

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR- ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR- ANNABA



جامعة باجي مختار - عنابة
2018-2019

FACULTE DES SCIENCES
DEPARTEMENT DE CHIMIE

THESE

Présentée en vue de l'obtention du diplôme de Doctorat en Sciences
Option : Chimie Analytique et Environnement

Modélisation des indices de rétention d'une série de composantes alimentaires et polluantes de l'environnement : Méthodes MLR, LAD et PLS Approches simple et hybride.

Présenté par: **Fatiha Mebarki**

Devant le jury composé de :

Président:	Ali Larkem	Professeur	U.B.M Annaba
Directeur de thèse :	Salima ALI- MOKHNACHE	Professeur	U.B.M Annaba
Examineur:	Khoreif Nacereddine Abdelmalek	MCA	ENSET Skikda
Examineur:	Hicham Lahmer	MCA	U.M.S.B.Y Jijel

Remerciements

Mes premiers remerciements vont au responsable du laboratoire LASEA ; Monsieur le professeur MESSADI Djelloul, pour le temps et la patience qu'il m'a accordés tout au long des années de recherche en Magister et en Doctorat, Merci pour ses orientations scientifiques.

-Mes Vifs remerciements s'adressent à madame Salima ALI-MOKHNACHE qui m'a fait l'honneur de diriger ce travail.

Je tiens également à remercier les membres du jury de thèse, pour avoir pris le temps de lire ce manuscrit et de juger mon travail :

* Professeur Ali Larkem pour avoir accepté la présidence de ce Jury ; et

* Dr Abdelmalek Khoreif Nacereddine

* Dr Hichem Lahmer

Pour avoir accepté de faire partie de mon Jury de thèse et d'examiner ce travail.

Je dédie ce travail

À mes parents, mes frères et mes sœurs,

*À mes nièces (Darine, Selsabule, Ranime, Djana, Mohamed amine, Ahmed Abed Bassite et
Mohamed Rassime et tamime).*

À ma petite famille Surtout OMER et HOUDAIFA,

À mes amis Vous êtes toujours présents dans mon cœur et mon esprit.

Résumé

Les Relations quantitative structure-rétention (QSRR) ont été appliqués pour la modélisation des indices de rétention des composés pyrazines séparés par CPG sur deux colonnes différentes l'une non polaire (OV-101) et l'autre polaire (CRW-20M).

Les différentes méthodes d'estimation (MLR, LAD, PLS) sur les modèles linéaires obtenus montrent quels descripteurs jouent un rôle important dans la variation de l'indice de rétention de ces Pyrazines, ainsi que leur comparaison ; pour élucider les problèmes majeurs des données aberrantes et de la multicolinéarité par des équations des hyperplans et graphiques. L'élution des pyrazines sur la colonne polaire (CRW-20M) est la plus adaptée.

Mots clés: Régression LAD, Robustesse, Observations aberrantes, Points levier, tests statistique, PLS.

Abstract

The Quantitative Structure- retention Relationship (QSRR) was applied for modeling retention index for compounds of pyrazines eluted by CPG on two different column one nonpolar (OV-101) and the other polar (CRW-20M).

The Different methods of estimation (MLR, LAD, PLS) for obtained linear models show that descriptors play important role in the variation of retention index of these Pyrazines in their comparison to illustrated the major problems of the aberrant data and the multicollinearity by the hyperplans equations and graphic .Finally we can observation that The elution of pyrazines on the polar column (CRW-20M) is adapted.

Keywords: Least Absolute Deviation Regression, Robustness, Outliers, Leverage points, tests statistics, PLS, environmental.

ملخص:

تم تطبيق العلاقة الكمية بنية مؤشرات (QSRR) لنمذجة مؤشرات الاستبقاء لمركبات من البرازين

مفصولة بـ (CPG) على عمودي استقطاب مختلفين الاولي غير قطبي OV-101 و الاخر قطبية

.Carbovax-20M,

تبين طرق التقدير المختلفة (MLR, LAD, PLS) على النماذج الخطية المأخوذة الصفات التي تلعب

دور مهم في اختلاف مؤشرات الاحتفاظ لبيرازين بالإضافة الى مقارنتها لتوضيح مشكل النقاط الشاذة

و multicolinéarité بمعادلات وبيانات.

فصل البرازين على العمودي القطبي هو الانسب.

الكلمات الدالة: MLR- الانحراف المطلق LAD- المتانة - الملاحظات الشاذة و الرافعة - الاختبارات

الاحصائية - PLS - البيئية.

SOMMAIRE

REMERCIEMENTS

DEDICACES

RESUMES

LISTE DES TABLEAUX

LISTE DES FIGURES

SYMBOLES ET ABREVIATIONS

INTRODUCTION GENERALE

PARTIE I : Généralités

I-Définition :

I-1-Définition de l'environnement	4
I-2-Formation des hétérocycles volatils dans nos aliments	4
I-2-1-Les diazines	5
I-2-1-1-La pyrazine	6
I-2-1-2-propriétés de la pyrazine	7
I-2-1-3-Hybridation de La pyrazine	7
I-2-1-4-Mécanismes de formation de la pyrazine	8
I-2-1-5-La formation de pyrazines dans les aliments	8
I-2-1-6-Les arômes dégagés	9

II –Chromatographie

II.1.Définition de la chromatographie en phase gazeuse(CPG)	10
II-2- Phases stationnaires	10
II-2-1-La phase stationnaire liquide (Phase polaire) de type « Carbowax »	11
II-2-1-1-Le Carbowax 20M	11
II-2-2-La phase stationnaire liquide (Phase apolaire) OV-101	11
II-3-Indice de rétention (Ir)	12
II-3-1-Indice de rétention de Kováts	13
II-3-2-Indice de rétention de van den Dool et Kratz	14
II-3-3- Constantes de McReynolds « Détermination de la polarité des colonnes »	15
Références bibliographiques	16

PARTIE II : Aspects théoriques de la modélisation moléculaire

I-Modélisation

I-1-Modélisation de QSAR, QSPR et QSRR	20
I-1-1-Historique	20
I-1-2- QSPR (Quantitative Structure-Property Relationships)	20
I-1-3- Relation quantitative structure /rétention chromatographique QSRR	21
I-2-Préparation de la base des données	21
I-2-1-Les Logiciels utilisés	21
I-2-1-1-Chemoffice 2008	21
I-2-1-2-Logiciels « Hyperchem 7.5	22
I-3- Calculs des descripteurs moléculaires	23
I-3-1-Le Logiciel « DRAGON»	23
I-3-1-1- Les descripteurs	24
1 - Descripteurs moléculaires	26

1-1- Descripteurs 1D	26
1-2- Descripteurs 2D	26
* Indices constitutionnels	26
* Indices topologiques	26
1-3- descripteurs 3D	27
* Descripteurs géométriques.	27
* Descripteurs électroniques	28
* Descripteurs spectroscopiques	28
1-4- Descripteurs électrostatiques	28
1-5-Descripteurs quantiques	29
*Descripteurs liés à la distribution de charge	29
*Descripteurs liés à l'énergie,	30
1-6-Descripteurs thermodynamiques	30
I-4- Développement de modèles QSAR/QSPR	31
I-4-1-Sélection d'un sous- ensemble de variables par algorithme génétique (GA-VSS)	31
I-4-2- Méthodes utilisées pour le développement de modèles QSAR/QSPR	32
I-4-2-1-Régression multiple	33
1-Régression LAD multiple	34
1-2-Tests d'hypothèses sur les β_j	37
2- Régression linéaire multiple (MLR)	37
3-Analyse en composantes principales (ACP)	39
4- Régression en composantes principales (RCP)	40
5- La régression PLS	41
5-1-Les données statistiques et les représentations graphiques pour la fonction Analyse en composantes principales	43
5-1-1-Valeur propre	43
5-1-2-Proportion	43
5-1-3-Cumulé	43
5-1-4-Composantes principales (CP)	43
5-1-6-Diagramme en cône	44
5-1-7-Diagramme des contributions	44
5-1-9-Diagramme des valeurs aberrantes	44
5-1-10-Diagramme de sélection des modèles	44
5-1-11-Diagramme des réponses	45
5-1-12-Diagramme des coefficients normalisés	45
6-Tests sur le modèle linéaire	45
6-1-Test de la signification globale de la régression (F-Fisher)	45
6-2-Test de signification de chaque paramètre (chaque descripteur) t-Student	45
I-5- DEVELOPPEMENT ET EVALUATION DE MODELE.	46
I-5-1 Sélection d'un sous-ensemble de descripteurs	46
I-5-2 Principe.	46
I-5-3 Initialisation aléatoire du modèle.	47
I-6 Développement des modèles	47
I-6-1 Paramètres d'évaluation de la qualité de l'ajustement	47

I-6-2 Robustesse du modèle	47
I-6-3 Domaine d'application	48
I-6-4 Validation externe	49
I-7- Test de Durbin-Watson	50
I-7-1-Principe	50
I-8-Tests de normalité	51
I-8-1-Test graphique	51
I-8-1-1-Q-Q-plot des résidus	51
I-8-1-2-Approches empiriques et graphiques	52
1- Histogramme de la distribution	52
1-1- Histogramme des résidus	52
I-8-2-Test statistique	52
I-8-2-1-Test d'Anderson-Darling	52
I-9-Diagramme de travail	54
Références bibliographiques	56
Partie III : Résultats et discussions	
I. Résultats et discussion	62
II-1- Cas de la colonne non polaire (OV-101)	62
II-2- Cas de la colonne polaire (CRWAX-20M)	86
Références bibliographiques	121
CONCLUSION GENERALE	127
ANNEXE I Présentation des données	129
ANNEXE II Programmes de calculs	145
ANNEXE III Publications	



LISTE DES TABLEAUX

	Titre	page
	partie I	
Tableau I	valeurs des propriétés chimiques et physiques de la pyrazine	7
Tableau II	Pyrazines dans les produits alimentaires	9
Tableau III	Aromatisation avec les pyrazines	9
Tableau IV	Constantes de McReynolds (ΔI) de quelques phases stationnaires	15
	partie II	
Tableau V	valeurs critiques AD pour différents niveaux de risques	53
Tableau VI	formule Mathématique pour les valeurs de probabilité P à partir de la statistique transformée Am	54
	PARTIE III	
	Cas de La Colonne OV-101	
Tableau VII	Les descripteurs sélectionnés pour la modélisation d'IR.	62
Tableau VIII	Matrice de corrélation	63
Tableau IX	Évaluations de MLR pour le model	64
Tableau X	Évaluations de LAD pour le model	67
	Cas de La Colonne CRW-20M	
Tableau XI	Les descripteurs sélectionnés pour la modélisation d'IR.	86
Tableau XII	Matrice de corrélation	87
Tableau XIII	Évaluations de MLR pour le model	87
Tableau XIV	Evaluations de LAD pour le model	91
Tableau XV	Evaluations de MLR pour le model.	108
Tableau XVI	la méthode et les nombres des composants par PLS	109
Tableau XVII	Choix du Modèle et Validation pour IR(cw) par PLS	109
Tableau XVIII	Analyse des valeurs propres de la matrice de corrélation	110
Tableau XIX	Analyse des vecteurs propres de la matrice de corrélation	111
Tableau XX	Évaluations de MLR pour le model	117



LISTE DES FIGURES

	Titre	page
	partie I	
Figure 1	Quelques hétérocycles	5
Figure 2	Les diazines	6
Figure 3	Structure de la pyrazine	6
Figure 4	Différentes formes limites et charges partielles sur les tomes de pz.	7
Figure 5	Formation suggérée de l'alkyl pyrazine par réaction de Maillard	8
	Partie II	
Figure 6	système d'optimisation par la méthode MM+ et la méthode AM1 pour le pyrazine	22
Figure 7	Forme principale du logiciel de DRAGON.	24
Figure 8	Diagramme de notre travail.	55
	Partie III	
Figure 9	Structure générale de la pyrazine	61
	Cas de la Colonne OV-101	
Figure 10	Diagramme des scores normaux	65
Figure 11	Diagramme de valeur de Durbin Watson	66
Figure 12	Diagramme de VIF pour chaque descripteur	67
Figure 13	Diagramme de comparaison des coefficients de régression entre les deux Méthodes	69
Figure 14	Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes	70
Figure 15	valeurs estimées en fonctions de valeurs observées sur les deux méthodes	72
Figure 16	Diagramme de comparaison des coefficients de régression entre les deux Méthodes.	74
Figure 17	Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes	75
Figure 18	Diagramme de comparaison des coefficients de régression entre les deux Méthodes.	78
Figure 19	Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes	79
Figure 20	valeurs estimées en fonction des valeurs observées par les deux méthodes.	81
Figure 21	Diagramme des scores normaux (Calibration) pour la méthode MLR.	82
Figure 22	Diagramme des scores normaux (validation) pour la méthode MLR.	83
Figure 23	Diagramme des scores normaux (Calibration) pour la méthode LAD.	83
Figure 24	Diagramme des scores normaux (validation) pour la méthode LAD.	84
	Cas de la Colonne CRW-20M	
Figure 25	Diagramme des scores normaux	88
Figure 26	Diagramme de valeur de Durbin Watson	89
Figure 27	Diagramme de VIF pour chaque descripteur	90
Figure 28	Diagramme de principe pour les deux méthodes	92

Figure 29	Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes	93
Figure 30	valeurs estimées en fonctions des valeurs observées par les deux méthodes	95
Figure 31	Diagramme de comparaison des coefficients de régression entre les deux Méthodes.	97
Figure 32	Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes	98
Figure 33	Diagramme de comparaison des coefficients de régression entre les deux Méthodes.	100
Figure 34	Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes	101
Figure 35	valeurs estimées en fonction des valeurs observées sur les deux méthodes.	103
Figure 36	Diagramme des scores normaux (Calibration) par la méthode MLR.	104
Figure 37	Diagramme des scores normaux (validation) par la méthode MLR.	105
Figure 38	Diagramme des scores normaux (Calibration) par la méthode LAD.	105
Figure 39	Diagramme des scores normaux (validation) par la méthode LAD.	106
Figure 40	Diagramme de sélection de modèle vers le nombre des composants	110
Figure 41	Diagramme en cône des valeurs propres et les proportions vers le nombre de composantes	111
Figure 42	Diagramme des contributions de corrélation pour les descripteurs	113
Figure 43	Diagramme de coefficient vers le nombre de composants	114
Figure 44	Diagramme des réponses pour les valeurs calculées vers les valeurs observées	115
Figure 45	Diagramme des valeurs Résiduel vers les valeurs de levier.	116
Figure 46	Digramme de VIF pour chaque descripteur	118
Figure 47	valeurs estimées en fonctions de valeurs observées sur les deux méthodes	118
Figure 48	Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes	119



SYMBOLES ET ABBREVIATIONS

AD	: Anderson Darling
CW-20M	: Carbowax-20M
DW	: test de Durbin-Watson
EQMC	: Ecart quadratique moyen calculé sur l'ensemble de calibration
EQMP	: Ecart quadratique moyen de prédiction
EQMPext	: Ecart quadratique moyen calculé sur l'ensemble de validation externe.
e_i	: Erreur résiduelle, $e_i = Y_i - \hat{Y}_i$.
F	: Statistique de Fisher.
GA	: Algorithme génétique (Genetic Algorithm).
H0	: Hypothèse nulle
H1	: Hypothèse alternative
h_i	: Eléments diagonaux de la matrice chapeau.
IC	: Interval de confiance
IR	: Indice de rétention
IX	: Indice de rétention de Kovats d'un composé x.
i	: Indice de l'observation.
LAD	: Méthode des moindres écarts en valeurs absolues (Least Absolute Deviations)
LOO	: Leave-One-Out: Validation croisée par omission d'une observation.
LS	: Méthode des moindres carrés (Least Squares).
MED	: Médiane.
MLR	: Régression linéaire multiple
N	: Taille de la population (échantillon).
Next	: Dimension de l'ensemble de validation.
n-2	: Nombre de degrés de liberté.
OV-101	: OhioValley Speciality company 101
p	: Nombre de paramètres.
p-1	: Nombre de descripteurs.
PM3	: Parameterized model number 3
PRESS	: Somme des carrés des erreurs de prédiction.
Q ²	: Coefficient de prédiction
QSAR	: Quantitative Structure/ Activity Relationships (Relations Quantitatives Structure/ Activité).
QSPR	: Quantitative Structure/ Property Relationships (Relations Quantitatives Structure/ propriété)
QSRR	: Quantitative Structure/ retention Relationships (Relations Quantitatives Structure/ rétention).
R ²	: Coefficient de détermination.
S	: Erreur standard.
SCE	: Somme des carrés des écarts.
SCE	: Somme des carrés des écarts
SCT	: Somme des carrés totale.
SCT :	: Somme des carrés totale.
t_{calc}	: Valeur de Coefficient t de Student calculée.
t_{obs}	: Valeur de Coefficient t de Student observée.
W_i	: Poids associé à la ième observation.
X	: Variable explicative.
Y	: Variable à expliquer.

Symboles et abréviations

Y_i	: Valeur observée
\hat{Y}_i	: Valeur estimée
α	: Niveau de confiance.
β_0 ou b	: Ordonnée à l'origine.
β_1	: Pente.
B_j	: jème coefficient de régression.
$\sum e_i $: Somme d'erreurs en valeurs absolues.
σ^2	: La Variance
$\sum e_i^2$: Somme carrée d'erreurs



INTRODUCTION GENERALE

Depuis les années 1970 le terme environnement est utilisé pour désigner le contexte écologique global, c'est-à-dire l'ensemble des conditions physiques, chimiques, biologiques, climatiques, géographiques et culturelles au sein desquelles se développent les organismes vivants, et les êtres humains en particulier. L'environnement inclut donc l'air, la terre, l'eau, les ressources naturelles, la flore, la faune, les hommes et leurs interactions sociales [8].

Quelques composés alimentaires sont des hétérocycles volatils qui sont retrouvés de façon naturelle dans notre environnement et l'attrait que les hommes éprouvent pour les arômes ne s'est jamais démenti au cours des siècles ayant un intérêt dans de multiples domaines, notamment dans l'alimentation. Leur présence dans les aliments résulte principalement d'un processus nécessitant une étape de cuisson (partielle ou complète). La civilisation égyptienne les utilisait déjà pour la cuisine [29].

Les hétérocycles volatils constituent également une famille importante de molécules odorantes, particulièrement intéressantes dans le domaine de la chimie des arômes et l'odeur peut être considérée comme une pollution locale et une nuisance limitée à la population riveraine des sources potentielles. Ils représentent plus d'un quart des 5 000 composés volatils isolés et caractérisés à ce jour dans nos aliments [8].

Les pyrazines sont des hétérocycles très présents dans nos aliments. Plus de 80 dérivés de pyrazines ont été identifiés dans un grand nombre d'aliments cuits, comme le pain, la viande, le café torréfié, le cacao ou les noisettes ; ce sont des composés aromatisants très puissants [8].

Les pyrazines (1,4-diazines) sont des hétérocycles azotés très largement distribués dans la nature, aussi bien dans le règne animal que végétal. Le domaine de l'alimentation reste celui dans lequel les pyrazines sont le plus étudiées. Elles sont considérées comme des composés

hétérocycliques les plus largement représentés dans l'arôme des aliments [1]. On peut les classer en trois groupes selon leurs origines : formées par traitement à la chaleur, par des microorganismes ou présentes à l'état naturel dans les végétaux [9] .

Le couplage chromatographie gazeuse /spectrométrie de masse, s'il facilite souvent les questions d'identification peut être inefficace dans l'analyse des isomères ou des composés mineurs d'un mélange complexe .Les relations structure / rétention peuvent, dans ces conditions, aider à l'identification. Il s'agit de relier les réponses obtenues pour un ensemble d'évaluation à des propriétés physico-chimiques expérimentales ou théoriques, et/ou des descripteurs moléculaires de différents types fournis par divers logiciels spécialisés[.]

On peut, a priori, supposer des corrélations linéaires, ou des corrélations plus compliquées. Dans le premier cas on fera appel à des techniques comme la régression linéaire multiple (MLR) ou la projection sur les structures latentes par la méthode des moindres écarts en valeurs absolues LAD (Least Absolute Deviation) ou les moindres carrés partiels (PLS),

A notre connaissance un seul article , qui remonte à 20 ans, s'est consacré aux relations quantitatives structure/rétention (QSRR: pour Quantitative Structure Rétention Relationship) de pyrazines séparées par chromatographie gazeuse à température programmée, sur deux colonnes de polarités très différentes. Les résidus obtenus pour les indices de rétention, particulièrement avec la colonne Carbowax 20M dépassent souvent 36 unités d'indice. Dans ces conditions, et contrairement à ce qu'affirment les auteurs de l'article [2] les modèles ne peuvent être considérés comme quantitatifs.

Il y a diverses méthodes robustes pour l'évaluation les paramètres de régression, La robustesse de la méthode LAD par rapport aux observations aberrantes, et sa susceptibilité aux points levier ont été largement étudiées en littérature [3 -4].

Le but poursuivi dans cette étude est de comparer différentes méthodes d'estimation des paramètres dans les modèles linéaires lorsque l'on est confronté à deux problèmes majeurs en

analyse de régression, à savoir le problème des données aberrantes et celui de la multicollinéarité

Notre mémoire comporte trois parties en plus d'une l'introduction et d'une conclusion générale.

Le premier partie englobe des généralités sur les composés alimentaires étudiés (pyrazines) séparés par CPG sur deux colonnes différents OV-101 et CRW-20M.

La deuxième partie englobe L'introduction des molécules et l'optimisation de leur géométrie pour le calcul des descripteurs moléculaires à l'aide de différents logiciels de modélisation, après un développement théorique des connaissances de base de différents technique (MLR, LAD, PLS) englobent la deuxième partie.

La troisième partie : partie expérimentale englobe : la collecte des données avec l'étude de différentes techniques pour la détermination (le choix) de chaque paramétré caractérisant le meilleur modèle et une étude comparative entre les différents méthodes d'estimation MLR LAD et PLS avec un traitement statistique.

Des Annexes en finalité.

PARTIE I

GENERALITES

- *Définition de :*
- *l'environnement*
- *pyrazine et ses dérivés*
- *Chromatographie*



I-/Définitions :**I-1/Définition de l'environnement [5,6]:**

L'environnement est l'ensemble des éléments constituant le voisinage d'un être vivant ou d'un groupe d'origine humaine, animale ou végétale, et susceptibles d'interagir avec lui directement ou indirectement. C'est ce qui entoure, ce qui est aux environs.

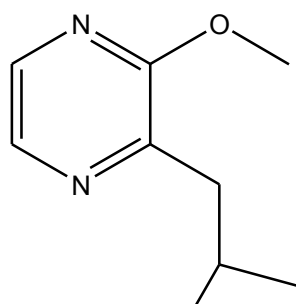
Depuis les années 1970 le terme environnement est utilisé pour désigner le contexte écologique global, c'est-à-dire l'ensemble des conditions physiques, chimiques, biologiques climatiques, géographiques et culturelles au sein desquelles se développent les organismes vivants, et les êtres humains en particulier. L'environnement inclut donc l'air, la terre, l'eau, les ressources naturelles, la flore, la faune, les hommes et leurs interactions sociales.

