test [1-4].

This parameter gives the knowledge necessary for understanding the fundamental physical and chemical processes of combustion. Moreover, it is of importance in practice for safety conditions in the storing, the processing and the handling of a given compound. And it is one of the major flammability characteristics used to assess the fire and explosion hazards of organic compounds [5].

The flash point of most compounds can be measured by two currently accepted experimental methods, which are the closed cup test and the open cup test [6]. However, for many other compounds, the experimental flash point values are scarce and too expensive to obtain. Moreover, it is even more difficult to make the experimental determination of the flash point of toxic, volatile, explosive and radioactive compounds. Hence, the development of estimation methods which are desirably convenient for predicting the flash point is required.

There are many methods for prediction of FP in the literature. Vidal et al. have presented a review of the most important methods for the prediction of the flash point [7]. Mainly, prediction methods for this property can be categorized as the group contribution method (GCM), the principal component analysis (PCA) and the quantitative structure-property relationship (QSPR).

A simple correlation for predicting the flash point of a large data set consisting various types of cyclic and acyclic hydrocarbons including the studied compounds where the proposed method was based on the number of carbons and hydrogen atoms and some specific molecular moieties, which can easily be used for any type of hydrocarbons [8].

Another method was introduced for the prediction of the flash point of different classes of unsaturated hydrocarbons showing that the number of carbons and hydrogen atoms can be used as a core function that may be revised by a correcting function. Correcting function contains two correcting terms that can be determined on the basis of molecular structure that can be determined on the basis of the molecular structure of unsaturated hydrocarbons [9].

The aim of this work is to build a new QSPR model that can be used for predicting flash points of 173 unsaturated hydrocarbons [9] from their molecular structure. In this work, after obtaining the most statistically significant descriptors by means of genetic algorithm (GA) based on variable selection approach, the multiple linear regression behavior of these molecular descriptors for predicting flash point of these compounds was studied.

MATERIALS AND METHODS

The data set

The experimental flash point dataset was taken from literature [9].

The set of the studied compounds is formed of different classes of unsaturated hydrocarbons including alkenes, alkynes and aromatics. Flash point values are in a range from 137 to 451 K. The dataset was randomly divided into two groups, a training set of 139 compound and a test set of 34 compound.

Descriptor generation

The chemical structure of each compound was sketched on a PC using the Hyperchem program [10] and preoptimized using MM+ molecular mechanics method (Polack-Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi-empirical PM3 method at a restricted Hartree-Fock level with no configuration interaction, applying a gradient norm limit of $0.01 \text{ kcal} \cdot \text{Å}^{-1} \cdot \text{mol}^{-1}$ as a stopping criterion.

The output files exported from Hyperchem were transferred into Dragon software [11], to calculate a large number of molecular descriptors on the basis of the geometrical and electronic structure of the molecules. Constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (when there was more than 95 % pairwise correlation, one variable was deleted), and the genetic algorithm was applied for variables selection to a final set of 269 descriptor.

Data splitting

In order to check the predictive capability of the proposed model, before model generation the data set was randomly split into a training set of 139 compounds from which the model is built and an external test set of 34 compounds on which to evaluate its prediction power as it is shown on Table 1.

Table 1: Names, structures and FP values of the studied set

N°	Names	FP(K)	N°	Names	FP(K)
1	Benzene	262	88	2-Heptyne	275
2*	Toluene	280	89	3-Heptyne	257
3*	Ethylbenzene	288	90*	3-Methyl-1-hexyne	268
4*	P-Xylene	300	91	1,7-Octadiyne	296
5	O-Xylene	303	92	2,6-Octadiene	307
6	Propylbenzene	303	93	1-Octyne	289
7	Cumene	304	94	2-Octyne	301
8	m-Ethyltoluene	311	95*	4-Octyne	291
9	1,2,3-Trimethylbenzene	324	96*	1,8-Nonadiyne	314
10*	1,2,4-Trimethylbenzene	321	97	1-Nonyne	306
11	1,3,5-Trimethylbenzene	317	98	1-Decyne	323
12	o-Ethyltoluene	312	99	1-Undecyne	338
13*	p-Ethyltoluene	309	100	4-Undecyne	341
14	Naphthalene	360	101	1-Dodecyne	352
15	Butylbenzene	331	102	1-Tridecyne	366
16	1,2,4,5-Tetramethylbenzene	346	103	Cyclobutene	202
17	2-Ethyl-p-xylene	329	104	Cyclopentene	244
18*	3-Ethyl-o-xylene	338	105	Cyclohexene	256
19*	4-Ethyl-m-xylene	330	106	4-Methylcyclopentene	243
20*	tert-Butylbenzene	307	107	Cycloheptene	267
21	P-Cymene	320	108	4-Methylcyclohexene	272
22	o-DiEthylbenzene	322	109*	3-Methylcyclohexene	270
23	m-DiEthylbenzene	324	110	4-Ethylcyclohexene	286
24	p-DiEthylbenzene	328	111	Ethylene	137
25	4-Ethyl-1,2-dimethylbenzene	331	112	Propene	165
26	1-Methylnaphtalene	355	113	Propadiene	177
27	n-Pentylbenzene	339	114	1,2-Butadiene	197
28	IsoPentylbenzene	335	115	1,3-Butadiene	197
29	Pentamethylbenzene	364	116*	Butene	194
30	p-tert-Butyltoluene	321	117	Cis-2-Butene	200
31	2-Phenyl-2methylbutane	338	118	Isobutylene	197
32	1-Ethylnaphtalene	380	119	1,2-Pentadiene	233
33	2-Ethylnaphtalene	377	120	2,3-Pentadiene	235
34	1,3-Dimethylnaphtalene	382	121	Cis-1,3-Pentadiene	232
35*	1,2-Dimethylnaphtalene	374	122*	2-Methylbutadiene	225
36	Hexylbenzene	356	123*	Pentene	229
37	Hexamethylbenzene	377	124	2-Pentene	253
38	3,5-Dimethyl-tert-butylbenzene	357	125	Cis-2-Pentene	227
39	1,2,4-Trimethylbenzene	349	126	trans-2-Pentene	225

N°	Names	FP	N°	Names	FP
40*	1,3,5-Triethylbenzene	354	127	Isopentene	211
41	1,4-Diisopropylbenzene	354	128	1,4,-Hexadiene	248
42	m-Diisopropylbenzene	350	129*	2,4-Hexadiene	264
43	n-Heptylbenzene	368	130	1,5-Hexadiene	246
44	1,2,3,4-Tetraethylbenzene	367	131	2,3-Dimethyl-1,3-butadiene	251
45	2-Phenyltoluene	373	132	3-Methyl-1,4-pentadiene	239
46	n-Octylbenzene	380	133	2-Methyl-2,3-pentadiene	255
47*	n-Nonylbenzene	390	134	1-Hexene	253
48	1,3,5-Triisopropylbenzene	359	135	Cis-2-Hexene	252
49	Decylbenzene	380	136*	Cis-3-Hexene	261
50	Pentaethylbenzene	386	137	Trans-3-Hexene	261
51*	n-Undecylbenzene	409	138*	Isohexene	241
52	Dodecylbenzene	418	139	2,3-Dimethyl-1-butene	255
53	1,2,4,5-tetraisopropylbenzene	397	140	2,3-Dimethyl-2-butene	256
54	1,3,5-Tri-tert-butylbenzene	372	141	3,3-Dimethyl-1-butene	244
55	Tridecylbenzene	385	142	2-Methyl-1-pentene	241
56*	1-Methylantracene	430	143	2-Methyl-2-pentene	246
57	2-Methylantracene	431	144	4-Methyl-2-pentene	241
58	9-Methylantracene	431	145	3-Methyl-1-pentene	244
59	1-methylphenanthrene	431	146	Trans-3-Methyl-2-pentene	266
60	7-isopropyl-1-methylphenanthrene	451	147*	2-Ethyl-1-butene	243
61	Phenylacetylene	303	148	1,6-Heptadiene	263
62	Styrene	304	149	1-Heptene	264
63*	2-Vinyltoluene	320	150	Cis-2-Heptene	265
64	3-Vinyltoluene	324	151	Trans-2-Heptene	267
65	3-Phenyl-1-propene	310	152	Trans-3-Heptene	266
66	beta-Methylstyrene	333	153	2-Methyl-1-hexene	267
67*	Cis-1-Propenylbenzene	325	154*	4-Methyl-1-hexene	258
68	Isopropenylbenzene	313	155*	2-Ethyl-1-pentene	263
69*	trans-1-phenyl-1-propene	331	156*	2,4-Dimethyl-2-pentene	264
70	m-Divinylbenzene	338	157	2,3,3-Trimethyl-1butene	256
71	p-Divinylbenzene	337	158*	Cis-5-Methyl-2-Hexene	268
72	1-Butenylbenzene	341	159	Trans-5-Methyl-2-Hexene	268
73	3-Ethylstyrene	333	160	Trans-3-Octene	282
74	4-Ethylstyrene	335	161	Trans-4-Octene	281
75	2,4-dimethyl-1-vinylbenzene	333	162	Cis-4-Octene	294
76	Acetylene	155	163*	1,8-Nonadiene	299
77	Propyne	186	164	2-Ethyl-1-hexene	279
78	1-Pentyne	230	165	1-Nonene	298
79	2-Pentyne	253	166*	1-Undecene	336
80	3-Methyl-1-butyne	221	167	Dodecene	351
81	1-Hexyne	252	168	2-Methyl-1-undecene	345
82*	2-Hexyne	263	169	1-Tridecene	352
83	3-Hexyne	259	170	1-Tetradecene	383
84	3,3-Dimethyl-1-butyne	239	171	1-Pentadecene	386
85	4-Methyl-1-Pentyne	249	172	1-Hexadecene	402
86	1,6-Heptadiyne	282	173	1-Heptadecene	408
87	1-Heptyne	263			