I-2/Formation des hétérocycles volatils dans nos aliments [7,8,9] :

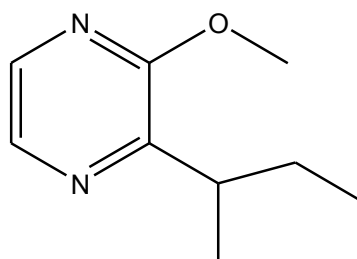
Les précurseurs des composés hétérocycliques sont les constituants fondamentaux de nos aliments : acides aminés, peptides, glucides, lipides et vitamines. Deux voies principales sont à l'origine de leur formation :

- les réactions enzymatiques ou la fermentation,
- les réactions de brunissement non enzymatiques, plus connues sous le nom de réactions de Maillard qui surviennent lors des différents traitements subis par nos aliments : cuisson, torréfaction, conservation, grillage. A ces réactions, on peut également associer les réactions de dégradation thermique des sucres, des acides aminés et des vitamines.

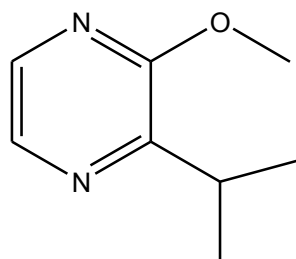
Les réactions enzymatiques ont lieu principalement dans les fruits , les légumes, les produits laitiers et les boissons fermentées.



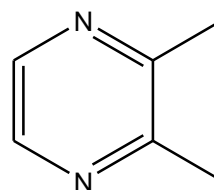
Iso-butyl-2 méthoxy-3 pyrazine(poivon)



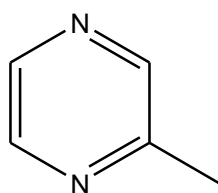
Sec-butyl-2 méthoxy-3 pyrazine (Carotte crue)



Iso- propyl -2 methoxy -3 pyrazine(petit pois)



Diméthyl pyrazine(Odeur de cuivre frais)



Méthyl pyrazine(Odeur de brulé)

Figure 1:- Quelques hétérocycles a compose de Pyrazine**I-2-1/ Les diazines**

Les diazines sont des hétérocycles aromatiques à six chaînons comportant deux atomes d'azote. Ces cycles, en raison de la présence d'atomes d'azote plus électronégatifs que le carbone, sont déficitaires, ce qui rend leur fonctionnalisation plus difficile, ainsi les substitutions électrophiles ne sont pas possibles sur de tels cycles. Selon la position respective des deux atomes d'azote, on distingue: la pyridazine, la pyrimidine, et la pyrazine (figure 2) :

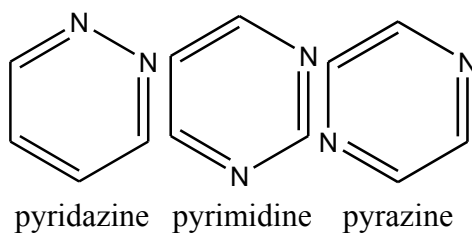


Figure 2 : Structure type de pyrazine

De nos jours, de nombreuses diazines polysubstituées sont connues pour leurs propriétés dans des domaines variés comme l'alimentaire (arôme et parfum), l'agrochimie et la pharmacologie.

I-2-1-1/La pyrazine :

La **pyrazine** (ou 1,4-diazine), de formule brute $C_4H_4N_2$, est un composé hétérocyclique simple et fondamental qui se rapproche de la structure du benzène où deux des groupements CH sont remplacés par des atomes d'azote. Elle est l'isomère de position de la pyrimidine (1,3-diazine) et de la pyridazine (1,2-diazine).

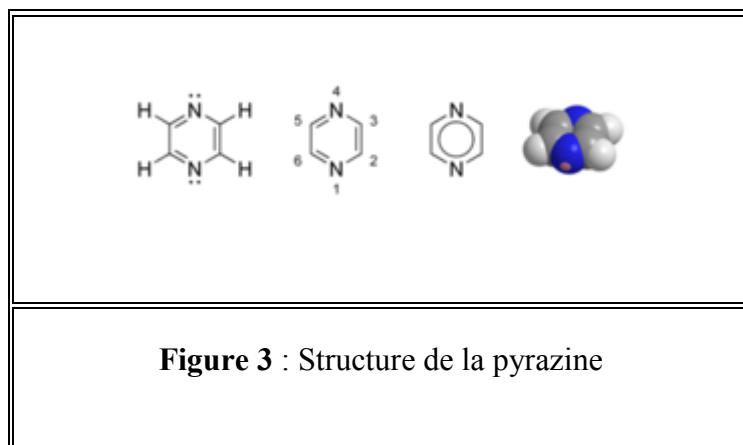


Figure 3 : Structure de la pyrazine

I-2-1-2/propriétés de la pyrazine :

Table I : valeurs des propriétés chimiques et physiques de la pyrazine.

Propriétés chimiques	
Formule brute	C ₄ H ₄ N ₂ [Isomères]
Masse molaire	80,088194 g·mol ⁻¹ C 59,99%, H 5,03%, N 34,98%,
pKa	0,51 à 20°C[1]
Propriétés physiques	
T° fusion	55°C[1]
T° vaporisation	115°C[1]
Solubilité	Soluble dans l'eau.
Masse volumique	1,031
Point d'éclair	55°C

I-2-1-3/Hybridation de La pyrazine :

La pyrazine (1, 4-diazine) est de symétrie plus élevée (D_{2h}) que la pyridine. Ses quatre atomes de carbone, tout comme ses quatre atomes d'hydrogène, sont équivalents. Il y a trois formes limites pour cette molécule). Les charges partielles sur les atomes ont été calculées (méthode ab initio) et sont représentées sur la figure suivante.

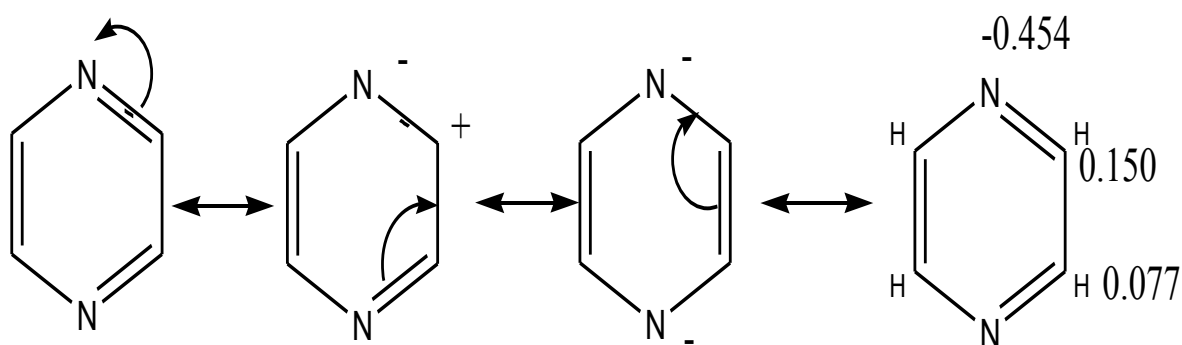


Figure 4: Les différentes formes limites et les charges partielles sur les tomes de pz.

I-2-1-4/Mécanismes de formation de la pyrazine [7,9] :

il existe beaucoup de mécanismes pour la formation de la pyrazine .Ils ont été étudiés au cours des 40 dernières années Suivant .:

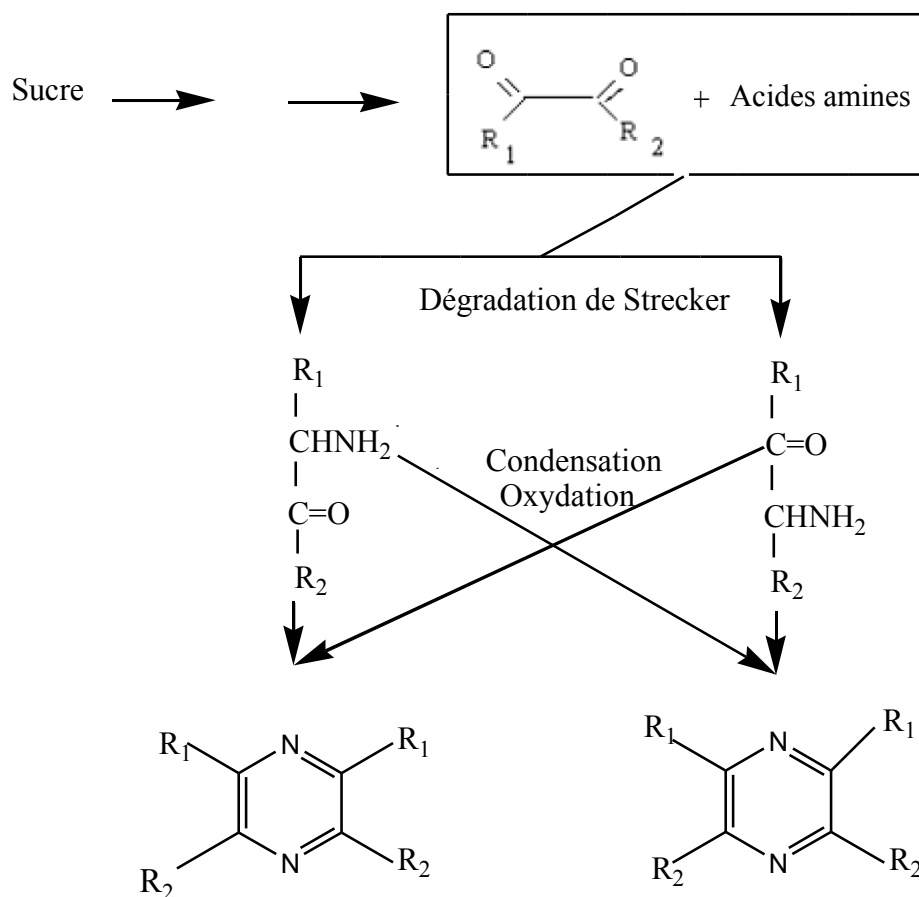


Figure 5 : Formation suggérée de l'alkyl pyrazine par réaction de Maillard

I-2-1-5/Formation de la pyrazines dans les aliments [9] :

Les pyrazines peuvent être synthétisées principalement de deux façons. D'une part, en chauffant les aliments. Il est donc possible de générer la formation de pyrazines grâce à la réaction de Maillard³. Cette réaction permet, grâce à divers réactions complexes, de former des pyrazines à partir d'acides aminés et de sucres : les acides aminés servent alors de source d'azotes et les sucres (glucides) fournissent les carbones qui viendront former la molécule de pyrazine (Figure 1). Cependant, afin d'éviter la dégradation de ces composés, il est essentiels

de ne pas dépasser une température supérieure à 150°C ,il est préférable de la réaliser à 115°C.

I-2-1-6/Les arômes dégagés [5,10] :

Les arômes produits par cette famille sont aussi variées que les composés sont nombreux. Par exemple les 2, 5-diméthylpyrazine et 2, 6-diméthylpyrazine sont responsables des arômes du chocolat, des noix grillées et de la pomme de terre frite. Le goût des cacahuète grillée est également caractéristique de divers dérivés de pyrazines tels que la 2-méthylpyrazine, la 2-éthylpyrazine, la 2-éthyl-6-méthylpyrazine et la 2,3-diéthyl-5-méthylpyrazine. En général, l'arôme d'un aliment est définit par un mélange de composés odorants. C'est le cas par exemple dans le sirop d'érable, où parmi les 27 pyrazines détectées, 15 possèdent des odeurs caractéristiques.

Tableau 2 : Pyrazines dans les produits alimentaires

Substituants	Qualité aromatique
2-méthyl-3-éthyle	Brûlé
Acétyle-	maïs rôti
2-éthyl-3,5-diméthyle	pommes de terre
2-éthyl-3,6-diméthyle	pommes de terre
2,3-diéthyl-5-méthyle	pommes de terre
2-sec-butyl-3-méthoxy-	Terreux
2-isobutyl-3-méthoxy-	fort, paprika

Tableau 3 : Aromatisation avec les pyrazines

Substance	Produit alimentaire	Arôme
2-éthyl-3-vinylpyrazine (6)	café instantané	Terreux
2-éthyl-3,5-diméthyl-pyrazine	sirop de glucose	amandes grillées
2-éthyl-3,6-diméthyl-pyrazine	sirop de glucose	Noisette
Formylpyrazine	café instantané	note de "grillé"
2-éthoxy-3-méthyl-pyrazine	Glaces	noix grillées
2-éthyl-3-méthoxy-pyrazine	produits à base de pommes de terre	pommes de terre

II –Chromatographie [11,12] :

La chromatographie est la méthode de séparation la plus générale, la plus puissante, et la plus simple qui soit actuellement disponible. Elle représente un des procédés physico-chimiques de séparation, au même titre que la distillation, la cristallisation ou l'extraction fractionnée, des constituants d'un mélange homogène liquide ou gazeux. Les applications de cette méthode sont donc potentiellement très nombreuses, d'autant plus que beaucoup de mélanges hétérogènes ou sous forme solide peuvent être mis en solution par emploi d'un solvant. Ce procédé hydrodynamique a donné naissance à une méthode analytique instrumentale, présentant un très grand domaine d'applicabilité et par suite se trouve très répandue.

II.1.Définition de la chromatographie en phase gazeuse(CPG)

Cette technique découverte en 1952 par James et Martin, a fait des progrès vertigineux, particulièrement grâce à la découverte des détecteurs ultrasensibles. C'est également le seul type de chromatographie qui utilise un gaz comme phase mobile [13].

Ce type de chromatographie peut être subdivisé selon le phénomène mis en œuvre:

- Chromatographie gaz/liquide(CGL) : La phase mobile est un gaz, et la phase stationnaire est un liquide fixé par imbibition d'un support inerte.
- Chromatographie gaz/solide(CGS) : La phase stationnaire est un solide poreux (carbone graphite ou gel de silice ou alumine) et la phase mobile est un gaz [14].

Les phases stationnaires garnissent des tubes métalliques ou en verre de faible diamètres (colonne à garnissage) ou sont déposées sur les parois internes d'un tube de très faible diamètre (colonnes capillaires).

II-2- Phases stationnaires :

Les phases actuelles correspondent à deux principaux types de composés: les polysiloxanes et les polyéthylèneglycols, chaque catégorie pouvant faire l'objet de modifications structurales mineures.

II-2-1-La phase stationnaire liquide (Phase polaire) de type « Carbowax » :

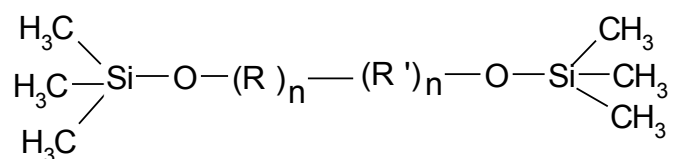
Les colonnes capillaires de type « Carbowax » sont pourvues d'une phase stationnaire à base de polyéthylène glycol très polaire. Cette phase permet une bien meilleure séparation des composés polaires que les phases apolaires ou légèrement polaires. Il existe diverses colonnes de type « Carbowax » caractérisées par des températures d'utilisation max./min., une inertie et une sélectivité différentes du fait des modifications apportées à la phase pour permettre l'analyse de familles spécifiques de composés. Ces colonnes polaires sont couramment utilisées pour les séparations de glycols, solvants, impuretés organiques volatiles dans les médicaments, acides gras, parfums, arômes, ... Les colonnes Stabilwax[®]-DA et Stabilwax[®]-DB sont respectivement destinées à l'analyse des composés acides ou basiques.

Le degré de polarité lié au nombre d'hydroxyles est indiqué par un chiffre qui représente la masse moléculaire. Les masses de Carbowax qui peuvent aller de 300 à 20.000 sont dénommées par ces valeurs extrêmes Carbowax 300 et Carbowax 20M.

II-2-1-1-Le Carbowax 20M : la colonne de masse moléculaire la plus élevée((20.000), est donc le moins polaire de la série. Ces phases stationnaires possédant de nombreux oxygènes sont classées parmi les phases stationnaires les plus polaires, et elles sont utilisées pour séparer les molécules de fortes polarités comme celles possédant des fonctions alcool, aldéhyde, ou cétone, Alkylamines, diamines, composés hétérocycliques azotés

II-2-2-La phase stationnaire liquide (Phase apolaire) OV-101 :

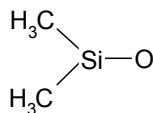
La phase stationnaire liquide OV-101 appartient à la famille des silicones qui répondent à la formule générale :



R et R' sont des groupements silylés pouvant posséder des groupements polaires ou polarisables, et par la même déterminer les propriétés de ces phases.

Dans la phase liquide OV-101:

R = R' =



La température maximale d'utilisation de la colonne OV-101 est comprise entre 300 et 350 °C. Les colonnes SE 30, OV1, OV101, diméthylsiloxanes sont très peu polaires, elles se différencient par leur nombre de groupements diméthylsilyle. La colonne OV-101 en possédant moins, a une viscosité plus faible et une température maximum d'utilisation plus basse. Elles sont utilisées pour séparer les dérivés méthyles ou silylés.

II-3-Indice de rétention (Ir) :

La chromatographie en phase gazeuse peut permettre l'identification d'un produit dans un mélange complexe. A cette fin on utilise le temps de rétention (t'_R) ou plus communément l'indice de rétention (Ir) qui dépend de la structure du composé [14].

L'introduction de ces paramètres a au moins trois objectifs:

- Caractériser tout composé par une grandeur plus générale que son temps de rétention dans des conditions définies. Il en résulte le système des indices de rétention qui est un moyen efficace et peu coûteux pour éviter certaines erreurs d'identification.
- Suivre l'évolution dans le temps des colonnes et par suite leurs performances.
- Classer entre elles les phases stationnaires connues pour faciliter le choix d'une colonne bien adaptée pour tout problème nouveau de séparation, sachant que la polarité ou la nature chimique d'une phase stationnaire ne permettent pas, seules, de prévoir sa réelle aptitude séparatrice [14].

Le calcul peut se faire pour une expérimentation à température constante par interpolation logarithmique: indices de Kováts [15], ou en programmation de température par interpolation linéaire indices de rétention, ou indices de van den Dool et Kratz [16]. Bien que dans la grande majorité des cas, chaque molécule possède des indices de rétention sur colonne apolaire et polaire qui lui sont propres, deux molécules peuvent fortuitement être co-éluées et présenter des indices de rétention identiques [17].

II-3-1-Indice de rétention de Kováts :

L'indice de rétention de Kováts (1958) est une mesure de la rétention d'un composé par rapport à la rétention des alcanes normaux (hydrocarbures à chaîne droite) à une température constante. Ils sont utilisés comme référence car ils sont non polaires, chimiquement inertes et solubles dans la plupart de phases stationnaires. L'indice de rétention de Kováts d'un composé (x) sur la colonne considérée se calcule comme suit :

$$I_x = 100 \times n + 100 \frac{\log t'_{R(x)} - \log t'_{R(n)}}{\log t'_{R(n+1)} - \log t'_{R(n)}} \quad (1)$$

I_x : Indice de rétention d'un composé x.

a $t'_{R(x)}$: temps de rétention du composé x.

$t'_{R(n)}$: temps de rétention réduit de l'alcane élué avant x.

$t'_{R(n+1)}$: temps de rétention réduit de l'alcane élué après x.

Les deux temps de rétention se rapportent à deux alcanes successifs (n et n+1 atomes de carbone), ou à deux composés de même type.

-Le temps de rétention réduit du composé t'_R est la différence entre son temps de rétention et le temps de rétention nulle t_0 , comme le montre l'équation :

$$t'_R = t_R - t_0 \quad (2)$$

Le temps de rétention nulle ou le temps de rétention d'un composé non retenu t_0 : correspond au temps que met un constituant pour traverser l'ensemble du système chromatographique sans interaction avec la phase stationnaire.

Un des désavantages de l'indice de Kováts est qu'il n'est pas utilisable en chromatographie gazeuse à température programmée (CGTP). Pour combler cette lacune van den Dool et Kratz proposèrent de calculer une grandeur (I_p), semblable à l'indice de Kováts, en remplaçant dans l'expression de ce dernier le logarithme du temps de rétention réduit directement par le temps (ou la température) de rétention.

II-3-2-Indice de rétention de van den Dool et Kratz :

$$\frac{I_p}{100} = n + \frac{T_{R(x)} - T_{R(n)}}{T_{R(n+1)} - T_{R(n)}} \quad (3)$$

Avec $TR(n) < TR(x) < TR(n+1)$;

$TR(x)$ est la température de rétention du soluté x , $TR(n)$ et $TR(n+1)$ celles des n -alcanes de référence à n et $(n+1)$ atomes de carbone l'encadrant sur le chromatogramme [17].

L'indice de rétention est une donnée facilement accessible avec précision, qui est indépendante des caractéristiques de l'appareil (paramètres de colonne) et n'est fonction que du soluté, de la température et de la phase stationnaire. L'utilisation simultanée des indices de rétention sur 2 colonnes de polarité différente conduit à la notion d'incrément d'indice.

$$\Delta I_r = I_r^p - I_r^a \quad (4)$$

D'après Kováts le ΔI_r d'un composé résulte des différences d'interactions entre les molécules des 2 phases stationnaires et les groupements fonctionnels que possède le soluté i .

I_r^p : Indice de rétention d'un composé i sur la phase polaire ;

I_r^a : Indice de rétention d'un composé i sur la phase apolaire.

II-3-3- Constantes de McReynolds « Détermination de la polarité des colonnes »

Pour caractériser le comportement d'une phase stationnaire on compare les indices de Kovats de cinq composés témoins appartenant à des types structuraux différents sur la phase étudiée d'une part et sur le squalane d'autre part, « phase qui a été choisie comme référence pour ce calcul ». Les cinq indices pour la colonne au squalane, la seule phase apolaire qui soit reproductible car formée d'un produit pur, ont été établis une fois pour toutes (tableau 4). Les cinq constantes de McReynolds pour une phase donnée, s'obtiennent en calculant les différences observées pour chacune des substances testées entre l'indice de Kovats sur une phase squalane (Isqualane) et l'indice sur la phase étudiée (Iphase):

Constante de McReynolds = $\Delta I = (I_{\text{phase}} - I_{\text{squalane}})$

La somme de ces 5 valeurs calculées d'après la formule, a été retenue pour définir la polarité globale de la phase testée.

On admet que chacun des composés test apporte une information particulière sur la phase stationnaire. Le benzène pour l'effet inducteur, la pyridine pour l'effet accepteur de H⁺, le butanol pour les liaisons hydrogène, le nitropropane pour les interactions dipolaires. Ces constantes, qui dépendent des structures moléculaires, permettent d'apprécier les forces soluté/phase stationnaire en fonction de quelques grandes classes de composés. Un indice dont la valeur est élevée indique que la phase étudiée retient fortement les composés porteurs de la fonction organique correspondante, ce qui normalement conduit à une sélectivité accrue [14].

Tableau IV : Constantes de McReynolds (ΔI) de quelques phases stationnaires.