* compounds of the test set .

Model development and validation

Once the molecular descriptors are generated, multiple linear regression (MLR) analysis and variable selection were performed by the software Mobydigs [12] using the Ordinary Least Square (OLS) regression method and Genetic Algorithm _Variable Subset Selection (GA-VSS) [13].

The outcome of the application of the genetic algorithms is a population of 100 regression models, ordered according to their decreasing internal predictive performance, verified by Q^2 . First of all, models with 1-2 variables were developed by the all-subset-method procedure in order to explore all the low dimension combinations. The number of descriptors was subsequently increased one by one, and the new models were formed. The best models are selected at each rank, and the final model must be chosen from among them. This has to be sufficiently correlated and at the same time, protect against any over parametrization, which would lead to a loss of predictive power for molecular outside training set. From a statistical view point the ratio of the number of samples (n) to the number of descriptors (m) should not be too low. Usually, it is recommended that $n/m \geq 5$ [14]. The GA was stopped when increasing the model size did not increase the Q^2 value to any significant degree.

Particular attention was paid to the collinearity of the selected molecular descriptors by applying the QUIK (Q Under Influence of K) rule [15] a necessary condition for the model validity. Acceptable models are only with a global correlation of [X+Y] block (K_{xy}) greater than the global correlation of the X block (K_{xx}) variable, X being the molecular descriptors and Y the response variable. Therefore, when there were models of similar performance, those with higher ΔK ($K_{xy} - K_{xx}$) were selected and further verified.

The goodness of fit of the calculated models were assessed by the means of the multiple coefficient R^2 , and the standard deviation error in calculation (SDEC).

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

Cross validation techniques allow the assessment of internal predictivity (bootstrap) in addition to the robustness of model (Q_{LOO}^2 cross validation).

Cross validation methods consist in leaving out a given number of compounds from the training set and rebuilding the model, which is then used to predict the compounds left out. This procedure, is repeated for all compounds of the training set, obtaining a prediction for everyone. If each compound is taken away once each time the cross validation procedure is called leave-one-out technique (LOO technique). An LOO correlation coefficient, generally indicated with Q^2 , is computed by evaluating the accuracy of these "test" compounds prediction.

$$Q^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{PRESS}{TSS} \quad (2)$$

The "hat" of the variable y , as is the usual statistical notation, indicates that it is a predicted value of the studied property, and the sub index "i/i" indicates that the predicted values come from the model built without the predicted compound.

The predictive residual of squares (PRESS) measures the dispersion of the predicted values. It is used to define Q^2 and the standard error in prediction (SDEP).

$$\text{SDEP} = \sqrt{\text{PRESS}/n} \quad (3)$$

A value of $Q^2 \geq 0.5$ is generally considered satisfactory, and a value greater than 0.9 is excellent [16,17]. However, studies have indicated that while Q^2 is a necessary condition for high predictive power of a model, is not sufficient.

In bootstrap validation technique K n -dimensional groups are generated by a randomly repeated selection of n -objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 5000 times for each validated model [18].

Obtaining a robust model does not give real information about its prediction power. This is evaluated by predicting the compounds included in the test set.

The external Q^2_{ext} for the test set is determined [19] with the equation (4):

$$Q^2_{\text{ext}} = 1 - \frac{\sum_{i=1}^{n_{\text{ext}}} (\hat{y}_{i/i} - y_i)^2 / n_{\text{ext}}}{\sum_{i=1}^{n_{\text{tr}}} (y_i - \bar{y}_{\text{tr}})^2 / n_{\text{tr}}} \quad (4)$$

Where y_i and $\hat{y}_{i/i}$ are, respectively, the measured and predicted (over the prediction set) values of the dependent variable, and \bar{y} the averaged value of the dependent variable for the training set. n_{tr} and n_{ext} are the number of objects in the external set, respectively.

Other useful parameters are R^2 , calculated for the validation chemicals by applying the model developed on the training set, and an external standard deviation error of prediction (SDEP_{ext}), defined as:

$$\text{SDEP}_{\text{ext}} = \sqrt{\frac{1}{n_{\text{ext}}} \sum_{i=1}^{n_{\text{ext}}} (y_i - \bar{y})^2} \quad (5)$$

Where the sum runs over the test set objects (n_{ext}).

According to [20] a QSPR model is successful if it satisfies several criteria as follows:

$$R^2_{\text{CV}_{\text{ext}}} > 0.5 \quad (6)$$

$$r^2 > 0.6 \quad (7)$$

$$(r^2 - r^2_0) / r^2 < (r^2 - r^2_0) / r^2 < 0.1 \quad (8)$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15 \quad (9)$$

$$Ab = |r^2 - r_0'^2| < 0.3 \quad (10)$$

Here:

$$r = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (11)$$

$$r_0^2 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{t_0})}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \quad (12)$$

$$r_0'^2 = 1 - \frac{\sum (y_i - y_i^{t_0})^2}{\sum (y_i - \bar{y})^2} \quad (13)$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \quad (14)$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad (15)$$

$$T1 = \frac{(r^2 - r_0^2)}{r^2} \quad (16)$$

$$T2 = \frac{(r^2 - r_0'^2)}{r^2} \quad (17)$$

Where r^2 is the correlation between the calculated and the experimental values in the test set; r^2 (calculated versus observed values) and r'^2 (observed versus calculated values) are the coefficients of determination; k and k' are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively. $y_i^{t_0}$, $\tilde{y}_i^{t_0}$ are defined as $y_i^{t_0} = k\tilde{y}_i$ and $\tilde{y}_i^{t_0} = k'y$ and the summations run over the test set.

QSPR model Applicability Domain (AD)

The applicability domain ability (AD) [19,17] is a theoretical region in the space defined by the descriptors of the model and the method response, for which a given QSPR should make reliable predictions. In this work, the structural AD was verified by the leverage (hii) approach [21].

The warning leverage h^* is, generally, fixed at $3(m+1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation.

The presence of both the response outliers (Y outliers) and the structurally influential compounds (X outliers) was verified by the Williams plot [22]. The plot of standardized residuals versus leverage values.

RESULTS AND DISCUSSION

Several acceptable MLR models of different dimensions, based on various descriptors, were obtained. The best one, taking into account the parsimony principal regarding the complexity of the models, is a model of 4 descriptors with a high predictive power.

The equation (18) the optimal model is given as:

$$FP = - 235 + 12.5 \text{ nsK} + 416 \text{ FDI} - 83.3 \text{ Mor26v} + 19.4 \text{ R5u} \quad (18)$$

Here, nsK is a constitutional descriptor (block1); representing the number of non-hydrogen atoms [11].

FDI is folding degree index; belongs to the list of geometrical descriptors calculated by dragon (block12). Geometrical descriptors are defined in several different ways but always derived from the three-dimensional structure of the molecule. Generally, geometrical descriptors are calculated either on some optimized molecular geometry obtained by the methods of the computational chemistry or on crystallographic coordinates. The folding degree index is the largest eigen value of the distance/distance matrix, normalised dividing it by the number of atoms nAT. This index tends to one for linear molecules (of infinite length) and decreases in correspondence with the folding of the molecule. Thus, it can be thought of as a measure of the folding degree of the molecule because it indicates the degree of departure of a molecule from strict linearity [11].

Mor26v is a 3D-Morse descriptor (block 14); 3D-Molecule Representation of Structures based on Electron diffraction) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves. The following expression is used for 3D-MORSE descriptor calculation:

$$\text{Morsw} = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} W_i W_j \frac{\sin(s.r_{ij})}{s.r_{ij}} \quad (19)$$

Where Morsw is the scattered electron intensity, w is an atomic property, r_{ij} are the interatomic distances and nAT is the number of atoms. The term s represents the scattering in various directions by a collection of nAT atoms.