Phase stationnaire	benzène	butan-1-ol	pentan-2-one	nitropropane	pyridine
Squalane	0	0	0	0	0
SPB—Octyl	3	14	11	12	11
SE-30(OV-1)	16	55	44	65	42
Carbowax 20M	322	536	368	572	510
OV-201	146	238	358	468	310
indice de Kovats des 5 composés témoins (X' Y' Z' U' S') sur squalane/squalane	653	590	627	652	699



REFERENCES BIBLIOGRAPHIQUES

- [1] Vernin G. et Vernin G. 1982. Heterocyclic aroma compounds in foods: occurrence and organoleptic properties. In Chemistry of heterocyclic compounds in flavours and aromas, vernin G. Ed., Ellis Horwood Limited, England, 72-150.
- [2] Stanton, D.T. and P.C. Jurs, 1989. Computer-assisted predict of gas chromatographic retention indexes of pyrazines. Anal. Chem., 61: 1328-1332.
- [3] Dodge, Y. 1987. Statistical Data Analysis Based on the Li-Norm and Related Methods. 1st Edn., North-Holland, Amsterdam, ISBN-10: 0444702733, pp: 464.
- [4] Dodge, Y. 1997. L1-Statistical Procedures and Related Topics. 1st Edn., Institute of Mathematical Statistics, Hayward, ISBN-10: 0940600439, pp: 498.
- [6] Fabier locher,grégory quenet.2009.Revue d'histoire modern et contemporaine.
- [7] Benoît Otte Jacques Nicolas.2005.Thèse de doctorat Les odeurs dans l'environnement: Sources d'odeur en Région Wallonne. Université de liège, campus d'Arlon, département des sciences et gestion de l'environnement.
- [8] Fatiha Mebarki, Khadija Amirat, Salima Ali Mokhnach et Djellol Messadi .2017. Least Absolute Deviation Regression and Least Squares for Modeling Retention Indices of Set Compounds Food and Pollutants of the Environment. American Journal of Applied Sciences :14 (5): 592.606.
- [9] Diego Tho . 2012/2013. Mémoire de Master 2 Synthèse Bibliographique en Biologie et Biotechnologie Voies de métabolisme des pyrazines dans les produits alimentaires. Biologie Gestion Universite de rennes UFR Sciences de la vie et de l'environnement. France.

- [10] Céline Niquet. 2007. These doctorat Identification de La structure des Mécanismes de formation de quelques produits de Maillard non Issus de l'ammoniac role precurseur de la glutamine.
- [11] Dr Ing. Antoine Audrin, Dr Jürg Löliger, Prof. Dr Ing et Werner Bauer. Livre Aromes et Colorants.
- [12] A. Dari.2015/2016. Techniques chromathgraphie.
- [13] Salghi R .2012. Cours d'analyses physico-chimique des denrées alimentaires II, GPEE, ENSA Agadir.URL
- [14] Rouessac F. Rouessac A .2004.Analyse chimique : Méthodes et techniques instrumentales modernes .6e édition, Paris,Dunod,
- [15] Kováts E. 1958. Chromatographische charakterisierung organischer verbidugen. Teil:Retention indices aliphatischer halogenide, alkohole, aldehyde und ketones. Helvetica Chimica Acta, vol 41(7): 1915–1932.
- [16] Van den Dool H. Kratz P. 1963. A generalization of the retention index system including linear temperature programmed gas–liquid partition chromatography.Journal of Chromatography, vol 11: 463-471
- [17] Sutour S. 2010.Thèse de doctorat. Étude de la composition chimique d'huiles essentielles et d'extraits de menthes de corse et de kumquats. Université de corse pascal Paoli. p 22
- [18] Lourici .L et Messadi .D. 2009 .méthodes non linéaires pour le calcul des indices de rétention en chromatographie gazeuse à programmation linéaire de température . Lebanese Science Journal, Vol. 10(10) :77-86.

- [19] Mahuzier G. Hamon M . Ferrier D et Prognon P .1999.Chimie analytique :Méthode de séparation .Tome 2,3^eédition ,paris,Masson.
- [20] Li, W., C.L. Heth and S.C. Rasmussen, 2014. Thieno[3,4-b] pyrazine-based oligothiophenes: Simple models of donor-acceptor polymeric
- [21] Buchbauer, G.2000. Threshold-based structure-activity relationships of pyrazines with bell-pepper Flavor. *J. Agric. Food Chem.*,48:4273-4278. PMID: 10995349
- [22] Mebarki, F., K. Amirat, S.A. Mokhnache and D. Messadi,2016. Treatment by alternative methods of regression gas chromatographic retention indices of 35 pyrazines. *Int. J. Instrument. Control Syst.*, 6: 1-14.
- [23] Small, G.W. and P.C. Jurs. 1983. Interactive computer system for the simulation of carbon-13 nuclear magnetic resonance spectra. *Anal. Chem.*, 55: 1121-1127. DOI: 10.1021/ac00258a033.
- [24] Gonin, R. and A.H. Money. 1989. *Linear LP-norm Estimation*. 1st Edn.,
- [25] Judge, G.G, W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.C. Lee.1985. *The Theory and Practice of Econometrics*. 2nd Edn, Wiley, New York, ISBN- 10: 047189530X, pp: 1019.
- [26] Mihara, S. and N. Enomoto. 1985. Calculation of retention indices of pyrazines on the basis of molecular structure. *J. Chromatogr.* 324: 428-430.
- [27] Eriksson. L, J. Jaworska, A. Worth, M. Cronin and R.M. Mc Dowell et al. 2003. Methods for reliability, uncertainty assessment and applicability evaluations of regression based and classification QSARs. *Environ. Health Perspect.*, 111: 1361-1375.

[28] Tropsha, A, P. Gramatica and V.K. Grombar.2003. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Combi. Sci.,22: 69-76.

[29] Fatiha Mebarki, Khadidja Amirat, Salima Ali Mokhnache, Djelloul Messadi.2016. Modeling Retention Indices of a Series Components Food and Pollutants of the Environment: Methods; OLS, LAD. International Journal of Engineering Research Volume No.5, Issue No.1, pp : 75-82



PARTIE II

I-Modélisation :

La modélisation par apprentissage consiste à trouver le jeu de paramètres qui conduit à la meilleure approximation possible de la fonction de régression, à partir des couples entrées/sortie constituant l'ensemble d'apprentissage (ou de calibration) ; le plus souvent, ces couples sont constitués d'un ensemble de vecteurs de variables (descripteurs dans le cas de molécules) $\{ x^i, i = 1.....N\}$, et un ensemble de mesures de la grandeur à modéliser $\{y(x^i), i= 1.....N\}$. La détermination des valeurs de ces paramètres nécessite la mise en œuvre de méthodes d'optimisation qui diffèrent selon le type de modèle choisi.

I-1-Modélisation QSAR, QSPR et QSRR**I-1-1- Historique :**

Les premiers essais de modélisation d'activités de molécules datent de la fin du 19^{ème} siècle, lorsque Crum-Brown et Frazer [1] postulèrent que l'activité biologique d'une molécule est une fonction de sa constitution chimique. Mais ce n'est qu'en 1964 que furent développés les modèles de "contribution de groupes", qui constituent les réels débuts de la modélisation QSAR. Depuis, l'essor de nouvelles techniques de modélisation par apprentissage, linéaires d'abord, puis non linéaires, ont permis la mise en place de nombreuses méthodes ; elles reposent pour la plupart sur la recherche d'une relation entre un ensemble de nombres réels, descripteurs de la molécule, et la propriété ou l'activité que l'on souhaite prédire.

I-1-2- QSPR (Quantitative Structure-Property Relationships):

Est le procédé par lequel des liens quantitatifs sont établis entre la structure moléculaire d'un ensemble de composés avec une propriété physico-chimique. Les grandes phases de développement d'un modèle QSPR peuvent être décrites comme suit :

-Choisir des descripteurs adaptés au problème structure-propriété,

- Exploiter les valeurs des descripteurs comme variables, afin de définir une relation qui les corrèle à la propriété en question, à l'aide de machines d'apprentissage. C'est la fouille de données.
- Établir des critères de performance et de validation qui aideront au choix des meilleurs modèles pour le problème posé et estimer des incertitudes de prédiction [2].

I-1-3- Relation quantitative structure /rétention chromatographique QSRR :

Parmi toutes les méthodes, la relation quantitative entre la structure et la rétention chromatographique (QSRR) est la plus populaire. En QSRR, la rétention des systèmes chromatographiques donnés est modélisée en fonction de descripteurs moléculaires [3].

QSRR est une technique d'analyse utile capable de lier le temps, ou l'indice de rétention chromatographique à la structure chimique. L'objectif principal de QSRR est de prévoir la rétention des données à partir de la structure moléculaire [4]. Cette technique permet de relier les variations dans une (ou plusieurs) variable de réponse (variables Y) aux variations de plusieurs descripteurs (variables X), avec des buts prédictifs ou au moins explicatifs. Les variables Y sont souvent appelées dépendantes et les variables X variables indépendantes. Les variables Y devraient être liées à la rétention chromatographique, les variables X doivent coder la structure moléculaire [5].

I-2-Préparation de la base des données

I-2-1-Les Logiciels utilisés :

I-2-1-1-Chemoffice 2008:

ChemDraw fournit des chimistes avec un ensemble d'outils riche, faciles à utiliser pour créer des publications prêt, dessins scientifiquement significatifs de molécules et de réactions[6,7] .

I-2-1-2-Logiciels « Hyperchem 7.5 » :

HyperChem rassemble dans une même interface un ensemble d'outils dédiés à la modélisation moléculaire, connu pour sa qualité, sa flexibilité, et sa facilité d'usage.

□ Fonctionnalités:

L'HyperChem est le logiciel qui vous permet de faire réellement de la modélisation : il possède plus de méthodes de calculs (mécanique moléculaire, semi-empirique et ab-initio) pour qu'on puisse calculer plus de propriétés, Pour stabiliser la forme de la structure de chaque molécule.

L'Hyperchem est utilisé dans cette étude pour construire et optimiser les molécules aromatique de pyrazine, chaque molécule est enregistrée comme un fichier nommé "Hin" après l'optimisation. On applique la méthode MM+ suivis par la méthode semi empirique AM1 pour l'optimisation présente dans figure suivante (Figure 8)

On a 114 molécules donc on obtient 114 fichiers Hin, ensuite en va calculer les descripteurs moléculaires à partir de ce fichier par logiciel DRAGON pour chaque molécule [6,8].

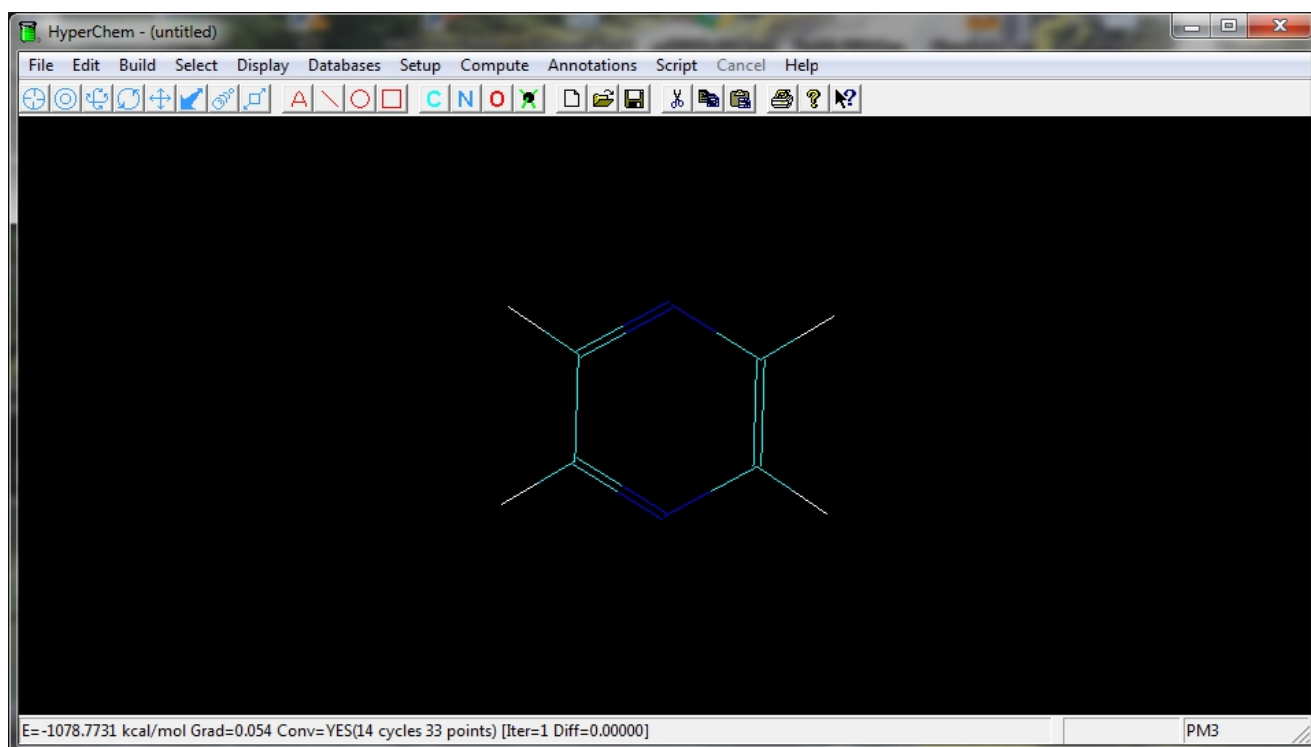


Figure 6 : système d'optimisation par la méthode MM+ et la méthode AM1 pour le pyrazine.

I-3- Calculs des descripteurs moléculaires :

Afin d'exploiter au maximum les informations contenues dans les structures moléculaires, celles-ci sont traduites en une série de grandeurs (en général scalaires) qui quantifient leurs caractéristiques physico-chimiques et structurelles. Dans la prochaine étape pour tous les 114 composés, des descripteurs moléculaires ont été calculés par le logiciel Dragon. Dragon peut calculer 1320 descripteurs moléculaires pour chaque structure dans notre jeu des données.

I-3-1-Le Logiciel « DRAGON» :

DRAGON: c'est une application pour le calcul des descripteurs moléculaires. Ces descripteurs peuvent être utilisés pour évaluer l'influence de la structure moléculaire ou des relations propriétés-structure, ainsi que pour l'analyse de la symétrie et de la projection des bases de données des molécules [6,10].

DRAGON fourni 1320 descripteurs moléculaires divisés en **20** blocs logiques quelque descripteurs comme : Les descripteurs topologiques et géométriques, Les indices de connectivité, Les Poids des valeurs propres, Les indices bord de contiguïté, Les indices de charge topologique, Les descripteurs topologiques, descripteurs de propriétés Moléculaires ,2D Autocorrélations, descripteurs constitutionnels, Profils moléculaires Randic, Descripteurs 3D- Morse, descripteurs RDF, descripteurs WHIM, descripteurs de charge descripteurs et géométriques descripteurs de GETAWAY.

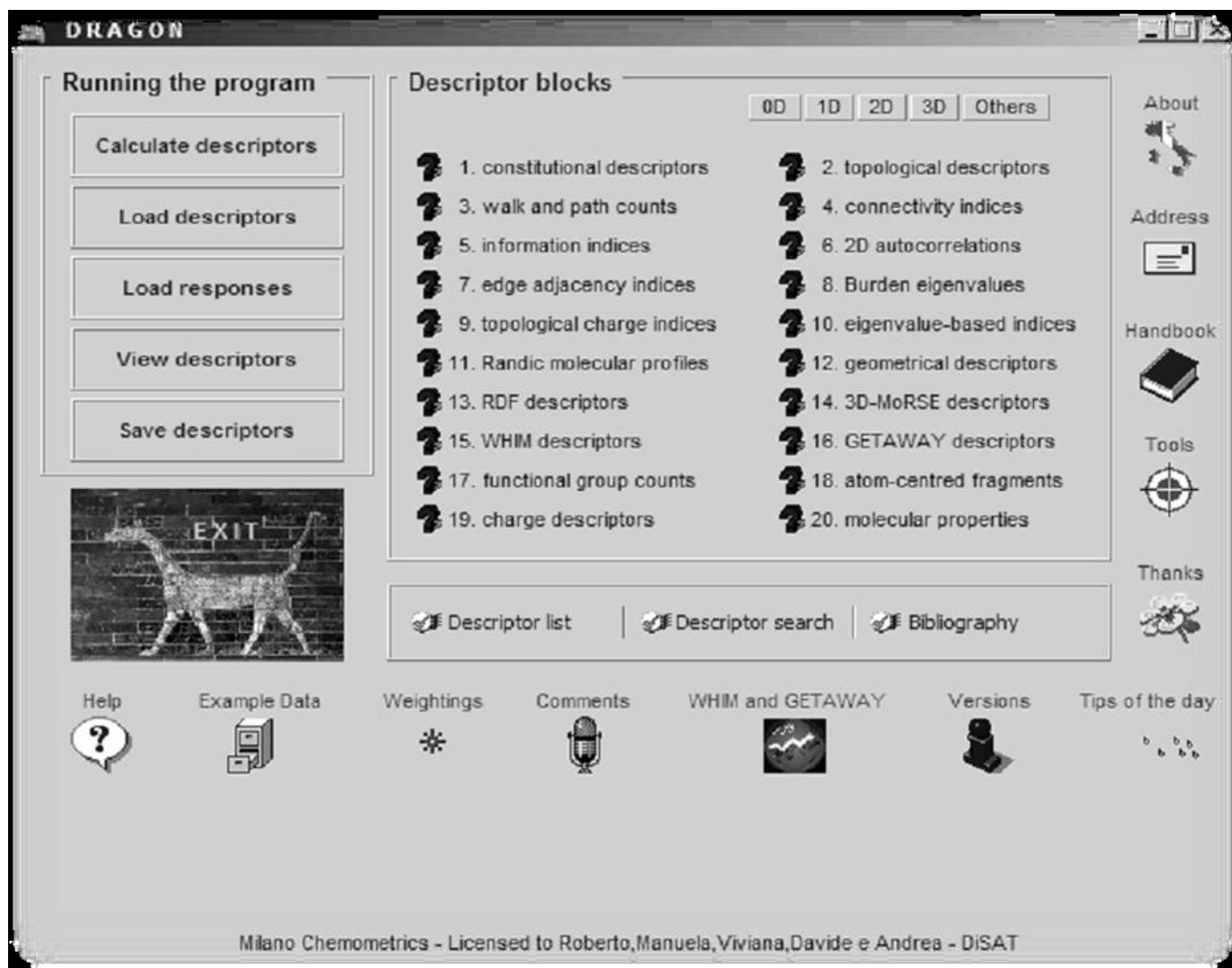


Figure 7 : Forme principale du logiciel de DRAGON.

I-3-1-1- Les descripteurs :

De nombreuses recherches ont été menées, au cours des dernières décennies, pour trouver la meilleure façon de représenter l'information contenue dans la structure des molécules, et ces structures elles-mêmes, en un ensemble de nombres réels appelés descripteurs ; une fois que ces nombres sont disponibles, il est possible d'établir une relation entre ceux-ci et une propriété ou activité moléculaire, à l'aide d'outils de modélisation classiques. Ces descripteurs numériques réalisent de ce fait un codage de l'information chimique en un vecteur de réels.

On en dénombre aujourd'hui plus de 3000 types, qui quantifient des caractéristiques physico-chimiques ou structurelles de molécules. Ils peuvent être obtenus de manière semi

empirique ou non semi-empirique, mais les descripteurs calculés, et non mesurés, sont à privilégier : ils permettent en effet d'effectuer des prédictions sans avoir à synthétiser les molécules, ce qui est un des objectifs de la modélisation. Il existe cependant quelques descripteurs mesurés : il s'agit généralement de données expérimentales plus faciles à mesurer que la propriété ou l'activité à prédire (indice de rétention, temps de rétention, coefficient de partage eau-octanol [36], polarisabilité, ou potentiel d'ionisation).

Avant toute modélisation, il est nécessaire de calculer ou de mesurer un grand nombre de Descripteurs différents, car les mécanismes qui déterminent l'activité d'une molécule ou une de ses propriétés sont fréquemment mal connus. Il faut ensuite sélectionner parmi ces variables celles qui sont les plus pertinentes pour la modélisation.

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 (HyperchemTM, 2000) pour représenter les molécules puis, à l'aide de la méthode semi empirique AM1, et PM3 obtenir les géométries finales. 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 ainsi optimisées ont été transférées dans le logiciel informatique DRAGON (TodeschiniR *et al.*, 2005) pour le calcul de plus de 1320 descripteurs (si l'on tient compte de ceux calculés à l'aide du logiciel Hyperchem) appartenant à 20 classes différentes.

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 ($R \geq 0,9$) et qui véhiculent une

information redondante. Pour chaque paire de descripteurs corrélés, on élimine automatiquement celui qui présente les plus hautes corrélations croisées avec les autres descripteurs.

1 - Descripteurs moléculaires :

Nous allons présenter les descripteurs moléculaires les plus courants, en commençant par les descripteurs les plus simples, qui nécessitent peu de connaissances sur la structure moléculaire, mais véhiculent peu d'informations. Nous verrons ensuite comment les progrès de la modélisation moléculaire ont permis d'accéder à la structure 3D de la molécule, et de calculer des descripteurs à partir de cette structure.

1-1-Descripteurs 1D :

Sont accessibles à partir de la formule brute de la molécule (par exemple $C_4H_4N_2$ pour le pyrazine), et décrivent des propriétés globales du composé. Il s'agit par exemple de sa composition, c'est-à-dire les atomes qui le constituent, ou de sa masse molaire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution.

1-2-Descripteurs 2D :

Sont calculés à partir de la formule développée de la molécule. Ils peuvent être de plusieurs types.

*** Indices constitutionnels :**

Caractérisent les différents composants de la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles...

*** Indices topologiques :**

Peuvent être obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications. Les plus fréquemment utilisés sont l'indice de Wiener [11], l'indice de Randić [12], l'indice de connectivité de valence de Kier-Hall [13] et l'indice de Balaban [14]. L'indice de Wiener permet de caractériser le volume moléculaire

et la ramification d'une molécule : si l'on appelle distance topologique entre deux atomes le plus petit nombre de liaisons séparant ces deux atomes, l'indice de Wiener est égal à la somme de toutes les distances topologiques entre les différentes paires d'atomes de la molécule.

L'indice de Randić est un des descripteurs les plus utilisés ; il peut être interprété comme une mesure de l'aire de la molécule accessible au solvant.

Ces descripteurs 2D reflètent bien les propriétés physiques dans la plupart des cas, mais sont insuffisants pour expliquer de façon satisfaisante certaines propriétés ou activités, telles que les activités biologiques. Des descripteurs, accessibles à partir de la structure 3D des molécules, ont pu être calculés grâce au développement des techniques instrumentales et de nouvelles méthodes théoriques.

1-3-Descripteurs 3D :

D'une molécule sont évalués à partir des positions relatives de ses atomes dans l'espace, et décrivent des caractéristiques plus complexes; leurs calculs nécessitent donc de connaître, le plus souvent par modélisation moléculaire empirique ou *ab initio*, la géométrie 3D de la molécule. Ces descripteurs s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés ou d'activités qui dépendent de la structure 3D. On distingue plusieurs familles importantes de descripteurs 3D :

*** Descripteurs géométriques :**

Les plus importants sont le volume moléculaire, la surface accessible au solvant, le moment principal d'inertie.

Ces descripteurs peuvent être obtenus expérimentalement ou par modélisation moléculaire, empirique ou *ab-initio*. Ils sont basés sur l'arrangement spatial des atomes constituant la molécule et sont définis par les coordonnées des noyaux atomiques et la grosseur de la molécule représentée. Ces descripteurs incluent des informations sur la surface moléculaire

obtenue par les aires de Van Der Waals et leur superposition. Les volumes moléculaires peuvent être obtenus par les volumes de Van Der Waals [16].