In order to obtain uniform length descriptors, the intensity distribution is made discrete, calculating its value at a sequence of evenly distributed values; in particular, in DRAGON, it is assumed that s takes integer values in the range 0 – 31 [11].

R5u (R autocorrelation of lag 5 / unweighted), is a GETAWAY descriptor (block16). GETAWAY descriptors have recently been proposed as chemical structure descriptors derived from a new representation of molecular structure [11].

The obtained statistical parameters are reported in table 2.

Table 2: Statistical parameters of the developed model

n_{tr}	n_{ext}	Q_{LOO}^2 (%)	R^2 (%)	R_{adj}^2 (%)	Q_{ext}^2	Q_{boot}^2
139	34	97.11	97.41	97.34	97.71	96.96
F	SDEC	SDEP	SDEP _{ext}	Kxy	Kxx	S
1261.62	10.09	10.66	9.50	49.77	34.57	10.28

The adjusted $R^2 (R_{adj}^2)$ is a better measure of the proportion of variance in the data explained by the correlation than R^2 , because R^2 is somewhat sensitive to changes in the number of samples of the training set and the number of descriptors involved in the correlation.

Statistical parameters show that the model (Eq.18) established a strong correlation between the selected variables and the studied property, characterized by an excellent coefficient of determination ($R^2 = 97.41\%$) that explains around 97.41% of data variation, in addition to a very large value of the Fisher F ($F=1261.62$), which indicates the excellence ability of the model in the prediction of FP values, and a good standard error ($s=10.28$). Equation (18) presents an R_{adj}^2 (%) =97.34 indicating excellent agreement between correlation and variation of the data.

The small difference between R^2 and Q_{LOO}^2 informs about the robustness of the model. The cross-validation prediction coefficient illustrates the reliability towards the elimination of the model focusing on the sensitivity towards the elimination of any 5 data. The value of Q_{boot}^2 (%) =96.96) confirms both the internal predictability and stability of the model.

A visual comparison of the predicted results of the new correlation with the experimental data is also shown in the plot of observed versus predicted values of FP (Figure 1) for the training and test sets confirmed that a linear model with very good fitting can be used to predict our studied property .

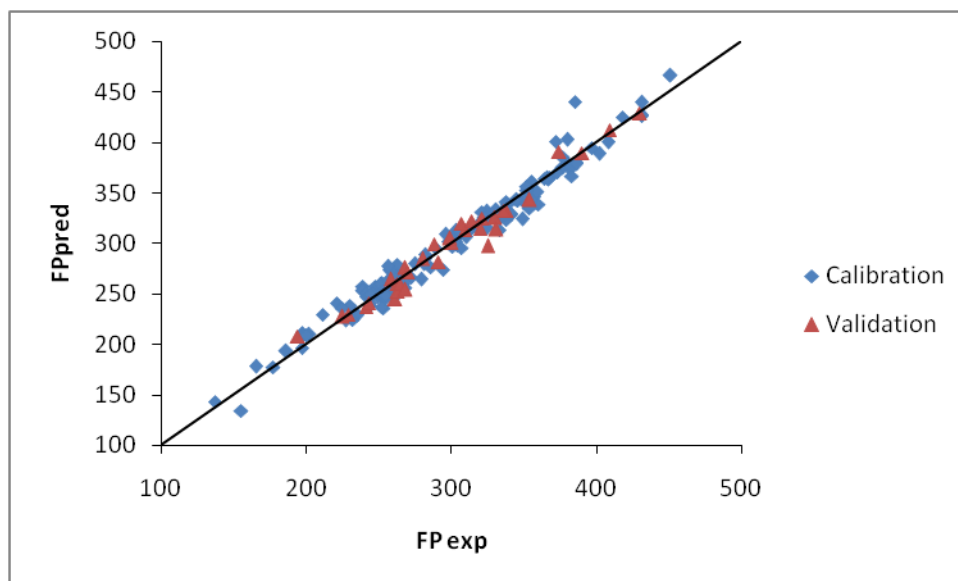


Figure 1: Experimental versus Predicted FP for the training and test sets.

Figure 2 represents the graph of statistical coefficients Q^2 and R^2 which allows comparing the results for randomized patterns (sign +) to the starting model (triangle) which is the real model. It is clear that the flash points statistics obtained for the modified vectors are smaller than those of the real QSPR model, to ensure that a real structure / property (FP) relationship has been established (Figure 2).

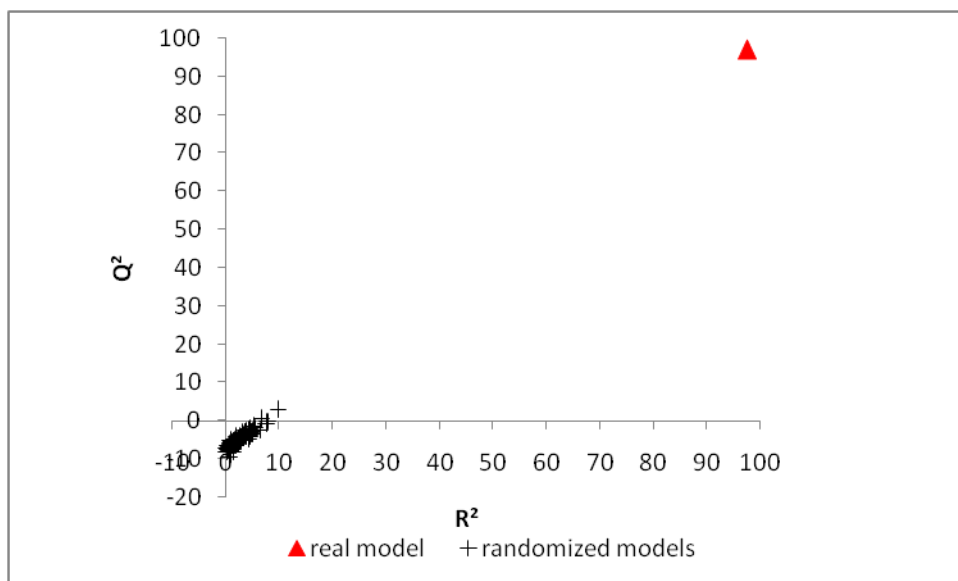


Figure 2: Randomization test associated to previous QSPR model

Signs + represent the randomly ordered flash points, and the triangle corresponds to the real flash point. The statistics for the modified FP vectors are clearly lower than the real QSPR model. Q^2 values are lower than 10 %, and for the major part one obtains even $Q^2 < 0$ for random models symbolized by sign +. This ensures that a real structure –property relationship has been found out.

Based on a previously described procedure [23], the relative contribution of the four descriptors to the model were determined as follow: $nsk (45.82\%) > FDI (18.55\%) > Mor26v (18.34\%) > R5u (17.29\%)$. As it is seen the nsk contribution is greater than FDI , $Mor26v$ and $R5u$ contributions, while the difference in the descriptor contribution is not significant, indicating that the nsk (number of carbon atoms) descriptor is more necessary in generating the predictive model than the other descriptors as seen on figure 3.

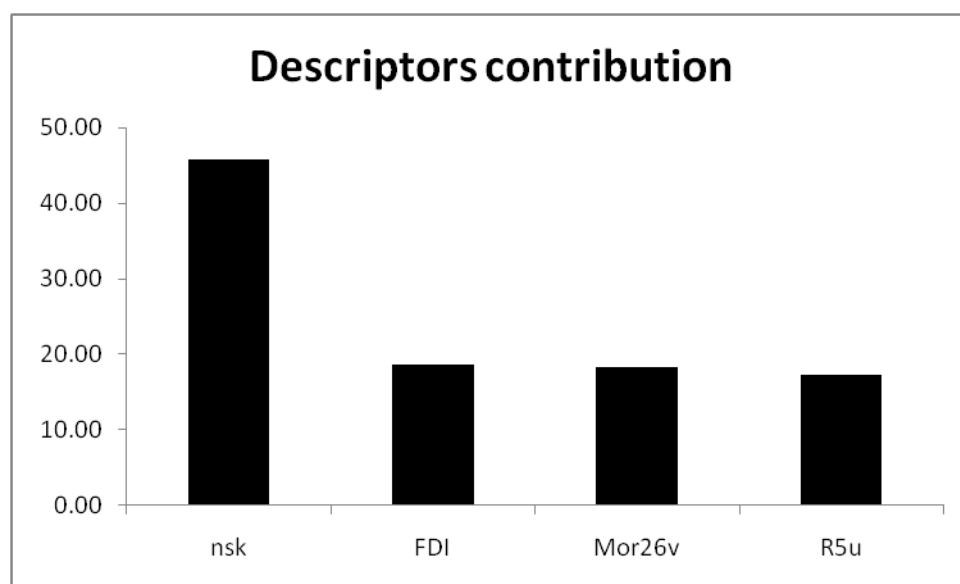


Figure 3: Relative contribution of the selected descriptors in the MLR model

The following statistical parameters according to Tropsha et al. reported in table 3, obtained for the external test set, check the generality accepted conditions which demonstrate the prediction power of the present model.