*** Descripteurs électroniques :**

Permettent de quantifier différents types d'interactions inter- et intramoléculaires, de grande influence sur l'activité biologique de molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie stérique est minimale, et fait souvent appel à la chimie quantique. Par exemple, les énergies de la plus haute orbitale moléculaire occupée et de la plus basse vacante sont des descripteurs fréquemment sélectionnés. Le moment dipolaire, le potentiel d'ionisation, et différentes énergies relatives à la molécule sont d'autres paramètres importants.

***Descripteurs spectroscopiques :**

Les molécules peuvent être caractérisées par des mesures spectroscopiques, par exemples par leurs fonctions d'onde vibrationnelles. En effet, les vibrations d'une molécule dépendent de la masse des atomes et des forces d'interaction entre ceux-ci ; ces vibrations fournissent donc des informations sur la structure de la molécule et sur sa conformation. Les spectres infrarouges peuvent être obtenus soit de manière expérimentale, soit par calcul théorique, après recherche de la géométrie optimale de la molécule. Ces spectres sont alors codés en vecteurs de descripteurs de taille fixe. Le descripteur EVA [15] est ainsi obtenu à partir des fréquences de vibration de chaque molécule. Les descripteurs de type

MoRSE (Molecule Representation of Structures based on Electron diffraction) [16] sont calculés à partir d'une simulation du spectre infrarouge ; ils font appel au calcul des intensités théoriques de diffraction d'électrons. Par ailleurs, le calcul de certains descripteurs demande une étape

1-4- Descripteurs électrostatiques :

Ces descripteurs reflètent les caractéristiques de la distribution de charge de la molécule. Les charges partielles empiriques dans la molécule sont calculées en utilisant l'approche proposée par Zefirov . Cette méthode est basée sur l'échelle d'électronégativité, Sur la base de ces charges partielles les descripteurs électrostatiques suivantes sont calculés :

* Les charges partielles minimales et maximales dans la molécule (q_{min} , q_{max})

*Les charges partielles minimales et maximales pour les atome (C, N ,O ,...)

* Les indices électroniques topologiques pour toutes les paires d'atomes et pour l'ensemble des liaisons d'atomes liés

*Les charges partielles de la zone du surface (Charged partial surface area (CSPA)) ont été développés par Jurs et al., ces descripteurs sont responsable sur des interactions entre les molécules polaires [17]..

1-5-Descripteurs quantiques :

Les descripteurs de chimie quantique donnent des informations importantes pour la molécule. Ces descripteurs permettent de quantifier différents types d'interactions inter-et intramoléculaires, de grande influence sur des propriétés physicochimiques de molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie est minimale, et fait appel aux approches de chimie quantique. On cite toutes les données énergétiques, vibrationnelles et orbitales du système par exemple, les énergies de la plus haute orbitale moléculaire occupée HOMO et de la plus basse vacante LUMO, le moment dipolaire, la polarisabilité, le potentiel d'ionisation. [17].

Les descripteurs quantiques sont classés comme suit :

*** Descripteurs liés à la distribution de charge :**

Ces descripteurs représentent ou dépendent directement aux calculs de chimie quantique de la distribution de charge dans les molécules, qui décrivent donc les interactions entre les molécules polaires ou leur réactivité chimique.

-Nombre de niveaux électroniques doublement remplis (nf)

- Les valeurs (minimum et maximum) des charges partielles sur les atomes présentés dans la molécule, par exemple, $dH(\min)$ est la charge minimum (négatif) partielle sur un atome d'hydrogène dans la molécule donnée.

Dans le cadre de la théorie LCAO-MO, les populations de Mulliken fournissent une méthode de calcul des charges atomiques.

-Le moment dipolaire total de la molécule (m), et son point de charge (mc).

*** Descripteurs liés à l'énergie :**

Ces descripteurs caractérisent l'énergie totale de la molécule et la distribution d'énergie intramoléculaire en utilisant différents schémas de partitionnement.

-La chaleur de formation de la molécule (DHF) donne l'énergie de la molécule dans l'état standard (T à 298.15K et P à 1 atm).

-L'énergie totale de la molécule (Etot).

-Le potentiel du premier et la deuxième ionisation de la molécule.

-HOMO - LUMO, La différence d'énergie entre ces deux niveaux HOMO et LUMO, appelé "gap_{HOMO-LUMO}" est un bon indicateur de la stabilité de la molécule[17].

1-6-Descripteurs thermodynamiques :

Les descripteurs thermodynamiques sont calculés sur la base de la fonction de partition totale Q de la molécule, La fonction de partition commode la façon avec laquelle l'énergie d'un système de molécules est répartie parmi les individus moléculaires.

Sa valeur dépend du poids moléculaire, de la température, du volume moléculaire, des distances inter nucléaires, des mouvements moléculaires et des forces intermoléculaires.

La fonction de partition est le point le plus commode entre les propriétés microscopiques des molécules individuelles (niveaux d'énergie, moments d'inertie) avec les propriétés macroscopiques (chaleur spécifique, entropie). La molécule peut accroître son énergie de translation, de vibration, de rotation de façon pratiquement indépendante [17].

I-4- Développement de modèles QSAR/QSPR :

I-4-1- Sélection d'un sous- ensemble de variables par algorithme génétique (GA- VSS) :

Les algorithmes génétiques fournissent des solutions aux problèmes n'ayant pas de solutions calculables en temps raisonnable de façon analytique ou algorithmique. Selon cette méthode, des milliers de solutions (génotypes) plus au moins bonnes sont créées au hasard puis sont soumises à un procédé d'évaluation de la pertinence de la solution mimant l'évolution des espèces : les plus "adaptés", c'est- à- dire les solutions au problème qui sont les plus optimales survivent davantage, que celles qui le sont moins et la population évolue par générations successives en croisant les meilleures solutions entre elles et les faisant muter, puis en relançant ce procédé un certain nombre de fois afin d'essayer de tendre vers la solution optimale.

Les algorithmes génétiques constituent une méthode de choix pour la sélection de sous-ensembles de variables explicatives.

Dans un algorithme génétique adapté à l'optimisation, une solution potentielle est considérée comme un individu dans une population. La valeur de la fonction de coût associée à une solution mesure « l'adaptation » de l'individu associé à son environnement. Un algorithme génétique simule l'évolution, sur plusieurs générations, d'une population initiale dont les individus sont mal adaptés au moyen d'opérateurs génétiques de reproduction et de mutation. Après un certain nombre de générations, la population est constituée d'individus bien adaptés,

autrement dit des solutions supposées « bonnes » au problème d'optimisation.

Dans le présent travail, la sélection des descripteurs a été réalisée par algorithme génétique, dans la version MOBY DIGS de Todeschini (Todeschini *Ret al.*, 2004)[18], en maximisant Q^2

LOO.

I-4-2- Méthodes utilisées pour le développement de modèles QSAR/QSPR :

L'application pratique des gammes des descripteurs moléculaires dans le développement de modèles QSAR/QSPR n'est pas une tâche aisée (Todeschini *Ret al.*, 2005)[82].

Tout d'abord, un très grand nombre (>3000) de descripteurs moléculaires, de différentes complexités et de conceptions diverses ont été imaginés et proposés au cours des (60 dernières) années. Ensuite, pendant ce temps, aucune règle stricte n'a été établie, ni même proposée, pour la sélection de descripteurs adaptés parmi la myriade de descripteurs disponibles. Ce choix a souvent été basé sur l'intuition chimique des chercheurs, ou en se pliant à la tradition.

Une autre difficulté dans la sélection des descripteurs QSAR/QSPR découle du non standardisation des gammes de descripteurs. Les gammes empiriques des constantes d'induction, de résonance et d'effet stérique des constituants, ou les échelles empiriques d'effets de solvant comportent des erreurs intrinsèques liées aux erreurs respectives des mesures expérimentales. Par ailleurs, les méthodes quanto- mécaniques appliquées aux calculs des descripteurs moléculaires et aux distributions de charges liés aux OM sont souvent basées sur différents paramètres semi- empiriques, ou l'utilisation de différents ensembles de base dans les calculs ab- initio. Naturellement, un descripteur construit à l'aide de différentes méthodes expérimentales ou théoriques, pour divers composés, ne peut être utilisé pour le calcul d'un modèle QSAR/QSPR unique. Une approche systématique pour la sélection de gammes de descripteurs pour le calcul de modèles QSAR/QSPR est basée sur la discrimination statistique entre de larges ensembles de descripteurs.

Dans ce qui suit nous passerons en revue diverses approches utilisées pour le développement des " meilleures" équations QSPR pour un espace plus grand de descripteurs.

En dernier ressort, les modèles QSAR/QSPR peuvent être développés selon des modèles mathématiques différents, généralement en relation avec l'analyse statistique multivariée. Les modèles ont été développés en utilisant la méthode des moindres carrés ordinaires (MCO) parce que la distribution d'erreur est normale mais la présence de autocorrélation pose le problème de la présence des points aberrants alors on applique avec la méthode des moindres carrés la technique de régression par les moindres absolus des erreurs (Least Absolute Deviation LAD) et complète le travail par l'utilisation de l'analyse factorielle ou l'analyse en composantes principales. L'intérêt de ces méthodes est qu'elles évacuent le problème de multicollinéarité inhérent aux méthodes de régression linéaires. Cependant, l'interprétation des équations QSAR/QSPR est alors entravée par la nature formelle des facteurs ou des composantes principaux. Une alternative aux méthodes très classiques de régression linéaire multiple (MLR) et d'analyse en composantes principales (ACP) est la technique de régression par les moindres carrés partiels (PLS)[19,20]), et Least absolute Deviation(LAD), le traitement d'analyse statistique et graphique pour les deux méthodes (MLR et PLS) on utilise logiciel MINITAB16.

Nous présenterons dans ce qui suit une courte vue d'ensemble des différentes méthodes mathématiques utilisées pour développer nos modèles.

Nous allons présenter les trois types des méthodes avec l'analyse statistique et graphique exploités dans cette thèse.

I-4-2-1-Régression multiple [42]:

Dans cette section, on considère le modèle de régression multiple donné par :

$$y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \varepsilon_i \quad (5)$$

Les x_{ij} et les y_i étant n données relatives à $(p-1)$ variables explicatives et à une variable expliquée. A partir d'estimations $\hat{\beta}_j$ des paramètres β_j , on obtient des valeurs estimées :

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{ij} \quad (6)$$

Qui ne trouvent pas sur une droite comme en régression simple, mais sur un hyperplan de dimension $(p-1)$, ainsi que des résidus :

$$e_i = y_i - \hat{y}_i \quad (7)$$

Dans ce travail on utilise trois méthodes sur régression multiples :

1-Régression LAD multiple [42]:

La méthode des moindres écarts en valeur absolue, dite la méthode LAD (Least Absolute Déviations), est l'une des principales alternatives à la méthode des moindres carrés lorsqu'il s'agit d'estimer les paramètres d'un modèle de régression. Elle a été introduite presque cinquante ans avant la méthode des moindres carrés, en 1757 par Roger Joseph Boscovich. Il utilisa cette procédure dans le but de concilier des mesures incohérentes dans le cadre de l'estimation de la forme de la terre. Pierre Simon Laplace adopta cette méthode trente ans plus tard, mais elle fut ensuite éclipsée par la méthode des moindres carrés principalement sur la simplicité des calculs. Mais aujourd'hui, avec les progrès de l'informatique, la méthode LAD peut être utilisée presque aussi simplement.

La méthode LAD à la régression multiple consiste à définir les estimateurs $\hat{\beta}_j$ qui minimisent :

$$\sum |e_i| = \sum |y_i - \hat{\beta}_0 - \sum_{j=1}^{p-1} \hat{\beta}_j x_{ij}| \quad (8)$$

Il s'agit tout d'abord de choisir (arbitrairement) p points de la donnée, que l'on indexe par $i=1, \dots, p$. Soit la matrice carrée :

$$A = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2p-1} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{p1} & x_{p2} & \dots & x_{pp-1} \end{pmatrix}$$

Et vecteur :

$$c = (y_1, \dots, y_p) \quad (9)$$

Notre première estimation $\hat{\beta}$ du vecteur des vecteurs des paramètres :

$$\beta = (\beta_0, \beta_1, \dots, \beta_{p-1}) \quad (10)$$

est déterminée de telle sorte que les p points en question satisfassent l'équation on donc :

A chaque étape de cet algorithme on va améliorer cet estimateur $\hat{\beta}$ par un estimateur $\hat{\beta}^*$ définir par :

$$\hat{\beta}^* = \hat{\beta} + t * d \quad (11)$$

Pour une certaine valeur de t et pour un certain vecteur d .

Pour un vecteur d donné, la valeur de t est déterminée de telle sorte qu'elle minimise :

$$\sum |y_i - x_i * (\hat{\beta} + t * d)| \quad (12)$$

Ou pour $i=1, \dots, n$

$$x_i = (1, x_{i1}, \dots, x_{i(p-1)}) \quad (13)$$

On notant :

$$z_i = y_i - x_i \hat{\beta} \quad (14)$$

et :

$$w_i = x_i d \quad (15)$$

L'expression est égale à :

$$\sum |y_i - x_i * (\hat{\beta} - tx_i * d)| = \sum |z_i - tw_i| \quad (16)$$

On a alors le même problème que régression simple. On considère les ratios z_i/w_i pour les $(n-p)$ points ayant des résidus non nuls, on les range par ordre croissant en les renumérotant de telle sorte que :

$$Z_{p+1}/w_{p+1} \leq Z_{p+2}/w_{p+2} \leq \dots \leq Z_n/w_n \quad (17)$$

et on détermine l'indice K qui satisfait :

$$|w_{p+1}| + |w_{p+2}| + \dots + |w_{k-1}| < T/2 \quad (18)$$

$$|w_{p+1}| + |w_{p+2}| + \dots + |w_{k-1}| + |w_k| > T/2 \quad (19)$$

$$T = \sum_{i=p+1}^n |w_i| \quad (20)$$

La solution pour t est alors :

$$t = z_k/w_k \quad (21)$$

D'autre part, on choisit le vecteur \mathbf{d} parmi les p vecteurs qui constituent les colonnes de A^{-1} et parmi les p vecteurs opposés. On a donc le choix entre $2p$ vecteurs pour \mathbf{d} . Parmi ces $2p$ vecteurs, on choisit celui pour lequel la quantité décroît le plus rapidement vers 0. Il s'agit alors de calculer pour chacun de ces vecteurs la dérivée à droite de en $t=0$. En utilisant la formulation de, ces dérivées sont égales à :

$$w = w_- + w_0 - w_+ \quad (22)$$

Où w_- est la somme des $|w_i|$ pour lesquels les ratios z_i/w_i sont négatifs, w_0 est la somme des $|w_i|$ pour lesquels $z_i=0$ et w_+ est la somme des $|w_i|$ pour lesquels les ratios z_i/w_i sont positifs.

Si toutes ces dérivées ne sont positives cela veut dire que notre solution $\hat{\beta}$ est l'estimateur LAD de β . Si ces dérivées ne sont pas positives, on choisit le vecteur \mathbf{d} pour lequel la dérivée est plus négative, et la valeur de t par. On recommence alors le processus avec $\hat{\beta}^*$ à la place de $\hat{\beta}$.

Notons que si le vecteur \mathbf{d} choisit correspond à la jème colonne de A^{-1} (ou à l'opposé de ce vecteur), l'hyperplan défini par la nouvelle estimation $\hat{\beta}^*$ contient les mêmes points que l'hyperplan défini par l'estimateur précédent $\hat{\beta}$, sauf que le jème point est remplacé par le $K^{\text{ème}}$ point (l'indice k étant celui utilisé pour définir t). L'algorithme revient donc à remplacer l'un des p points définissant $\hat{\beta}^{(en)}$ (l'occurrence le jème) par l'un des $(n-p)$ points restants (en l'occurrence le $k^{\text{ème}}$ pour définir l'hyperplan).

1-2-Tests d'hypothèses sur les β_j :

Nous allons voir dans cette section comment tester l'hyperplan nul :

$$H_0: \beta_1 = \dots = \beta_q = 0$$

(en renumérotant au besoin les variables explicatives), contre l'hyperplan H_1 stipulant qu'au moins un de ces paramètres est différent de 0.

Pour tester cette hypothèse avec la méthode LAD, on calcule une statistique similaire :

$$F_{LAD} = \frac{SA_{res}(H_0) - SA_{res}}{q * (\hat{\sigma}^2 / 2)} \quad (23)$$

$$SA_{res} = \sum |e_i| \quad (24)$$

Est la quantité minimisée par la régression LAD lorsque l'on utilise les $(p-1)$ variables explicatives, et où $SA_{res}(H_0)$ est cette même quantité lorsque l'on n'utilise que les $(p-1-q)$ variables explicatives restantes dans le modèle si H_0 est acceptée. Quant à l'estimation $\hat{\sigma}^2$, il est donné par, sauf prendre ici $m=n-p$ (ce qui correspond au nombre de résidus non nuls de la régression LAD). On rejette ici aussi l'hypothèse nulle au seuil de signification α lorsque la statistique F_{LAD} est supérieure à la valeur critique $F_{(\alpha, q, n-p)}$ issue d'une table de Fisher. Ce test est d'autant plus valide que la taille n de l'échantillon est grande.

2- Régression linéaire multiple (MLR) [51, 56,67] :

La régression linéaire multiple est la méthode la plus simple de modélisation. elle consiste à rechercher une équation linéaire par rapport à ses paramètres reliant la variable à modéliser au vecteur d'entrées $x = \{x_k, k = 1, \dots, q\}$. Ces entrées peuvent être des fonctions non paramétrées, ou à paramètres fixés, de ces variables. L'équation linéaire recherché est de la forme :

$$g(x, \theta) = \sum_{k=1}^q \theta_k x_k = X\theta \quad (25)$$

où $\theta = \{\theta_k, k=1, \dots, q\}$ est le vecteur des paramètres; X , matrice des observations de taille (N, q) , est définie comme la matrice dont les éléments de la colonne k prennent pour valeurs les N mesures de la variable k . Pour chaque élément i de la base d'apprentissage, le résidu est défini comme la différence entre la valeur de la grandeur à modéliser pour cet élément i et l'estimation du modèle :

$$R_i = y^i - g(x^i, \theta) \quad (26)$$

L'apprentissage est réalisé par minimisation de la fonction de coût des moindres carrés, qui mesure l'ajustement du modèle g aux données d'apprentissage :

$$J(\theta) = \sum_{i=1}^N (R_i)^2 = \sum_{i=1}^N [y^i - g(x^i, \theta)]^2 = \|y - X\theta\|^2 \quad (27)$$

La fonction $J(\theta)$ est une fonction positive quadratique en θ : son minimum est unique. Il est donné par :

$$\theta_{mc} = (X^T X)^{-1} X^T y \quad (28)$$

Les paramètres θ_k sont appelés coefficients de régression partielle ; chacun d'eux mesure l'effet de la variable explicative x_k concernée sur la propriété modélisée lorsque les autres variables explicatives sont maintenues constantes.

La régression linéaire est facile à mettre en œuvre, et les coefficients θ_k obtenus peuvent être interprétés : ils mesurent l'influence de chacune des variables sur les grandeur étudiée.

Cependant, il est souvent nécessaire d'avoir recours à des modèles de plus grande complexité.

3-Analyse en composantes principales (ACP)[27-32]:

L'analyse en composantes principales est une des techniques les plus anciennes et les plus connues de l'analyse multi-variée [21, 22, 23, 24, 25,26]. L'ACP a été « inventé » en 1901 par Karl Pearson [27]). Actuellement, l'ACP est utilisé comme outil d'exploration et d'analyse de données ainsi que pour la conception de modèles. L'analyse par composantes principales (PCA, pour Principal Component Analysis) consiste à transformer un jeu de variables corrélées entre elles en un nouveau jeu de variables, appelées composantes principales, moins nombreuses mais indépendantes. En utilisant ces nouvelles variables, la dimensionnalité du système est réduite en perdant un minimum d'information.

En général, une composante principale est une combinaison linéaire des variables :

$$p_i = \sum_{j=1}^v c_{ij} x_j \quad (29)$$

ou p_i est la i ème composante principale et c_{ij} le coefficient de la variable x_j . Il y a un nombre v de telles variables. La première composante principale d'un ensemble de données correspond à la combinaison linéaire des variables qui conduit à la droite la mieux ajustée aux données quand elles sont représentées dans l'espace de dimension v . Plus précisément, la première composante principale maximise la variance des données, de sorte que leur dispersion soit maximale sur la première composante principale. Le second axe principal, et les suivants, tiennent compte de la variance maximale des données non encore prises en compte par les précédents axes principaux. Chaque composante principale correspond à un axe dans l'espace de dimension v , et chaque composante principale est orthogonale à toutes les autres composantes principales. Le nombre de composantes principales possibles est égal à

la dimension des données originales et, effectivement, pour expliquer complètement la variabilité des données on est amené à incorporer toutes les composantes principales. Cependant, dans de nombreux cas, seul un petit nombre de composantes principales sera nécessaire pour expliquer une proportion significative de la variabilité des données. Si seulement une ou deux composantes principales permettent d'expliquer la plupart des données, alors une représentation graphique est possible.

Les composantes principales sont calculées en utilisant les techniques matricielles standards [36]. La première étape consiste à calculer la matrice de variance covariance. S'il y a s observations, chacune contenant v valeurs, alors l'ensemble des données peut être représenté par une matrice V avec v lignes et s colonnes. La matrice Z de variance-covariance est alors :

$$Z = R^T R = V^T V \quad (30)$$

Les vecteurs propres de Z sont les coefficients des composantes principales. Comme Z est une matrice carrée symétrique, ses vecteurs propres peuvent être orthogonaux (à condition qu'il n'y ait pas de valeurs propres dégénérées). Les valeurs propres et leurs vecteurs propres associés peuvent être obtenus en résolvant l'équation séculaire :

$$|Z - \lambda I| = 0 \quad (31)$$

, ou par triangulation

de matrice.

La première composante principale correspond à la plus grande valeur propre, la seconde composante principale à la 2ème plus grande valeur propre et ainsi de suite. La i ème composante propre tient compte d'une proportion : $\frac{\lambda_i}{\sum_{j=1}^v \lambda_j}$ de la variance totale des données.

4- Régression en composantes principales (RCP) :

La régression linéaire multiple ne peut s'appliquer à des ensembles de données où les variables sont hautement corrélées et /ou le nombre de variables excède celui des valeurs observées. Dans de telles situations deux méthodes sont largement utilisées : la régression en

composantes principales et les moindres carrés partiels.

Dans la régression en composantes principales on soumet d'abord les variables à une ACP, puis l'analyse de régression est opérée sur les premières composantes principales en nombre limité. Lorsqu'on réalise une régression en composantes principales par, disons, sélection progressive alors l'équation résultante ne s'exprimera pas nécessairement en fonction des composantes principales les plus basses. Ceci est dû au fait que l'ordre des composantes principales correspond à leur capacité à expliquer la variance des variables indépendantes, alors que l'analyse de régression concerne l'explication de la variable dépendante. En règle générale seules les composantes principales dont les valeurs propres sont inférieures à 1 seront insérées dans les régressions en composantes principales.

Lorsqu'une valeur propre est inférieure à 1, alors une des variables originales de l'ensemble est mieux à même d'expliquer la variance que la composante principale. Néanmoins, et c'est souvent le cas à la limite, les 2 premières composantes conduisent à la meilleure corrélation avec la variable dépendante. Un autre fait à souligner en RCP est que, lorsqu'on incorpore de plus en plus de composantes principales, les coefficients des régresseurs déjà présents ne changent plus. Ceci est dû à l'orthogonalité des composantes principales, et parce que le rôle de chaque nouvelle composante principale est d'expliquer la variance non encore couverte.

5- La régression PLS[29].

La régression PLS (Partial Least Squares regression) tire son origine des sciences sociales, plus précisément des sciences économiques par Herman Wold 1966 (Gauchi J P, 1995)[27] mais devient très populaire en chimie grâce au fils d'Herman, Svante. Cette méthode connaît un très grand succès dans le domaine de la chimie, particulièrement dans les applications concernant des données de chromatographie ou de spectrographie. De plus Svante Wold, Nouna Kettanech- Wold et leurs collaborateurs ont développé le logiciel d'analyse des données SIMCA-P for Windows centré sur la régression PLS. Signalons également l'avantage

de la régression PLS par rapport à d'autres méthodes de régression dans l'analyse des plans d'expériences non orthogonaux (Vancolen S, 2004)[29].

La régression PLS est une technique récente qui généralise et combine les caractéristiques de l'analyse sur composantes principales et de la régression multiple. Elle est particulièrement utile quand on a besoin de prédire un ensemble de variables dépendantes à partir d'un ensemble très grand de variables explicatives qui peuvent être très fortement corrélées entre elles. La PLS est donc une méthode pour construire des modèles de prédiction quand les facteurs sont nombreux et très colinéaires.

Notons que cette méthode met l'accent sur la prédiction de la réponse et pas nécessairement sur la mise en évidence d'une relation entre les variables. Ce qui signifie que la PLS n'est pas appropriée pour désigner les variables ayant un effet négligeable sur la réponse, mais quand le but est la prédiction et qu'il n'y a pas besoin de limiter le nombre de variables mesurées, la PLS est un outil très utile [30].

Comparé à d'autres méthodes de régression pour des données colinéaires, il a été établi que le plus grand avantage de la PLS est que l'information dans la variable Y est utilisée. Un avantage plus évident est que la méthode rend possible la combinaison de la prédiction avec l'étude d'une structure jointe latente dans les variables X et Y. Ainsi la méthode demande souvent moins de composantes que la PCR pour donner une bonne prédiction [28] .

La régression PLS est évidemment liée à la corrélation canonique et à l'analyse des facteurs multiples. Ces relations sont explorées en détail par Tenenhaus [30], Pagès et Tenenhaus [31]. La principale originalité de la régression PLS est de préserver l'asymétrie de la relation entre les prédicteurs et les variables dépendantes, contrairement aux autres techniques qui les traitent symétriquement.

La régression par les moindres carrés partiels (PLS, pour Partial Least Squares ou Projection on Latent Structures) [30] est en quelque sorte une version supervisée de la APC. Il

s'agit dans ce cas de considérer deux types de variables : une ou des variable(s) dépendante(s) Y_i – dans le cas d'une analyse QSPR, une ou des propriété(s) – dont la variance est expliquée par un nombre de variables indépendantes X_i – les descripteurs moléculaires.

La PLS repose sur une projection des X_i sur des composantes principales, comme dans le cadre de la ACP, à la différence près qu'ici, cette projection est guidée par leur relation avec les Y_i . Le système peut alors être analysé, comme dans le cas d'une PCA à partir de la matrice des coordonnées et des poids. En plus des informations données par les représentations de ces matrices, l'importance des variables dans le modèle est traduite au travers d'un indice, le VIP (pour Variable Importance in the Projection) (SIMCA-P, SIMCAP+, 2005 Minitab 16).

Un des avantages de cette méthode de régression réside dans le traitement de bases de données de grande taille présentant de nombreuses variables corrélées entre elles [31]. On trouve donc des utilisations de cette méthode pour d'autres types d'applications telles que le traitement d'images [32].

5-1-Les données statistiques et les représentations graphiques pour la fonction Analyse en composantes principales.

5-1-1-Valeur propre :

Les valeurs propres (également appelées valeurs caractéristiques ou racine latente) sont les variances des composantes principales [35].

5-1-2-Proportion :

La valeur Proportion désigne la proportion de la variabilité des données expliquée par chaque composante principale [35], pas suffisamment importante pour être incluse.

5-1-3-Cumulé :

La valeur Cumulé est la proportion cumulée de la variabilité de l'échantillon représentée par des composantes principales consécutives [35].

5-1-4-Composantes principales (CP) :

Les composantes principales sont les combinaisons linéaires des variables d'origine qui rendent compte de la variance des données. Le nombre maximal de composantes extraites est toujours égal au nombre de variables. Les vecteurs propres, constitués de coefficients correspondant à chaque variable, sont utilisés pour calculer les scores des composantes principales. Les coefficients indiquent la pondération relative de chaque variable dans la composante [5].

5-1-5-Diagramme en cône :

Le diagramme en cône organise les valeurs propres par ordre décroissant. Idéalement, la courbe doit d'abord décrire une pente forte, puis s'incurver avant de poursuivre en ligne droite. Utilisez les composantes correspondant à la partie abrupte de la courbe, c'est-à-dire avant le point marquant le début de la portion en ligne droite [35].

5-1-6-Diagramme des contributions :

Le diagramme des contributions représente les coefficients de chaque variable pour la première composante par rapport à ceux associés à la deuxième composante [35].

5-1-7-Diagramme des valeurs aberrantes :

Le diagramme des valeurs aberrantes affiche la distance de Mahalanobis pour chaque observation, ainsi qu'une ligne de référence permettent de détecter les valeurs aberrantes. La distance de Mahalanobis est la distance entre chaque point de données et le centre d'un espace multivarié (la moyenne globale). Etudier les distances de Mahalanobis est une méthode multivariée plus performante que l'examen des variables une par une pour détecter des valeurs aberrantes, car elle prend en compte les différentes échelles entre les variables, ainsi que leurs corrélations [35].

5-1-8-Diagramme de sélection des modèles :

Le diagramme de sélection des modèles est un nuage de points des valeurs de R^2 et de R^2 prévu en tant que fonction du nombre de composantes ajustées ou à validation croisée. Il s'agit d'une représentation graphique du tableau de sélection et validation de modèle. Si vous n'optez pas pour la validation croisée, les valeurs de R^2 prévu ne figurent pas dans votre diagramme. Minitab fournit un diagramme de sélection de modèle par réponse[35].

5-1-9-Diagramme des réponses :

Le diagramme des réponses est un nuage de points des valeurs ajustées en fonction des réponses réelles. Si vous réalisez une validation croisée, le diagramme inclut également les valeurs ajustées en fonction des valeurs ajustées à validation croisée. Minitab propose un diagramme de réponses pour chaque réponse [35].

5-1-10-Diagramme des coefficients normalisés :

Le diagramme des coefficients est un nuage de points projeté représentant les coefficients normalisés de chaque prédicteur. Minitab propose un diagramme des coefficients normalisés pour chaque réponse [35].

6-Tests sur le modèle linéaire [78] :

Comme pour le modèle linéaire simple, les hypothèses de régression linéaire doivent être vérifiées pour un modèle de régression multiple.

6-1-Test de la signification globale de la régression (F-Fisher) :

Ce test permet de connaître l'apport global de l'ensemble des variables X_1, \dots, X_p à la détermination de Y .

On veut tester l'hypothèse nulle:

$H_0: \beta_1 = \dots = \beta_p = 0$ contre H_a : il existe au moins un β_j parmi β_1, \dots, β_p non égal à 0.

On calcule la statistique de test $F = MS_{\text{model}} / MS_{\text{error}}$ (32)

6-2-Test de signification de chaque paramètre (chaque descripteur) t-Student :

Pour voir la contribution de chaque paramètre dans l'explication de la variable dépendante Y on utilise la statistique « t » définie auparavant en régression simple.

A partir de cette statistique, il est possible de tester un à un la nullité des différents paramètres du modèle de régression linéaire multiple et de construire des intervalles de confiance sur ces paramètres, très utiles lors de la phase d'interprétation du modèle.

On calcule t-test pour chaque paramètre $i \hat{\beta}_i$ en utilisant la formule ci-dessous

$t_{\text{observe}} = \hat{\beta}_i / s(\hat{\beta}_i)$ avec $s(\hat{\beta}_i)$ est l'erreur type du paramètre $\hat{\beta}_i$.

I-5- DEVELOPPEMENT ET EVALUATION DE MODELE**I-5-1 Sélection d'un sous-ensemble de descripteurs :**

Des logiciels spécialisés permettent le calcul de plus de 6000 descripteurs moléculaires appartenant à différentes classes. Plutôt que de chercher à expliquer la variable dépendante (grandeur d'intérêt) par tous les régresseurs (descripteurs moléculaires), on peut chercher seulement un ensemble réduit de régresseurs i qui donne une reconstitution aussi satisfaisante de la variable à expliquer. Parmi les stratégies mises en œuvre pour la sélection d'un ensemble réduit de variables explicatives, on peut citer : les méthodes de pas-à-pas (méthode descendante; méthode ascendante, et méthode dite stepwise), ainsi que les algorithmes génétiques.

En général, la comparaison se fait à l'avantage des algorithmes génétiques (GA) que nous avons appliqués dans le présent travail, et que nous rappelons succinctement.

I-5-2 Principe :

Dans la terminologie des algorithmes génétiques, le vecteur binaire \tilde{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" ; (cf. infra), 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.

I-5-3 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) ;

I-6 Développement des modèles :

Les techniques les plus courantes pour établir des modèles QSAR utilisent l'analyse de régression (régression linéaire multiple : RLM ; projection des structures latentes par les moindres carrés partiels : PLS).

I-6-1 Paramètres d'évaluation de la qualité de l'ajustement :

Deux paramètres sont couramment utilisés :

Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \quad (33)$$

où \hat{y}_i est la valeur estimée du paramètre physique, et \bar{y} la moyenne des valeurs observées.

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

$$\text{EQMP} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2} \quad (34)$$

I-6-2 Robustesse du modèle :

La stabilité du modèle a été explorée en utilisant la "validation croisée par omission d'une observation" (LOO : cross-validation by leave-one-out) [40]. Elle consiste à recalculer le modèle sur $(n - 1)$ composés de calibrage, le modèle obtenu servant alors à estimer la valeur de la propriété du composé éliminé noté $\hat{y}_{(i)}$. On répète le procédé pour chacun des n composés de l'ensemble de calibrage.

"La somme des carrés des erreurs de prédiction", désignée par l'acronyme PRESS (pour : Predictive Residual Sum of Squares) :

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 \quad (35)$$

est une mesure de la dispersion de ces estimations. On l'utilise pour définir le coefficient de prédiction :

$$Q_{\text{LOO}}^2 = \frac{\text{SCT} - \text{PRESS}}{\text{SCT}} \quad (36)$$

Contrairement à R^2 , qui augmente avec le nombre de paramètres de la régression, le facteur Q_{LOO}^2 affiche une courbe avec maximum (ou avec palier), obtenu pour un certain nombre de variables explicatives, puis décroît par la suite de façon monotone. Ce fait confère une grande importance au coefficient Q_{LOO}^2 . Une valeur de $Q_{\text{LOO}}^2 > 0,5$ est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [36].

I-6-3 Domaine d'application :

Le domaine d'application a été discuté à l'aide du diagramme de Williams (traité en détail dans [44 ,45], représentant les résidus de prédiction standardisés en fonction des valeurs des leviers h_i . L'équation (56) définit le levier d'un composé dans l'espace original des variables indépendantes (x_i)

$$h_i = \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \quad (37)$$

Où (x_i) est le vecteur ligne des descripteurs du composé i et X ($n \times p$) la matrice du modèle déduite des valeurs des descripteurs de l'ensemble de calibration ; l'indice T désigne le vecteur (ou la matrice) transposé (e).

La valeur critique du levier (h^*) est fixée à $3(p+1)/n$. Si $h_{ii} < h^*$, la probabilité d'accord entre les valeurs mesurée et prédite du composé i est aussi élevée que celle des composés de calibration. Les composés avec $h_{ii} > h^*$ renforcent le modèle quand ils appartiennent à l'ensemble de calibration, mais auront, sinon, des valeurs prédites douteuses sans pour autant être forcément aberrantes, les résidus pouvant être bas.

I-6-4 Validation externe :

En plus du test de randomisation, il est intéressant [37], pour juger la qualité du modèle, de considérer le coefficient de prédiction externe calculé comme suit :

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{next} (\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{ntr} (y_i - \bar{y}_{tr})^2 / n_{tr}} \quad (38)$$

la racine de l'écart quadratique moyen (RMSE pour Root Mean Squared Error), calculée sur différents ensembles:

- Ensemble d'estimation (appelée EQMC)
- Ensemble de prédiction externe (désignée par EQMPext).

Ces valeurs RMSE sont mieux adaptées, pour juger de la qualité d'un modèle que les valeurs de R^2 et Q^2 seules, qui constituent de bons tests uniquement pour des données réparties régulièrement.

$$EQMC = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (39)$$

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}{n_{ext}}} \quad (40).$$

I-7- Test de Durbin-Watson

I-7-1-Principe:

Alors le test de Durbin-Watson permet de vérifier si le résidu en i est non-corrélé au résidu en $(i+1)$: on parle d'autocorrélation d'ordre 1. Il est obtenu par l'option DW de l'instruction MODEL de REG.

On calcule ainsi le coefficient de Durbin-Watson à partir des résidus :

$$e_i = y_i - \hat{y}_i$$

$$DW = \frac{\sum_i (e_{i+1} - e_i)^2}{\sum_i e_i^2} \quad (41)$$

$$\text{En notant } \rho = \frac{\sum_i (e_{i+1} - e_i)}{\sum_i e_i^2} \quad (42)$$

si les résidus forment un processus autorégressif d'ordre 1, c'est-à-dire suivent le modèle

$$e_{i+1} = \rho * e_i - \eta \quad (43)$$

alors DW vaut à peu près $(2 - 2\rho)$, où :

$$DW \cong 2 * \left(1 - \frac{\sum_i (e_{i+1} - e_i)}{\sum_i e_i^2}\right). \quad (44)$$

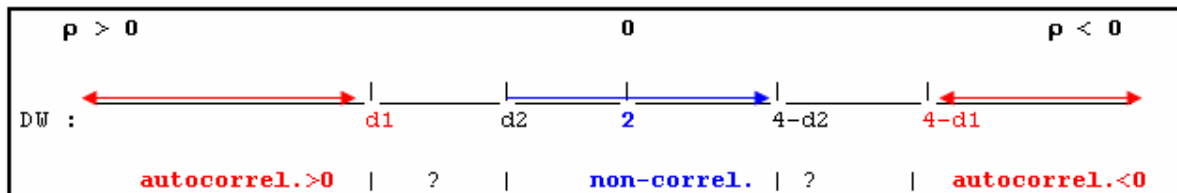
Liens entre les valeurs ρ et DW:

-Si $0 < \rho < 1 \Rightarrow$ DW compris entre 0 et 2.

-Si $0 > \rho > -1 \Rightarrow$ DW compris entre 2 et 4.

S'il n'y a pas d'auto-corrélation d'ordre 1 $\Leftrightarrow \rho$ proche de 0, donc DW proche de 2.

Il existe des tables dites de Durbin-Watson permettant de tester l'absence d'autocorrélation d'ordre 1 en fonction du niveau de confiance α , et de n (d'observations) et p (nombre de variables). On y lit deux valeurs d_1 et d_2 :



I-8-Tests de normalité :

En statistiques, les tests de normalité permettent de vérifier si des données réelles suivent une loi normale ou non. Les tests de normalité sont des cas particuliers des tests d'adéquation (ou tests d'ajustement, tests permettant de comparer des distributions), appliqués à une loi normale.

Ces tests prennent une place importante en statistiques. En effet, de nombreux tests supposent la normalité des distributions pour être applicables. En toute rigueur, il est indispensable de vérifier la normalité avant d'utiliser les tests. Cependant, de nombreux tests sont suffisamment robustes pour être utilisables même si les distributions s'écartent de la loi normale.

I-8-1-Test graphique

I-8-1-1-Q-Q-plot des résidus :

Afin de vérifier la quatrième hypothèse du modèle énoncée en section I, c'est –adire la normalité des erreurs, il est possible d'effectuer un QQ-plot des résidus. Il s'agit tout d'abord de renommer les façon a ce que e_1 soit le plus petit résidus, e_2 le second plus petit résidus, et ainsi de suite, en étant le plus grand des n résidus. On associe ensuite à un résidus e_i le

$i/(n+1)$ quantile q_i d'une loi normale centrée réduite. On représente alors sur un graphique, les résidus e_i en ordonnée et les quantiles q_i en abscisse. Si les erreurs e_i sont normalement distribuées, les points sur le graphe doivent être à peu près alignés sur la droite d'équation $e_i = q_i$. Dans la littérature francophone, ce dispositif est appelé Droite de Henry.

1-8-1-2-Approches empiriques et graphiques

1- Histogramme de la distribution :

Il est possible de visualiser la forme de la distribution des données à analyser en les représentant sous forme d'histogramme puis de comparer la forme de cet histogramme avec une courbe représentant une loi normale (les paramètres de cette loi étant calculés à partir des données à analyser).

1- Histogramme des résidus :

Il est également possible de représenter l'histogramme des résidus (c'est-à-dire la différence entre la distribution observée et la loi normale). Les résidus doivent suivre également une loi normale.

1-8-2-Test statistique [6, 7, 30 et 31]:

Un autre critère lié au choix des tests statistiques est celui de robustesse : la majorité des tests reposant sur un certain nombre d'hypothèses implicites, ou conditions d'application (normalité de la distribution des observations, etc.), la robustesse d'un test traduit sa tolérance à l'égard de la déviation par rapport à ces conditions d'application.

1-8-2-1-Test d'Anderson-Darling :

Le test de d'Anderson-Darling est une autre variante du test de Kolmogorov-Smirnov, à la différence qu'elle donne plus d'importance aux queues de distribution [68]. De ce point de

vue, elle est plus indiquée dans la phase d'évaluation des données précédant la mise en œuvre d'un test paramétrique (comparaison de moyenne, de variances, etc.) que le test de Lilliefors. Autre particularité, ses valeurs critiques sont tabulées déferrement selon la loi théorique de référence, un coefficient multiplicatif correctif dépendant de la taille d'échantillon n peut être aussi introduit.

Concernant l'adéquation à la loi normale, la statistique du test s'écrit :

$$A = -n-1/n * \sum_{i=1}^n (2i - 1) * [\ln(F_i) + \ln(1 - F_{n-l+1})] \quad (45)$$

où F_i est la fréquence théorique de la loi de répartition normale centrée et réduite associée à la valeur standardisée : $z(i) = (x_i - x_{moy})/s$.

Une correction est recommandée pour les petits effectifs [44.67.68], cette statistique corrigée est également utilisée pour calculer la p-value :

$$A_m = A * \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2} \right). \quad (46)$$

Les valeurs critiques A_{crit} pour différents niveaux de risques sont résumées dans le tableau suivant :

Tableau V : valeurs critiques AD pour différents niveaux de risques

α	A_{crit}
0,1	0,631
0,05	0,752
0,01	1,035

ils ont été produits par simulation et ne dépendent pas de l'effectif de l'échantillon :

L'hypothèse de normalité est rejetée lorsque la statistique A prend des valeurs trop élevées :

$$R.C. : A > A_{crit}$$

La différence est dans le calcul de la p-value .Minitab se contente de spécifier une plage de p-value en comparant la statistique aux seuils critiques relatifs aux différents niveaux de risque.

Dans le cas présent, il indique p-value > 0.10

Calcul de la p-value :

La p-value est calculée à partir de la statistique A_m par interpolation à partir d'une table décrite Nous donnons ici la règle de calcul :

1. calcule la statistique transformée A_m .
2. on utilise la règle suivante pour en déduire la p-value.

Tableau VI: la formule Mathématique pour les valeurs de probabilité P à partir de la statistique transformée A_m

A_m	valeur de P
$A_m < 0,2$	$1 - \exp(-13,436 + 101,14 * A_m - 223,73 * A_m^2)$
$0,2 \leq A_m < 0,34$	$1 - \exp(-8,318 + 42,796 * A_m - 59,938 * A_m^2)$
$0,34 \leq A_m < 0,6$	$\exp(0,9177 - 4,279 * A_m - 1,38 * A_m^2)$
$0,6 \leq A_m$	$\exp(1,2937 - 5,709 * A_m + 0,0186 * A_m^2)$

I-9- Diagramme de travail:

Dans notre travail, les données expérimentales sont issues de la littérature, De plus, les données doivent être obtenues suivant un protocole expérimental unique. Les étapes de travail de ce protocole suivent le schéma suivant :

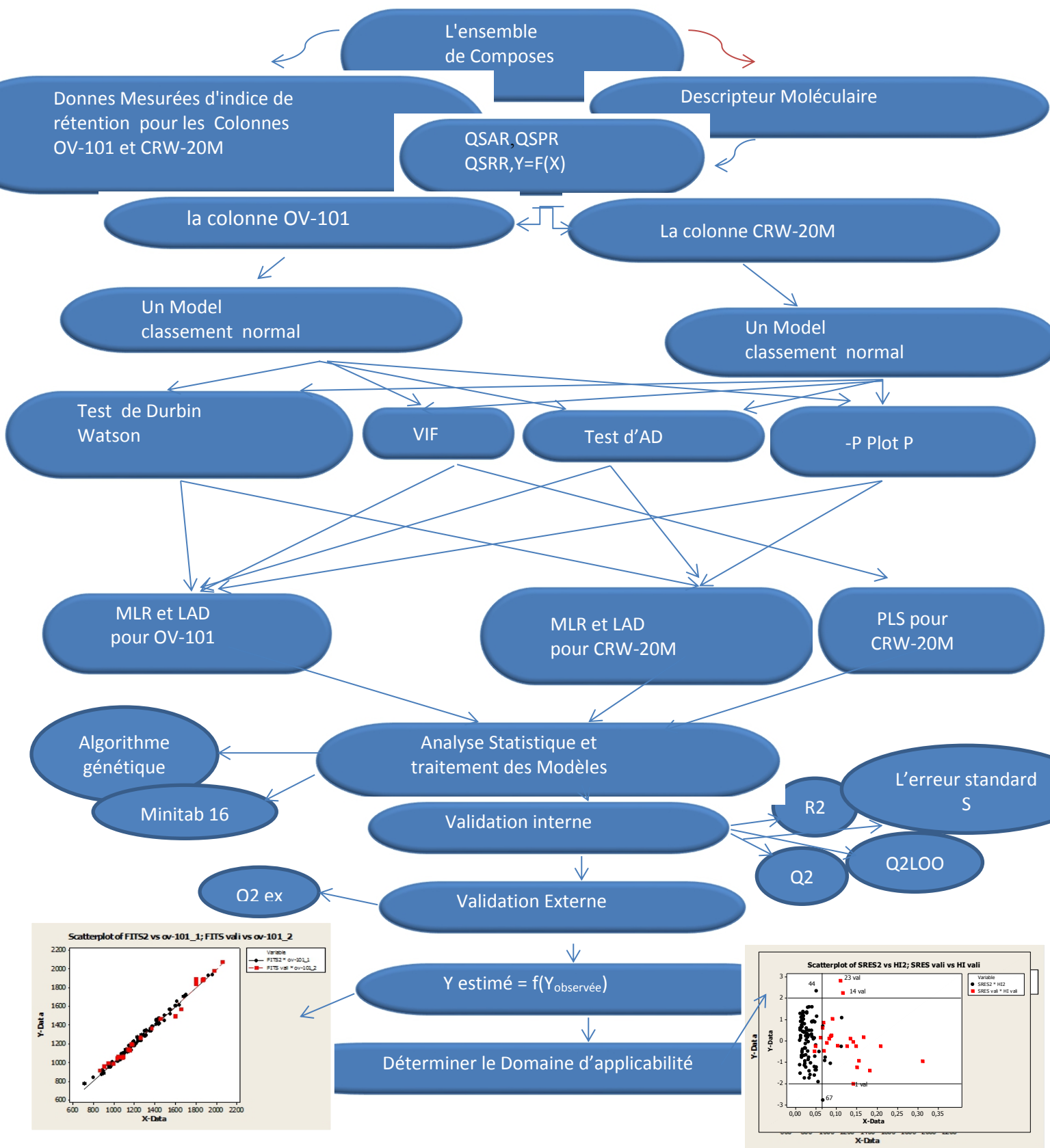


Figure 8: Diagramme de notre travail.