Table 3: Statistical parameters of Tropsha et al.

$R^2_{CV_{ext}}$	r^2	r_0^2	$r_0'^2$	T1
0.9668	0.968	0.9996	0.9993	-0.0326
T2	k	k'	Ab	
-0.0324	1.0035	0.9955	0.0313	

$$R^2_{CV_{ext}} = 0.9668 > 0.5 \quad ; r^2 = 0.968 > 0.6 \quad ; r_0^2 = 0.9996$$

$$r_0'^2 = 0.9993 \quad ; T1 = -0.0326 < 0.1, T2 = -0.0324 < 0.1$$

$$0.85 < k = 1.0035 < 1.15 \quad ; 0.85 < k' = 0.9955 < 1.15$$

$$|r^2 - r_0'^2| = 0.0313 < 0.3$$

The applicability domain is analyzed using the Williams plot, presented in figure 4 shows standardized residuals in prediction plotted against leverage (Hat diagonal) values of each compound used to evaluate the applicability domain (AD) of a QSPR model suggested by [24].

The plot makes possible to verify the presence of the outliers objects which are compounds with standardized residual greater than 3 standard deviation units and 9 compounds very influential in the determination of the model parameters which is the compounds with leverage greater than $h^* = 3(m+1)/n_{tr} = 0.1079$, where h^* is the warning leverage or the critical value (Figure 4).

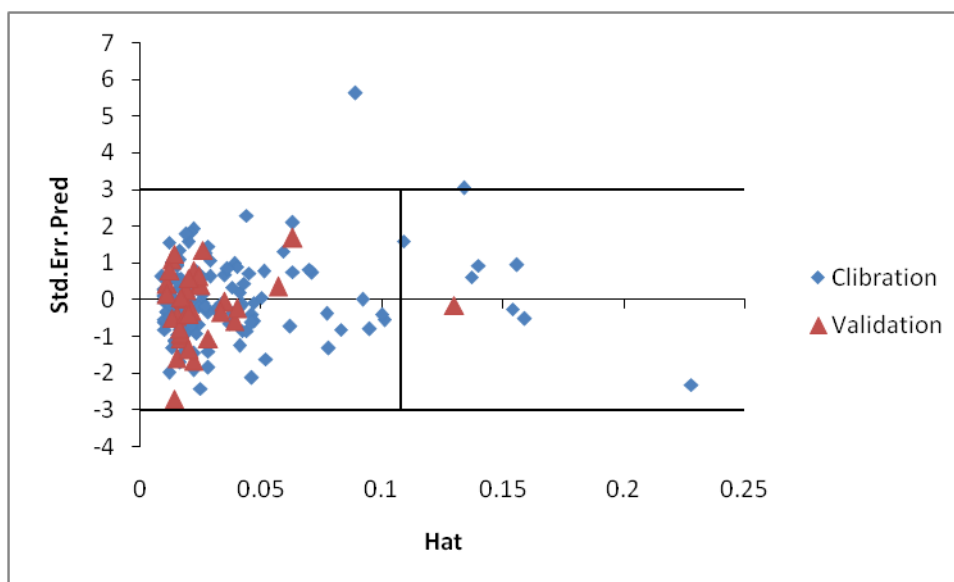


Figure 4: The Williams plot

As it is seen on figure 4 the only outlier object is Tridecylbenzene from the training set with a high FP value and considered as flammable substances. This compound is out of the AD of the QSPR model.

Nine compounds (Hexamethylbenzene, 1,2,4,5-Tetraisopropylbenzene, Ethylene, Acetylene, 2-Methylantracene, 9-Methylantracene, 7-Isopropyl-1-methylphenanthrene, 1,3,5-Tri-tert-butylbenzene)

from the training set and one object (1-Methylantracene) from the test set, are influential objects with Hat values greater than the critical Hat value, but they belong to the AD of the model.