REFERENCES BIBLIOGRAPHIQUES

- [1] Crum-Brown, A., et Frazer, T. 1868-69. On the connection between chemical constitution and Physiological action. Transactions of the Royal Society of Edinburgh , 25, p.151-203.
- [2] Regadi Dahmane. 2012 /2013 . Développement de modèle QSPR pour la prédiction des propriétés physique des quelques composés odorants. Mémoire Master Académique. Université Kasdi Marbah Ouargla.
- [3] Tropsha A, Gramatica P, and Grombar V.K. 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR & J Combinatorial Science., Vol 22. pp 69-76.
- [4] Sharma B.K, Singh P, Pilania P, Sarbhai K, Yenamandra S, and Prabhakar. 2011. CP-MLR/PLS directed QSAR study on apical sodium-codependent bile acid transporter inhibition activity of benzothiepinines J Mol Divers Vol 15. pp 135–147.
- [5] Todeschini R., Consonni V., Mannhold R. , Kubinyi H, and Timmerman H., eds., Wiley .2000. Handbook of Molecular Descriptors, VCH, Verlag Gmbh, Weinheim
- [6] Hamad Bechira , Khahla Samia .mémoire de Magister Académique , Faculté des Sciences et de la Technologie, Université D'el-oued.
- [7] Searching scientific literature directly with Chem Draw v14 Chemistry World .29 July 2014
- [9] Martin G., Laffort P. 1991, Odeurs et désodorisations dans l'environnement, Lavoisier, Tec&Doc, Paris..
- [10] Afnor, Qualité de l'air-Mesurage de l'odeur d'un effluent gazeux. Méthodes supraliminaires, NF X 43-103.
- [11] Wiener, H. 1947. Structural determination of paraffin boiling points. Journal of Chemical Information and Computer Sciences, 69, p. 17-20.
- [12] Randić, M. 1975. On characterization of molecular branching. Journal of the American Chemical Society , 97, p. 6609-6614.
- [13] Kier, L.B., et Hall, L.H. 1976. Molecular connectivity in chemistry and drug research. New-York : Academic Press,.
- [14] Balaban, A.T. 1982. Highly discriminating distance-based topological index. Chemical Physics Letters, 89, p. 399-404.
- [15] Heritage, T.W., *et al.* EVA: 1998. A novel theoretical descriptor for QSAR studies. Perspectives in Drug Discovery and Design, 9-11 (0), p. 381-398.
- [16] Chuur, J.H., Selzer, P., et Gasteiger, J. 1996. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. Journal of Chemical Information and Computer Sciences, 36 (2), p. 334-344..

- [17] Abdelkrim Guendouz .2015.-Élaboration des modèles QSPR prédictifs des propriétés physico-chimiques à l'aide des descripteurs moléculaires. Université Abou Bekr Belkaid De Tlemcen.
- [18] Todeschini R, Ballabio D, Consonni V, Mauri A, Pavan M. 2004. Moby Digs .Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, Release for Windows, Milano.
- [19] Kowalski B, Gerlach R, Wold H. 1982. Systems under indirect observation. (K, Jöreskog H, Wold, eds.), North Holland, Amsterdam, pp, 191-206.
- [20] Erikson L, Johansson E, Kettaneh-Wold N, .2001. Multi and megavariate data analysis principles and applications. Umetrics Academy, Umeå.
- [21] Malinowsky E R, Howery D G .1980. Factor analysis in chemistry. Wiley Interscience, New York.
- [22] Meloum M, Militky M, Forina M. 1992. Chemometrics in Analytical Chemistry, Ellis Horwood, New York.
- [23] Strouf, O. 1986. Chemical Pattern Recognition, Wiley, New York.
- [24] Lebart L, Morineau A, Piron M. 2004. Statistique exploratoire multidimensionnelle", 3ème ed, Dunod, Paris.
- [25] Jolliffe, I T. 1986. Principal Component Analysis", Springer- Verlag, Berlin.
- [26] Escofier B, Pages J. 1998. Analyses factorielles simples et multiples: Objectifs, méthodes et interprétation", 3ème ed, Dunod, Paris.
- [27] Gauchi J P. 1995. Utilisation de la régression PLS pour l'analyse des plans d'expériences en chimie de formulation", Rev, Stat, Appl, Vol, 43, pp, 65-89.
- [28] Gelada P, Kowalski B, R. 1986. Partial least-squares regression: tutorial. Anal, Chim, Acta, Vol, 185, pp, 1- 17.
- [29] Vancolen, S. 2004. la régression PLS, groupe de statistique, université de Neuchâtel, Suisse.
- [30] Tenenhaus M. 1998. la régression PLS, théorie et pratique Paris : Technip.
- [31] Pagès J, Tenenhaus M. 2001. Multiple factor analysis combined with PLS path modeling, applications to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgments", Chemometrics and Intelligent Laboratory Systems, Vol, 58, pp, 261- 273.
- [32] Mesquita D P O, Dias A M A, Dias A L, Amaral E C, Ferreira .2009. Correlation between sludge settling ability and image analysis information using partial least squares. Anal, Chim, Acta, Vol, 642, pp, 94-10
- [33] Errahoui née bellifa khadidja, thèse de doctorat département de chimie faculté des sciences, Université Kasdi Marbah Ouargla.

- [34] Wold H .1966.Estimation of principal component and related models by iterative least squares, multivariate analysis, ed, P, R, Krishnaiah, New York: Academic Press, pp, 391-420.
- [35] MINITAB, Release 16.1, Statistical Software, 2003
- [36] Leo Ghemtio.2010. Simulation numerique et approche orientee connaissance pour la decouverte de nouvelles molecules therapeutiques. Universit_e Henri Poincar_e - Nancy I, 2010.the de doctorat . Thèse dirigée par Dr Maignret Bernard.
- [37] Hohenberg , P., Kohn,W. (1964)- Inhomogeneous Electron Gas, j,Physical Review; vol. 136 No. 3B, pp.B 864 –B871.
- [38] Randić, M. 1975.On characterization of molecular branching. Journal of the American Chemical Society, 97, p. 6609-6614.
- [39] Ramachandran K.I , Deepa, G , Namboori, K.2008. Computational Chemistry and Molecular Modeling. Principles and Applications. DOI 10. 1007/978-3-540-77304-7.
- [40] Dodge Y., Rousson. .2004. Analyse de regression appliqué Dunold, Paris.
- [1]Montgomery,D.C , Peck, E.A. 1992.Introduction to Linear Regression Analysis. Snd Edition. John Wiley & sons. Inc.
- [41] Kubinyi H.1994. Quantitative Structure-activity Relationships., Vol 13 , pp 285.
- [42] Draper N.R, and Smith H., (1998)- Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New York.
- [43] Sharma B.K, Singh P, Pilia P, Sarbhai K, Yenamandra S, and Prabhakar. 2011.CP-MLR/PLS directed QSAR study on apical sodium-codependent bile acid transporter inhibition activity of benzothiepinines J Mol Divers Vol 15. pp 135–147.
- [44] Nornadiah Mohd Razali ,Yab Bee Yah .2011. Power Comp²araisons of shapiro-wilk, Kolmogorov-smornov,lillieffors and Anderson-Darling tests,journal of statistique Modelling and analytics .vol 2 No 1:21-33
- [45] Damodar N. Gujarati, Dawn C. Porter.2009.Basic Econometrics Fifth Edition
1. [46] Faria, S. and Melfi, G. 2006. Lad regression and nonparametric methods for detecting outliers and leverage points. Student, 5 :265– 272.
- [47] Gabriela Ciuperca. 2009.Estimation robuste dans un modèle paramétrique avec rupture. Bordeaux.
- [48] Gilbert Saporta. 2012. Régression robuste.
- [49] Ndèye Niang- Gilbert Saporta.2014.Régression robuste Régression non-paramétrique.
- [50] Matlab Ra 2009a

- [51] Pynnönen, Seppo and Timo Salmi 1994. A Report on Least Absolute Deviation Regression with Ordinary Linear Programming. Finnish Journal of Business Economics 43:1, 33-49.
- [52] Tiffany Machabert .2014 "Modèles en très grande dimension avec des outliers. Théorie, simulations, applications" paris
- [53] Dodge, Y. (2004). Statistique : Dictionnaire encyclopédique. Springer-Verlag France Paris.
- [54] Dodge, Y. and Jureckova, J. 2000. Adaptive Regression. Springer-Verlag New York.
- [55] Dodge, Y. et Valentin Rousson .2004. Analyses de régression appliquée.paris.
- [56] Yadolah Dodge.1997. LAD Regression for Detecting Outliers in Response and Explanatory Variables. journal of multivariate analysis 61, 144_158 .
- [57] -J. Paul Tsasa Vangu.2011.Econometrie.German
- [58] Oya Can Mutan , Birdal Şenoğlu.2009.A Monte Carlo Comparison of Regression Estimators When the Error Distribution is Long-Tailed Symmetric. Journal of Modern Applied Statistical Methods.Volume 8 | Issue 1.
- [59] Umaporn Chantasorn.2011. Efficiency Comparisons of Normality Test Using Statistical Packages. J. Sc. Tech., Vol. 16, No. 3,
- [60] Tron Foss, Ingunn Myrvtveit, Erik Stensrud. Yinbo Li, Gonzalo R. Arce.2004. AMaximum Likelihood Approach to Least Absolute Deviation Regression. Journal on Applied Signal Processing, 12, 1762–1769.
- [61] Amit Mitra , DebasisKundu. 2009.Genetic algorithms based robust frequency estimation of sinusoidal signals with stationary errors. Engineering Applications of Artificial Intelligence. Elsevier.
- [62] Chenlei Leng.2010. Variable Selection and Coefficient Estimation Via Regularized Rank Regression . Statistica Sinica.
- [63] Firas H. Thanoon.2015. Robust Regression by Least Absolute Deviations Method. International Journal of Statistics and Applications , 5(3): 109-112.
- [64] Soner Çankal,Samet hasan abaci.2015.AComparative Study of Some Estimation Methode in SimpleLinear Regression Model for Different Sample in Presence of outliers.Turkish Journal of Agriculture-Food Science and Technology .3(6);380-386.
- [65] Richard J. Butler,James B.McDonald,Ray.Nelson,Steven B.White .1990.Robust and Partially Adaptive Estimation of Regression Models.The Review of Economics and Statistics.Volume 72 ,Issue 2,321-327.
- [66] Philippe GROS. 1997. documents océanographiques. Paris Océanis 23(3) : 359 - 515.

- [67] Gilbert Colletaz. 2004 .Statistique non paramétrique élémentaire .Université Dorleans.
- [68] Ricco Rakotomalala. 2011. Tests de normalité Techniques empiriques et tests statistiques.Université Lumière Lyon 2.
- [69] claudio araujo.2013 .Micro économétrie stratégie d'estimation méthode de base.
- [70] Ricco Rakotomalala.2011. Pratique de la Régression Linéaire Multiple. -Université Lumière Lyon 2.
- [71] Daiel Borcard.2009. Régression multiple. Université de Montréal
- [72] Hansch, C., Leo, A., et Hoekmann, D. 1995.Exploring QSAR : hydrophobic, electronic and steric constants. Washington, DC : American Chemical Society.
- [73] Goulon -Sigwalt-ABRAM.2008.Thèse De Doctorat De L'université Paris 6 Pierre et Marie Curie. Une nouvelle méthode d'apprentissage de données structurées : applications à l'aide à la découverte de médicaments.
- [74] Dewar, M. J. S. and Thiel ,W. 1977. Ground States of Molecules.38. The MNDO Method. Approximations and Parameters, J of the American Chemical Society , vol .99 No 15, pp.4899-4907.
- [75] Christophe Morell .2006.Un nouveau descripteur de la réactivité chimique : Etude théorique et applications de quelques réactions chimiques. Université Joseph-Fourier- Grenoble I. Français.
- [76] Kubinyi H .1994. Quantitative Structure-activity Relationships., Vol 13 , pp 285.
- [77] Draper N.R, and Smith H.1998.Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New York.
- [78] Eriksson L., Jaworska J, and Worth A.P.2003. Perspective M.T.D., Vol 111(10), pp 1361-1375.
- [79] Erikson L., Jaworska J., Worth A. P., Cronin M. T. D., Mc Dowell R. M, and Gramatica P..2003. Methods for Reliability and uncertainty Assessment and for Applicability Evaluations of Classification-and Regression - Based QSARs. J Environmental Health Perspectives. Vol 111 (10). pp 1361-1375.
- [80] Todeschini R, ConsonniV, PavanM .2005.DRAGON, Software for the Calculation of Molecular Descriptors, Release 5.3 for windows, Milano.

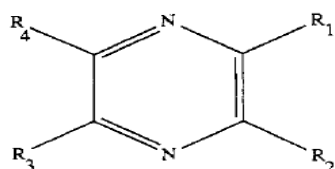
PARTIE III

RESULTATS ET DISCUSSION

Les pyrazines sont des hétérocycles très présents dans nos aliments. Plus de 80 dérivés de pyrazines ont été identifiés dans un grand nombre d'aliments cuits, comme le pain, la viande, le café torréfié, le cacao ou les noisettes ; ce sont des composés aromatisants très puissants. Leur identification se fait généralement par chromatographie gazeuse (CG) en comparant leurs pics à ceux obtenus pour les standards des composés suspectés.

Nous montrerons tout d'abord comment les molécules peuvent être représentées par des vecteurs de réels, et comment ces descripteurs sont sélectionnés. Nous introduirons ensuite les Outils de modélisation sans contraintes, les plus utilisés, c'est-à-dire la régression linéaire multiple, qui est fondée sur le calcul de descripteurs. Nous présenterons le problème de la sélection de modèle, ainsi que les stratégies les plus efficaces pour le résoudre. Nous décrirons alors la théorie statistique de l'apprentissage de Vapnik et les méthodes de modélisation sous contraintes, ainsi que leurs applications en QSAR et QSPR.

Les composés impliqués dans cette étude, présentent la structure générale suivante [1] :



R₁: H, alkyl, alkoxy, alkylthio, aryloxy, arylthio, acetyl, chloro.

R₂: H, alkyl, chloro.

R₃: H, alkyl.

R₄: H, alkyl.

Figure 9 : Structure général de pyrazine

Les Relations Quantitatives Structure/Rétention (QSRR) initiées par Hansch et Fujita (1964) [2], ont trouvés de nombreuses applications en chimie, en particulier dans la prédiction de la rétention chromatographique.

I. Résultats et discussion :

Le but est de trouver un modèle statistique pour la prédiction de l'indice de rétention de composés hétérocycles aromatique de pyrazine pour deux colonnes ; une colonne non polaire (OV-101) et une colonne polaire (CRW-20M). A cet effet, la relation entre les descripteurs moléculaires [3] reliée aux valeurs expérimentalement constatées (indices de rétention des composés), a été établie sur les deux colonnes. Les modèles QSRR ont été construit en utilisant les méthodes de la régression linéaire multiple (MLR), (LAD) et (PLS) et leurs performances validées. Les modèles obtenus montrent quels descripteurs jouent un rôle important dans la variation des IR de ces Pyrazines.

Les meilleurs modèles trouvés pour chaque phase stationnaire en employant le logiciel de Moby Digs [4].sont donnés ci-dessous.

I-1- Cas de la colonne non polaire (OV-101) :**I-1-1-La régression avec les moindres carrés :**

La sélection par algorithme génétique conduit à un bon modèle, la méthode MLR à trois descripteurs décrit au mieux l'indice de rétention. Le modèle retenu a pour équation

$$Y = -809.4 + 29.2454 * XMOD + 1028.3 * FDI + 70.453 * Mor06v \quad (47)$$

S = 18,42 R-Sq = 99,43% R-Sq(adj) = 99,41% ,Q²=99,37%,Q²boo=99,34%,Q²ext=95,38%,
F=4965,34.

Le tableau VII résume les trois descripteurs sélectionnés utilisé pour la modélisation.

Tableau VII : Les descripteurs sélectionnés pour la modélisation d'IR.

Descripteur	Définition	Classe
XMOD	indice de connectivité Randic modifié	indice de connectivité
FDI	indice de degré se pliant	descripteur géométrique
Mor06v	3D-MoRSE - signal 06/pondérée par les volumes atomiques de van der waals.	3D-MoRSE (Morsw)

Les trois descripteurs ont été obtenus en utilisant le logiciel Dragon [6]. On trouvera plus d'informations concernant ces descripteurs dans le guide de l'utilisateur du logiciel Dragon (Todeschini R *et al.*, 2005) [62] et les références afférentes.

Les valeurs de $R^2=99.43\%$ et $R^2_{adj}=99.41\%$ montrent la qualité de l'ajustement, alors que la petite différence entre R^2 et Q^2 ($=0.06$) renseigne sur la robustesse du modèle qui est en outre hautement significatif (grande valeur de la statistique F de Fisher= $4965,34$, $P=0.00$, $S=18.42$, $n=89$). La validation par $Q^2_{EX}=95.38\%$ confirme tout à la fois la bonne capacité de prédiction interne et la stabilité du modèle.

La matrice de corrélation, selon le tableau(VIII) suggère que :

Les deux descripteurs importants (XMOD) ($R=0.986$) et Mor06v($R=0.181$) présentent le plus grand impact sur l'indice de rétention (relation positive entre ce descripteur et l'indice de rétention), corrélation forte sur la phase OV-101.

Le premier descripteur important est FDI qui a une corrélation négative avec les valeurs de IR ($R=-0.039$), il est le plus petit impact sur l'indice de rétention (relation négative entre ce descripteur et l'indice de rétention), corrélation moindre sur la phase OV-101

Tableau VIII : Matrice de corrélation

	ov-101	XMOD	FDI
XMOD	0,986 0,000		
FDI	-0,039 0,715	-0,152 0,154	
Mor06v	0,181 0,089	0,059 0,582	0,274 0,009

Le tableau IX résume les résultats des caractéristiques (les évaluations) des descripteurs sélectionnés par L'estimation MLR.

Tableau IX: Évaluations de l'estimation MLR pour le model

Predictor	Coef	SE Coef	T	P	VIF
Constant	-809,4	107,2	-7,55	0,000	
XMOD	29,2454	0,2535	115,35	0,000	1,035
FDI	1028,3	108,5	9,48	0,000	1,116
Mor06v	70,453	6,266	11,24	0,000	1,094

- Les valeurs des probabilités de t pour les trois descripteurs sont nuls, ceci indique qu'ils sont hautement significatifs avec un risque d'erreur de première espèce $\alpha=0.05$, les paramètres sont tous significatifs parce que Leurs coefficients d'estimations sont de l'ordre : $\beta_0 = -809.4, \beta_1 = 29.245, \beta_2 = 1028.3$ et $\beta_3 = 70.453$.

La méthode MLR est la meilleure méthode de régression linéaire ; et la plus utilisée dans les études QSPR, Sous certaines conditions:

1-la normalité des résidus :

En utilisant un niveau de $\alpha=0.05$, l'essai de normalité avec le test de normalité d'Anderson-Darling) (figure 10) (A-Squared =0,134 ; $OV - 101, < v_{cri} = 0.752,$) indique que les données d'observation des 'erreurs suivent une distribution normale mais il' Ya quelques perturbations dues à la présence de points aberrants.

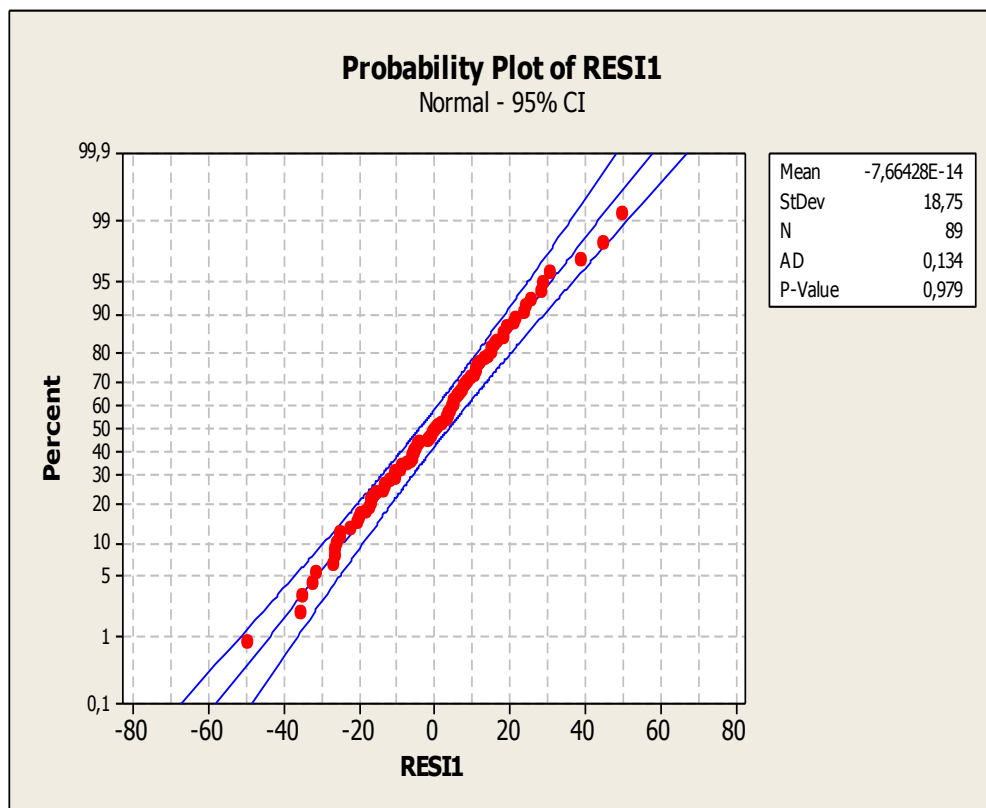


Figure 10 : Diagramme des scores normaux

2--auto correlation des residues:

Les valeurs des statistiques de Durbin-Watson (Durbin, et de Watson, 1951), [$d=1,47910$] sont inférieures aux valeurs données par les tables respectivement pour 3 régresseurs, et pour le risque raisonnable $\alpha = 0.05$ (valeur critique entre : ($d_L=1,55$, $d_U=1.72$)) ce qui exprime l'auto corrélation positive des résidus (présence de l'autocorrélation) (Figure 11) ceci indique que la présence des points aberrants pose un problème ; les coefficients de régression ne sont plus efficaces ; dans ce cas on compare la méthode MLR avec différentes méthodes d'estimation .

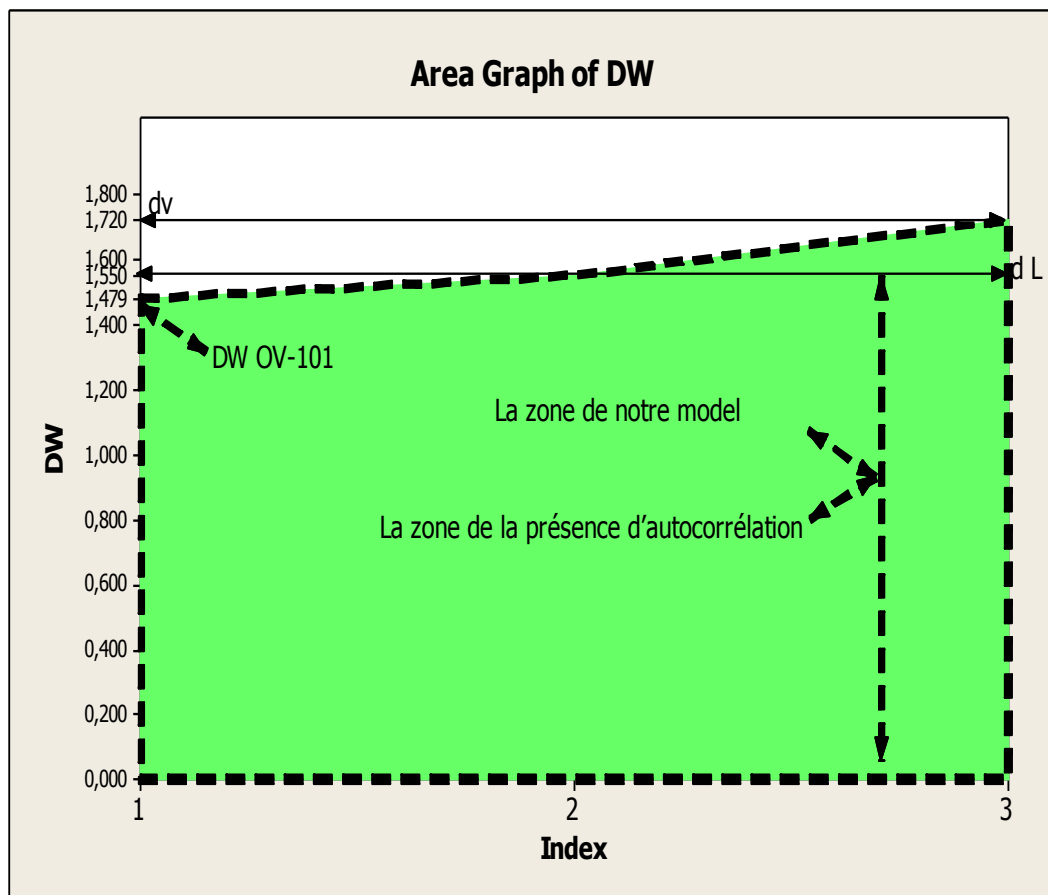


Figure 11: Diagramme de valeur de Durbin Watson.

3-Test de multicollinéarité :

Le facteur d'inflation de la variance (FIV) permet de décrire l'importance de la multicollinéarité (la corrélation entre les prédicteurs) dans une analyse de régression (Tableau IX) et figure(12), Les valeurs des facteurs d'inflation de la variance (FIV) pour les trois descripteurs sont égaux à 1, ceci indique qu'ils ne sont pas corrélés, donc le problème de multicollinéarité n'existe pas.

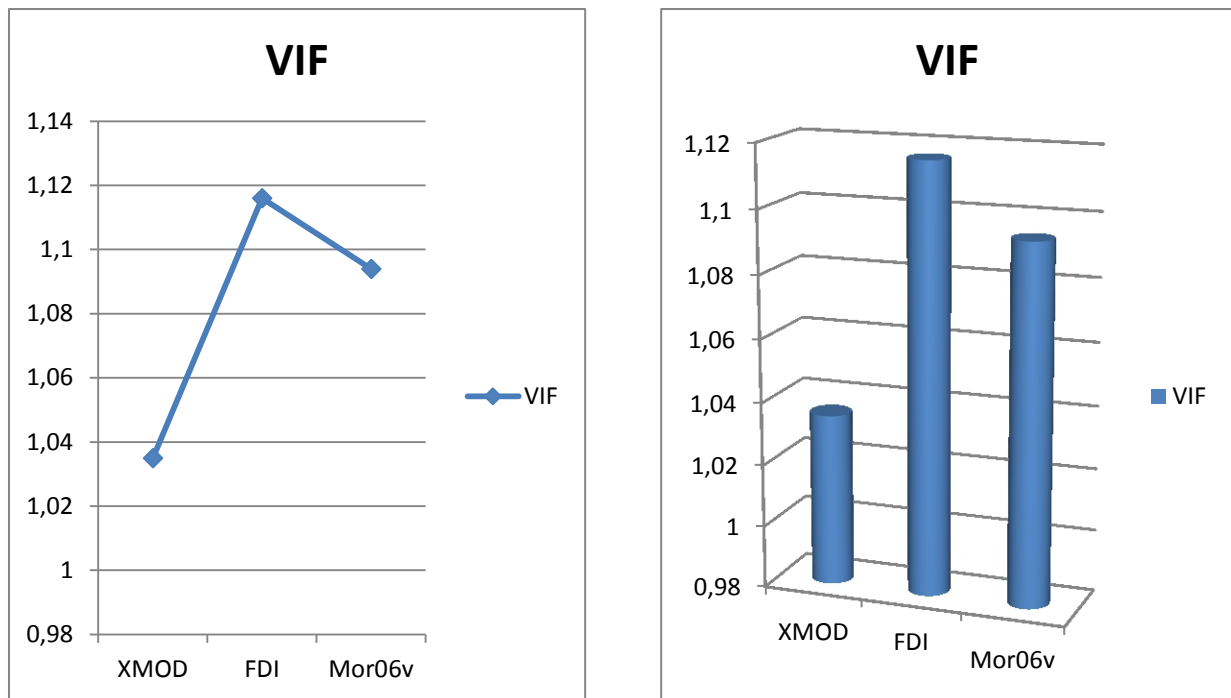


Figure 12: Diagramme de VIF pour chaque descripteur.

I-1-2-Modèle de régression par la méthode LAD :

I-1-2-1-Programmes de calcul d'équation d'hyperplan par l'estimation LAD.

Les programmes de calculs de l'équation d'hyperplan de régression multiple LAD par le logiciel Matlab (R2009a) [7] compatible sur PC, écrits en langage Turbo Pascal (voir l'annexe 2). Conduit à l'équation de l'hyperplan par l'estimation LAD :

$$y = -946 + 29,1 \text{ XMOD} + 1174,4 \text{ FDI} + 70,4 \text{ Mor06v} \quad (48)$$

Le tableau X résume les résultats des caractéristiques des descripteurs sélectionnés par l'estimation LAD.

Tableau X : Évaluations de l'estimation LAD pour le modèle.

Predictor	Coef	SE Coef	T	P
Constant	-946	100,237	-9,44	0,000
XMOD	29,1	5,216	5,58	0,000
FDI	1174,4	65,36	17,97	0,000
Mor06v	70,4	10,909	6,453	0,000

- Les valeurs des probabilités de t pour les trois descripteurs sont nuls, ceci indique qu'ils sont hautement significatifs avec un risque d'erreur de première espèce $\alpha=0.05$, les paramètres sont tous significatifs parce que Leurs coefficients d'estimations sont de l'ordre :

$$\beta_0 = -946, \beta_1 = 29.1, \beta_2 = 1174.4 \text{ et } \beta_3 = 70.4.$$

I-1-3- Comparaison de la Régression Robuste par la méthode MLR et la méthode LAD :

I-1-3-1-Comparaison des hyperplans par la méthode MLR et la méthode LAD :

2-La méthode MLR : A partir de l'équation (47)

$$Y = - 809 + 29, 2 XMOD + 1028 FDI + 70, 5 Mor06v$$

1-La méthode LAD: A partir de l'équation (48) :

$$y = - 946 + 29,1 XMOD + 1174.4 FDI + 70,4 Mor06v$$

On remarque que les coefficients β_0, β_2 calculés par la méthode MLR sont peu différents par rapport aux coefficients β_0, β_2 calculés par la méthode LAD, et les coefficients β_1 et β_3 calculés par la méthode MLR sont presque les mêmes que les coefficients β_1 et β_3 calculés par la méthode LAD (l'équation 47 et 48) et le Figure(13).

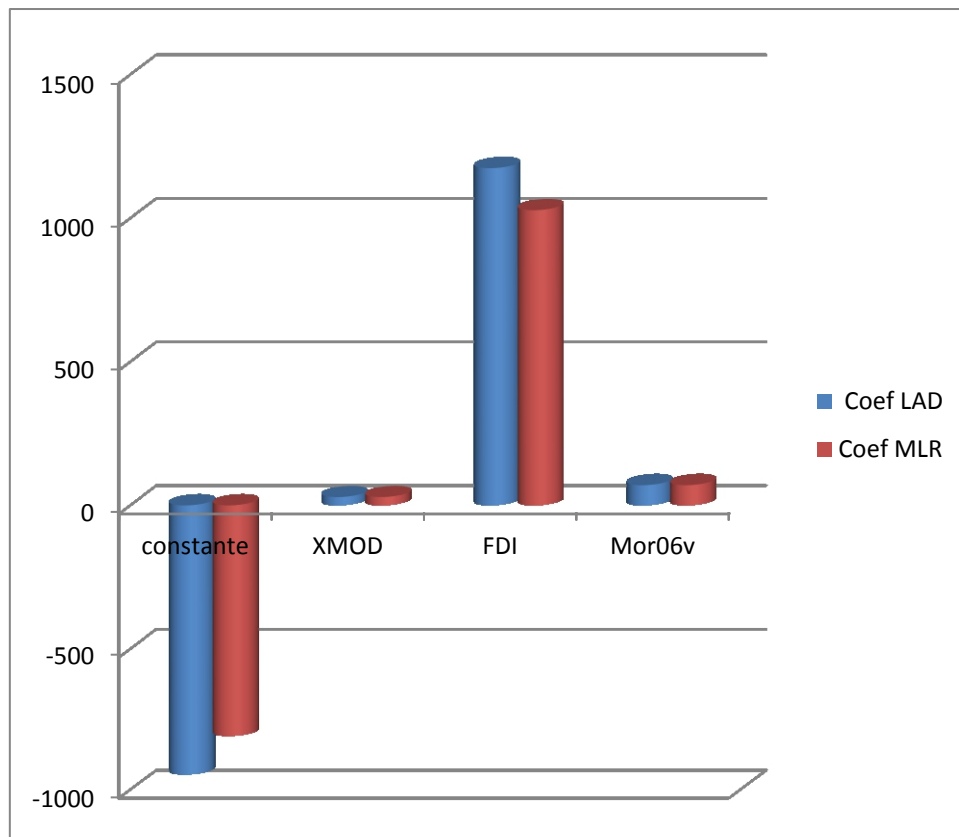


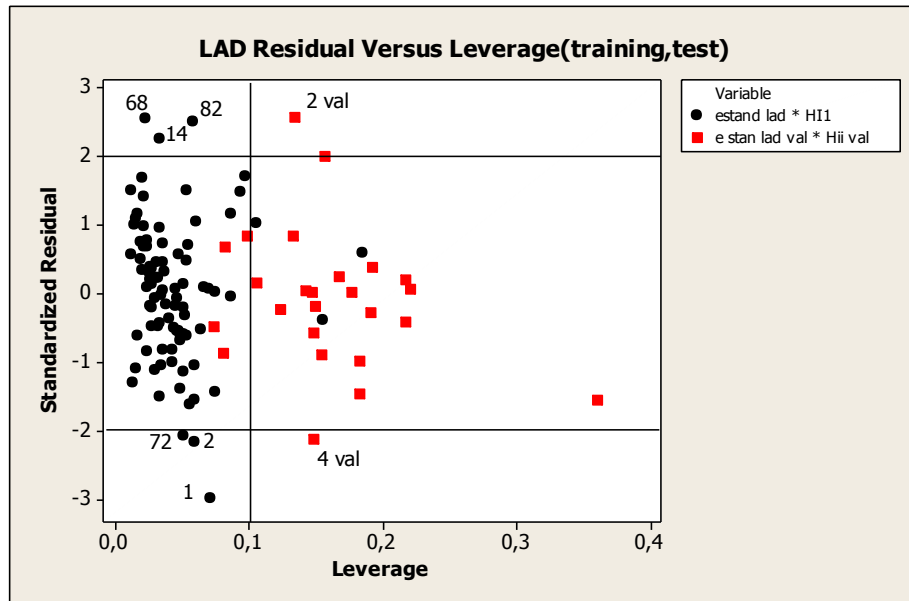
Figure 13: Diagramme de comparaison des coefficients de régression entre les deux Méthodes.

Il est donc pertinent de refaire une vérification de la présence de valeurs aberrantes, Puisque L'hyperplan de régression peut radicalement changer, avec le changement des coefficients de l'hyperplan.

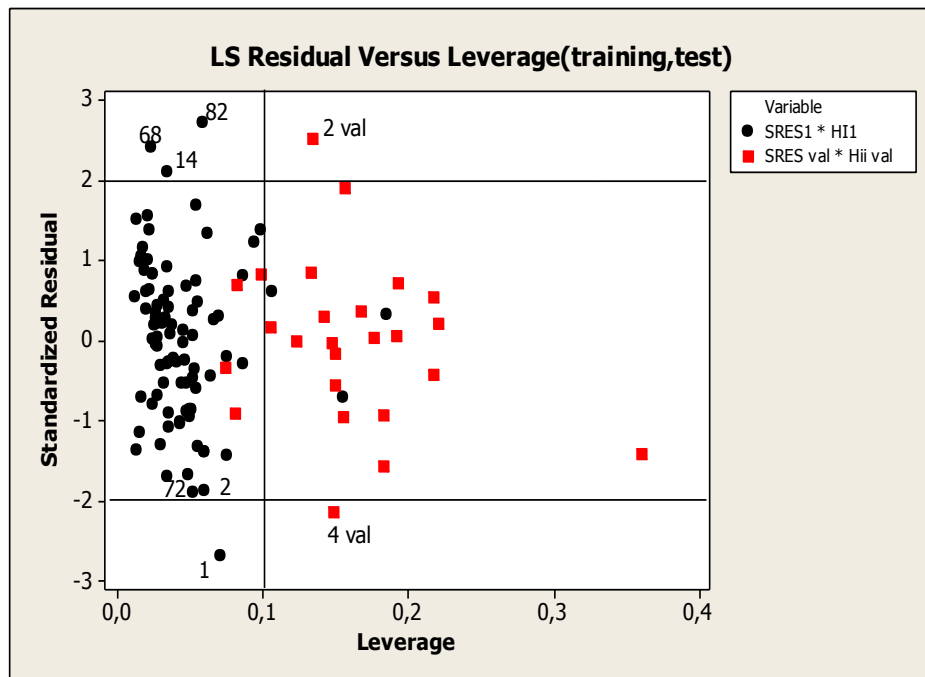
On trace le diagramme de Williams pour la méthode LAD et la méthode MLR pour pouvoir les comparer :

I-1-3-2-Comparaisons graphiques des modèles alternatifs de régression :

1-domaine d'application



La Méthode LAD (Calibration, validation).



La Méthode MLR (Calibration, validation)

Figure 14: Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes.

Dans ce diagramme (Figure 14) L'analyse des résidus pour les deux ensembles donne avec :

- l'estimation LAD :

- dans le cas de la Calibration : six points peuvent être considérés comme valeurs aberrantes :

1-Pyrazine

2 -Methylpyrazine

14 -2,3-diethylpyrazine

68 -(methylthio)pyrazine

72 -3-isopropyl-2-(methylthio)pyrazine

82 -2-ethylthio-5-isobutyl-3-methylpyrazine

Car ils se situent en dessous de la ligne de référence horizontale.

-dans le cas de la validation : deux points peuvent être considérés comme valeurs aberrantes :

2 -(phenylthio)pyrazine

4-5-isopropyl-3-methyl-2-(phenylthio)pyrazine

car ils se situent en dessous de la ligne de référence horizontale et situés à droite de la ligne verticale ($h_{ii} > 0.1$) ; ceci influe sur l'ajustement du modèle.

- l'estimation MLR :

- dans le cas de la Calibration : quatre points peuvent être considérés comme valeurs aberrantes:

1 -Pyrazine

14 -2,3-diethylpyrazine

68 -(methylthio)pyrazine

82 -2-ethylthio-5-isobutyl-3-methylpyrazine

Car ils se situent en dessous de la ligne de référence horizontale.

-dans le cas de la validation : deux points peuvent être considérés comme valeurs aberrantes :

2 -(phenylthio)pyrazine

4-5-isopropyl-3-methyl-2-(phenylthio)pyrazine

Car ils se situent en dessous de la ligne de référence horizontale et situés à droite de la ligne verticale ($h_{ii} > 0.1$) ; ceci influe sur l'ajustement du modèle.

2-La qualité de l'ajustement :

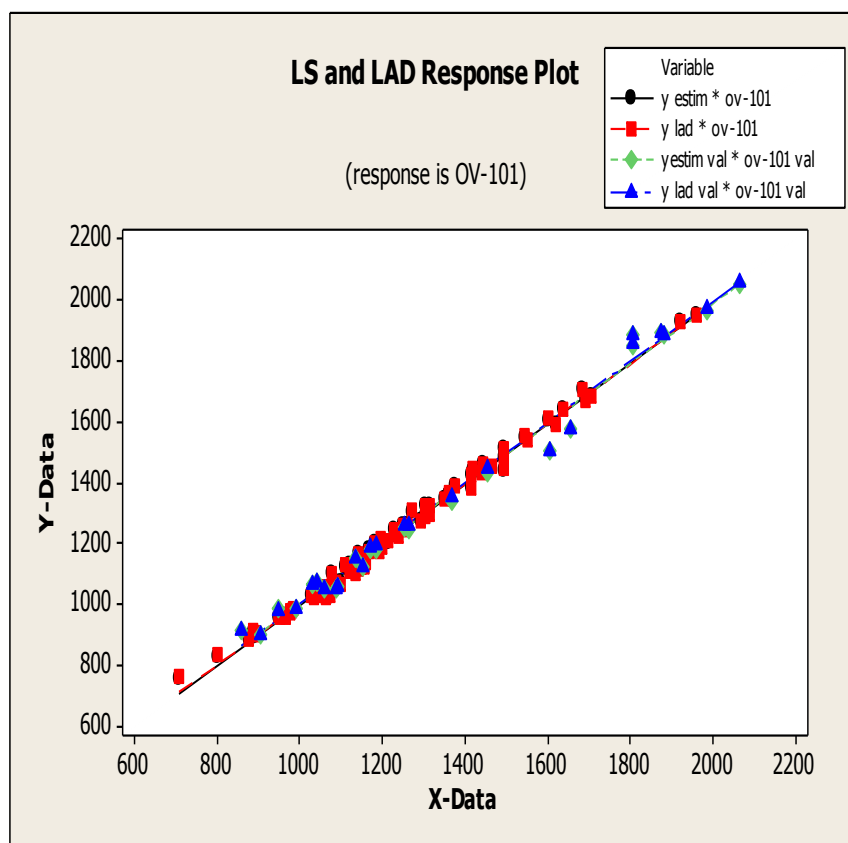


Figure 15 : Valeurs estimées en fonctions de valeurs observées pour les deux méthodes.

On remarque une relation linéaire entre les valeurs estimées et les valeurs observées pour les deux méthodes la Figure (15) montre que le modèle est correctement ajusté aux données pour les deux ensembles ; Ce qui prouve l'approche et la proportionnalité entre les deux méthodes.

Il est donc pertinent de refaire une seconde étude, après l'élimination des points aberrants communs entre les deux méthodes sur la colonne OV-101:

- dans le cas de la Calibration : on supprime quatre points communs :

1-Pyrazine

14-2,3-diethylpyrazine

68-(methylthio)pyrazine

82-2-ethylthio-5-isobutyl-3-methylpyrazine.

- dans le cas de la Validation : on supprime deux points communs:

2 -(phenylthio) pyrazine)

4 -5-isopropyl-3-methyl-2-(phenylthio) pyrazine.

I-1-3-3-Comparaison des hyperplans par la méthode MLR et la méthode LAD:

1- La méthode LAD:

$$y = - 946 + 29,15 \text{ XMOD} + 1174.4 \text{ FDI} + 70,43 \text{ Mor06v} \quad (49)$$

2- La méthode MLR :

$$Y = - 821 + 29,1 \text{ XMOD} + 1044 \text{ FDI} + 72,0 \text{ Mor06v} \quad (50).$$

On remarque que les coefficients β calculés par la méthode MLR sont peut différents que les coefficients β calculés par la méthode LAD ; sauf le coefficient β_1 est identique pour la méthode MLR et la méthode LAD (Figure 16 et L'équation (49 et 50)).

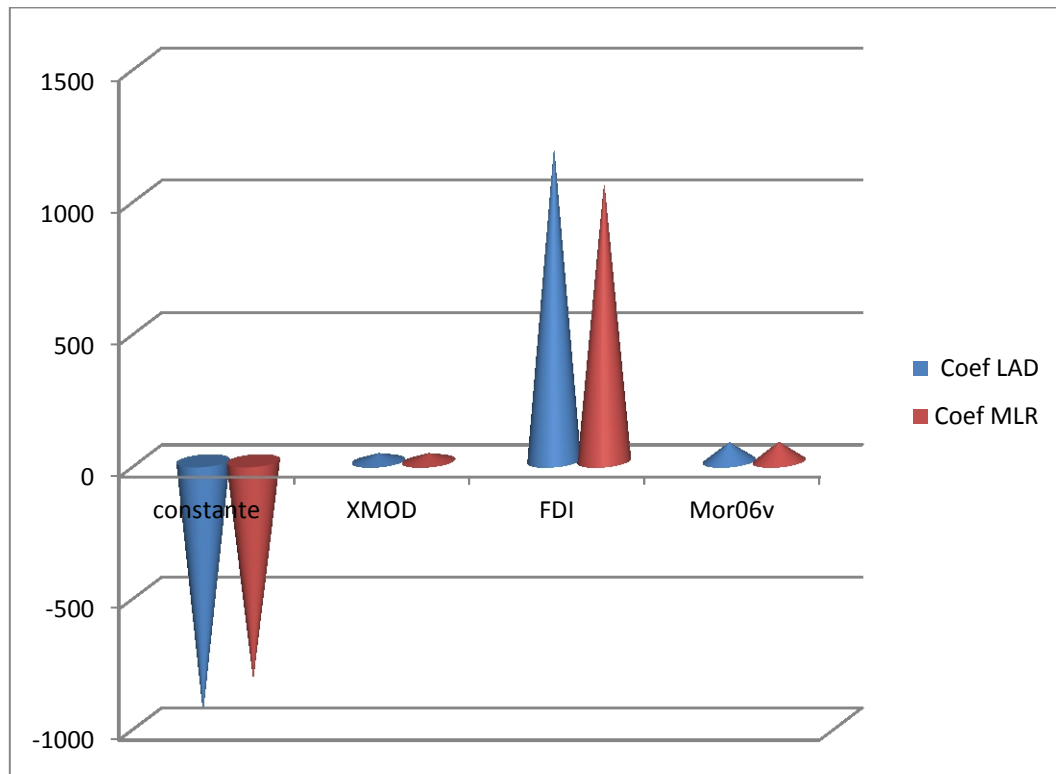


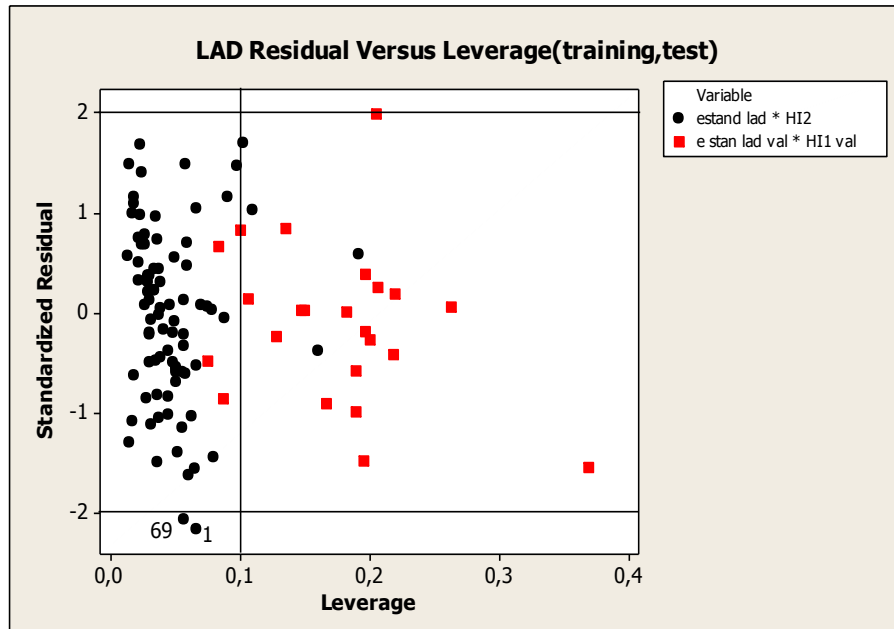
Figure 16 : Diagramme de comparaison des coefficients de régression entre les deux Méthodes.