CONCLUSION

A QSPR model for predicting flash points for 173 unsaturated hydrocarbons was established after applying successive steps beginning from the molecular structure generation to the model generation and the statistical analysis.

The obtained results ensure that the four molecular descriptors explain successfully the studied property which is the flash point. High correlation coefficient $R^2 = 0.9741$, high $Q_{\text{ext}}^2 = 0.9771$ and the low values of the prediction error (SDEP = 10.66 and SDEP_{ext} = 9.50) confirm the predictive ability of the obtained model.

The results showed that the predicted values of flash points agreed with the experimental values satisfactorily which can sometimes approach the accuracy of experimental flash point determination. Thus this QSPR model using MLR can be successfully used to estimate flash points for new organic compounds or for other unsaturated hydrocarbons for which experimental values are unknown. Furthermore, this work is of assistance to the further study on other flammability characteristics, such as auto ignition temperature and flammability limits, in order to predict the risks of environmental pollution.

REFERENCES

- [1] Evlanov SF, KhimZh, Prikl. J ApplChem-USSR (Engl. Transl.) 1991; 64: 747-752.
- [2] Tetteh J, Takahiro S, Metcalfe E, Howells S. J ChemInf Comp Sci 1999; 39: 491.
- [3] Katritzky AR, Petrukhin R, Jain R, Karelson M. J ChemInf Comp Sci 2001; 41: 1521-1530.
- [4] ASTM International, General test Method. 2004;14.02, (ASTM, West. Conshohocken, PA, 2004).
- [5] Liaw HJ, Lee YH, Tang CL, Hsu HH, Liu JH. J LOSS PREVENT PROC2002; 15: 429-438.
- [6] Lyman J, Reehl WF, Rosenblatt DH. Handbook of Chemical Property Estimation Methods. New York: McGraw Hill 1982: 751-752.
- [7] Vidal M, Rogers WJ, Holste JC, Mannan MS. A review of estimation methods for flash points and flammability limits. Process Saf Prog 2004; 23: 47-55.
- [8] Keshavars MH. Indian J Eng Mater S 2012; 19: 269-278.
- [9] Keshavarz MH, Ghanbarzadeh M. J Hazard mater 2011; 193: 335-341.
- [10] Hyperchem™, Release 6.03 for windows, Molecular Modeling system. 2000.
- [11] Todeschini R, Consonni V, Mauri A, and Pavan M. DRAGON Software for the Calculation of Molecular Descriptor. Version. 5.3 for Windows, Talete S. r. l., Milan. Italy. 2005.
- [12] Todeschini R, Ballabio D, Consonni V, Mauri A, and Pavan M. MOBYDIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 1.1 for windows, Milano. 2009.
- [13] Leardi R, Boggia R, Torrile M. J Chemometr 1992; 6: 267-281.
- [14] Xu J, Zang H, Lei Wang, Liang G, Wang L, Shen X, Xu W. SPECTROCHIM ACTA A 2010; 76: 239-247.
- [15] Todeschini R, Maiocchi A, Consonni V. ChemomIntell Lab Syst. 1999; 46: 13-29.
- [16] Eriksson L, Jaworska J, Worth A, Cronin M, McDowell RM, Gramatica P. Environ Health Perspect 2003; 111: 1361-1375.
- [17] Tropsha A, Gramatica P, Gombar VK. QSAR Comb Sci 2003; 22: 69-76.
- [18] Efron B. The jackknife, the Bootstrap and Other Resampling Planes, Society for Industrial and Applied Mathematics, Philadelphia, PA. 1994
- [19] Shi LM, Fang H, Tong W, Wu J, Perkias R, Blair RM, Branham WS, Dial SL, Moland CL, Sheehan DM. J ChemInf Comp Sci 2001; 41: 186-195.
- [20] Golbraikh A, Tropsha A. J Mol Graph Model 2002; 20: 269-276.
- [21] Weiberg S. Applied Linear Regression, 3rd edition. (John Wiley and sons, Inc., New Jersey); 2005.
- [22] SCAN-Software for Chemometric Analysis. Version 1.1-for Windows, Minitab USA; 1995.
- [23] Zheng F, Bayram E, Sumithran SP, Ayers JT, Zhen CG, Schmitt JD, Dwoskim LP, Crooks PA. Bioorg Med Chem 2006; 14: 3017-3037.
- [24] Gramatica P. QSAR Comb Sci 2007; 26: 694-701.