Il est donc pertinent de refaire une vérification sur la présence des valeurs aberrantes.

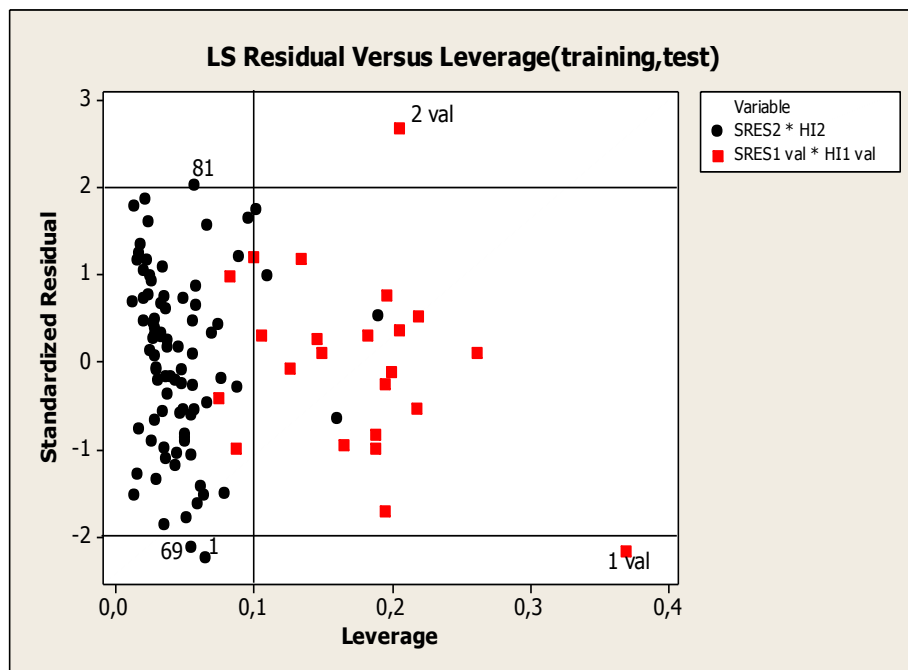
On trace le diagramme de Williams pour la méthode LAD et la méthode MLR pour pouvoir les comparer :

I-1-3-4--Comparaisons graphiques des modèles alternatifs de régression :

1-domaine d'application :



La Méthode LAD (Calibration, validation).



La Méthode MLR (Calibration, validation).

Figure 17: Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes.

Dans ce diagramme (Figure 17) L'analyse des résidus pour les deux ensembles donne avec

- l'estimation LAD :

- dans le cas de la Calibration : deux points peuvent être considérés comme valeurs aberrantes:

1 -Methylpyrazine

69 - 3-isopropyl-2-(methylthio) pyrazine

Car ils se situent en dessous de la ligne de référence horizontale

- dans le cas de la validation : toutes les observations entre l'intervalle (-2,2) ne présentent pas de valeurs aberrantes.

- l'estimation MLR :

- dans le cas de la Calibration : trois points peuvent être considérés comme valeurs aberrantes:

81 - Phenoxy pyrazine

69 - 3-isopropyl-2-(methylthio)pyrazin

1 -Methylpyrazine

Car ils se situent en dessous de la ligne de référence horizontale.

- dans le cas de la validation : deux points peuvent être considérés comme valeurs aberrantes:

1 - 3-methyl-5-(2-methylpentyl)-2-phenoxy pyrazine

2 - 3-methyl 2(phenylthio)pyrazine

Car ils se situent en dessous de la ligne de référence horizontale et situés à droite de la ligne verticale ($h_{ii} > 0.1$) ; ceci influe sur l'ajustement du modèle.

On remarque que les deux méthodes gardent les valeurs aberrantes, il est donc pertinent de refaire une seconde étude après l'élimination des points aberrants communs entre les deux méthodes et qui ne sont pas séparés sur la colonne OV-101:

- dans le cas de la Calibration : on supprime deux points aberrants communs :

1 - Méthylpyrazine)

69 - 3-isopropyl-2-(méthylthio) pyrazine)).

- dans le cas de la Validation : (il n'y a pas des points communs)).

I-1-3-5--Comparaison des hyperplans par la méthode MLR et la méthode LAD:

1- La méthode LAD:

$$y = -946 + 29,1 \text{ XMOD} + 1174,4 \text{ FDI} + 70,4 \text{ Mor06v} \quad (51)$$

2-La méthode MLR :

$$\text{ov-101} = -886 + 29,1 \text{ XMOD} + 1115 \text{ FDI} + 70,9 \text{ Mor06v} \quad (52)$$

On remarque que les coefficients β calculés, par la méthode MLR sont plus proches des coefficients β calculés, par la méthode LAD (Figure 18, l'équation 51 et 52).

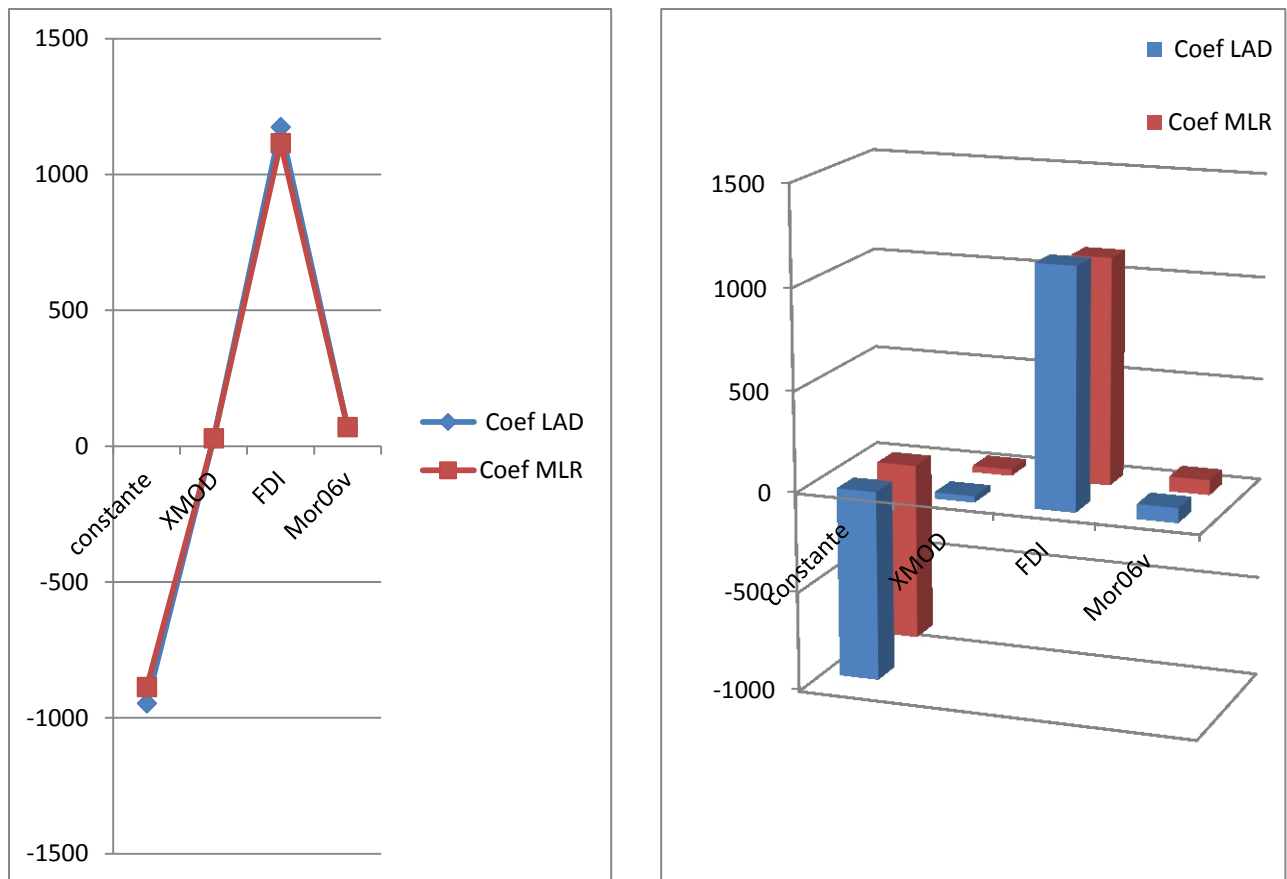


Figure 18 : Diagramme de comparaison des coefficients de régression entre les deux Méthodes.

Il est donc pertinent de refaire un test graphique pour confirmer l'état proche et la vérification de la présence des valeurs aberrantes.

On trace le diagramme de Williams pour la méthode LAD et la méthode MLR pour pouvoir les comparer :

I-1-3-6--Comparaisons graphiques des modèles alternatifs de régression :

1-Domaine d'application :

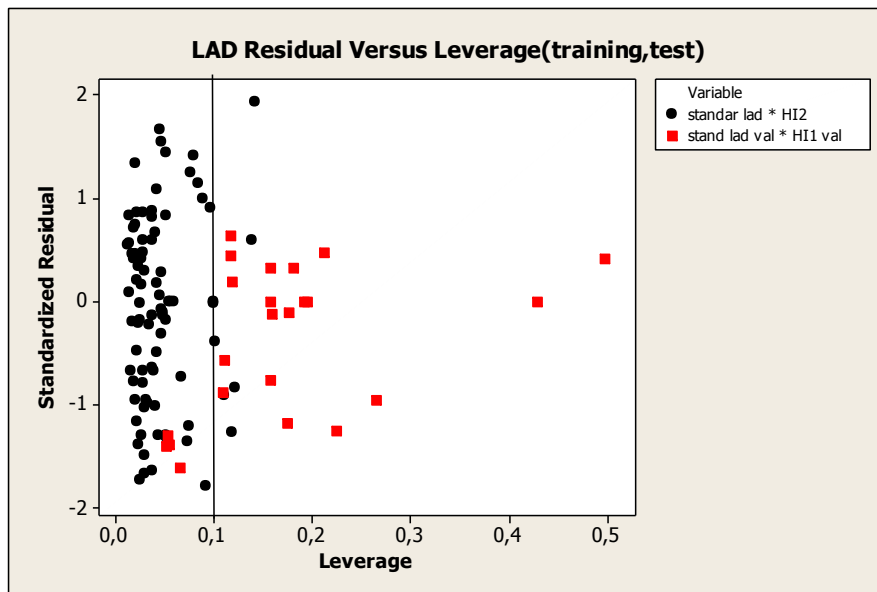
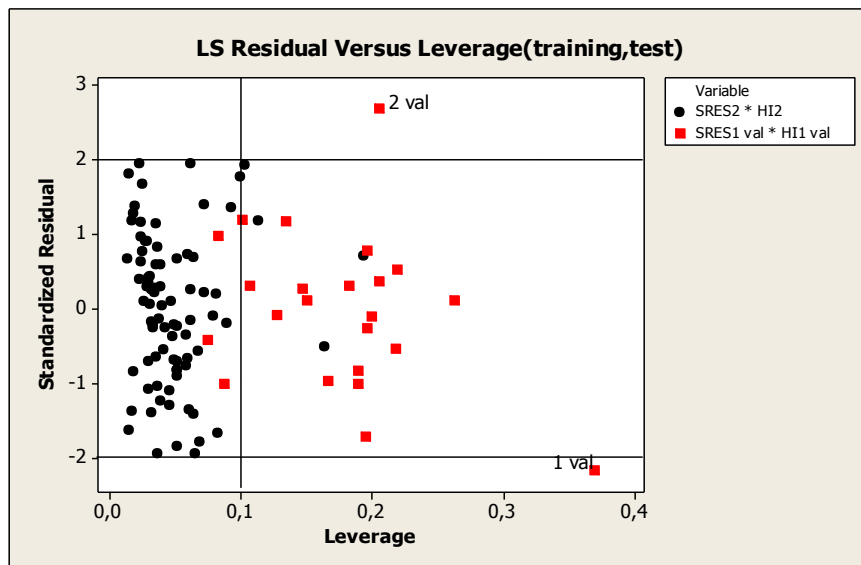
*La Méthode LAD (Calibration, validation).**La Méthode MLR (Calibration, validation).*

Figure 19: Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes.

Dans cette étape l'analyse de résidus (Figure 19) montre avec :

- **l'estimation LAD** : toutes les observations par la méthode LAD entre l'intervalle (-2,2) dans les deux ensembles (Calibration, Validation), ne présentent pas de points aberrants.

- **l'estimation MLR** :

- dans le cas de la Calibration : toutes les observations par la méthode MLR entre l'intervalle (-2,2), ne présentent pas de points aberrants.

- dans le cas de la validation : deux points peuvent être considérés comme valeurs aberrantes:

1- 3-methyl-5-(2-methylpentyl)-2-phenoxy pyrazine

2- 3-methyl-2-(phenylthio)pyrazine

Car ils se situent en dessous de la ligne de référence horizontale et situés à droite de la ligne verticale ($h_{ii} > 0.1$) ; ceci influe sur l'ajustement du modèle.

Sur cette colonne, 106 dérivés sur 114 ont été séparés avec la méthode LAD (83 dérivés sur 89 ont été séparés dans l'ensemble de Calibration et 23 dérivés sur 25 ont été séparés dans l'ensemble de validation), 106 dérivés sur 114 ont été séparés avec la méthode MLR (83 dérivés sur 89 ont été séparés dans l'ensemble de Calibration et 23 dérivés sur 25 ont été séparés avec la méthode MLR dans l'ensemble de validation (deux points aberrants restants).

Nous constatons que la méthode LAD est la plus efficace pour cette séparation chromatographique dans la stabilité et la robustesse par rapport à la méthode des moindres carrées après l'élimination de points aberrants.

La Colonne OV-101 est plus efficace pour cette séparation chromatographique avec la méthode LAD par rapport à la méthode MLR avec minimisation de valeurs aberrantes.

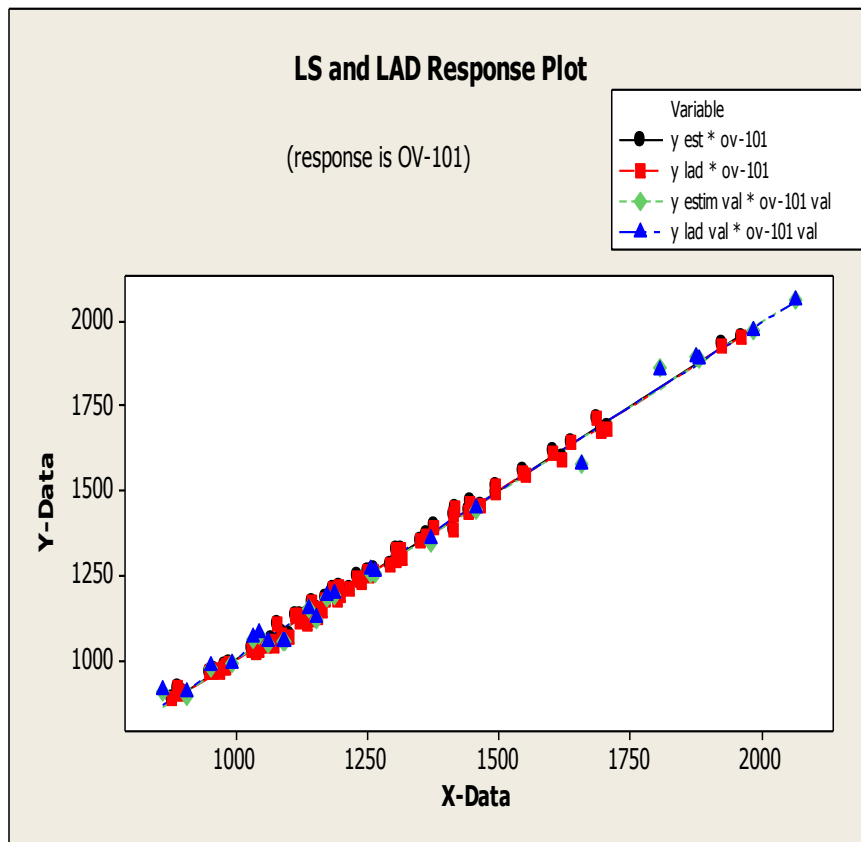
I-1-3-7- Confirmation de l'approche entre les deux méthodes : validation des résultats**1-La qualité de l'ajustement :**

Figure 20 : Valeur estimée en fonction des valeurs observées pour les deux méthodes.

On remarque une relation linéaire entre les valeurs estimées et les valeurs observées pour les deux méthodes (Figure 20), ceci indique que le modèle est correctement ajusté aux données pour les deux ensembles, ce qui prouve l'approche et la proportionnalité entre les deux méthodes.

2-Tests de normalité des erreurs :

Il existe un paquet de tests de normalité. En effet, grâce à la notion de robustesse, un test peut s'appliquer même si l'on s'écarte légèrement des conditions d'applications initiales. Dans ce point de vue, nous pouvons dès lors nous contenter de techniques simples (ex. statistique descriptives, techniques graphiques) pour vérifier si la distribution des données est réellement inconciliable avec la distribution normale.

2-1-Tests graphiques :

2-1-1-Probabilité Plot de l'erreur :

Pour vérifier la normalité des erreurs d'un modèle de régression on effectue la Probabilité plot des résidus.

1-La Méthode MLR

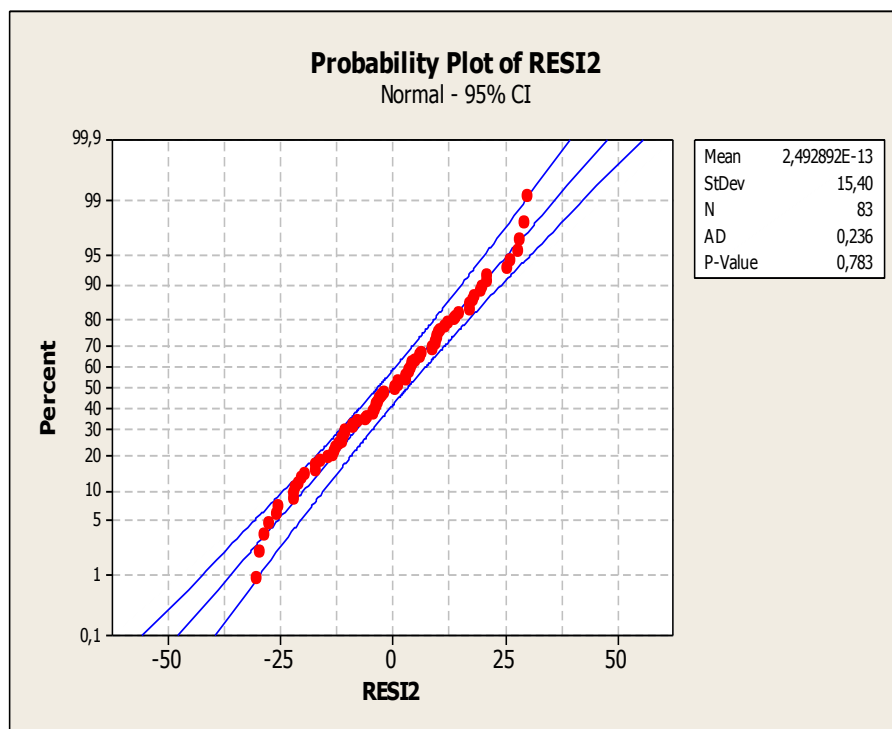


Figure 21: Diagramme des scores normaux (Calibration).

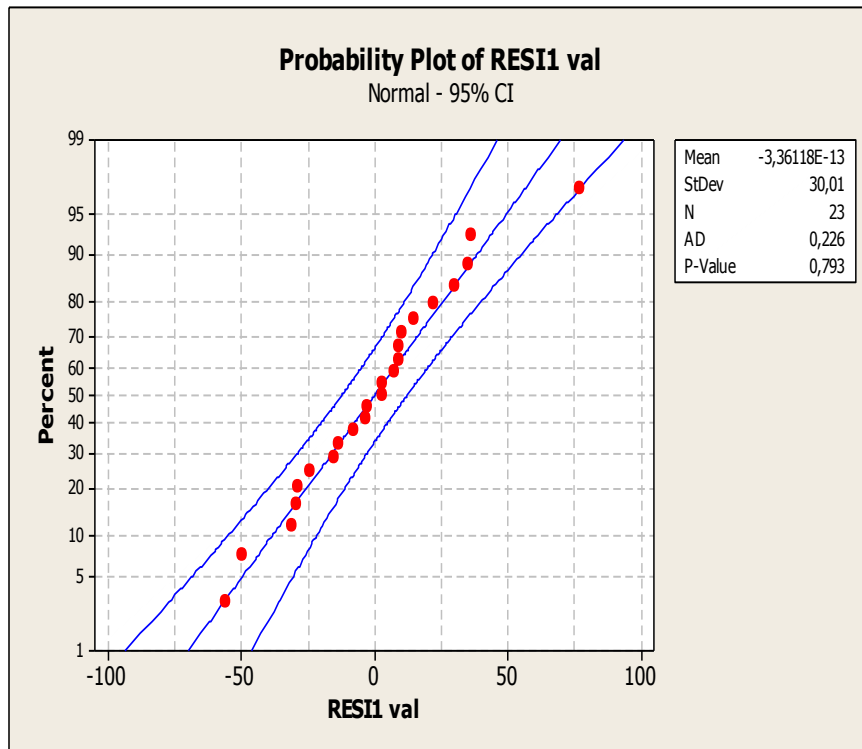
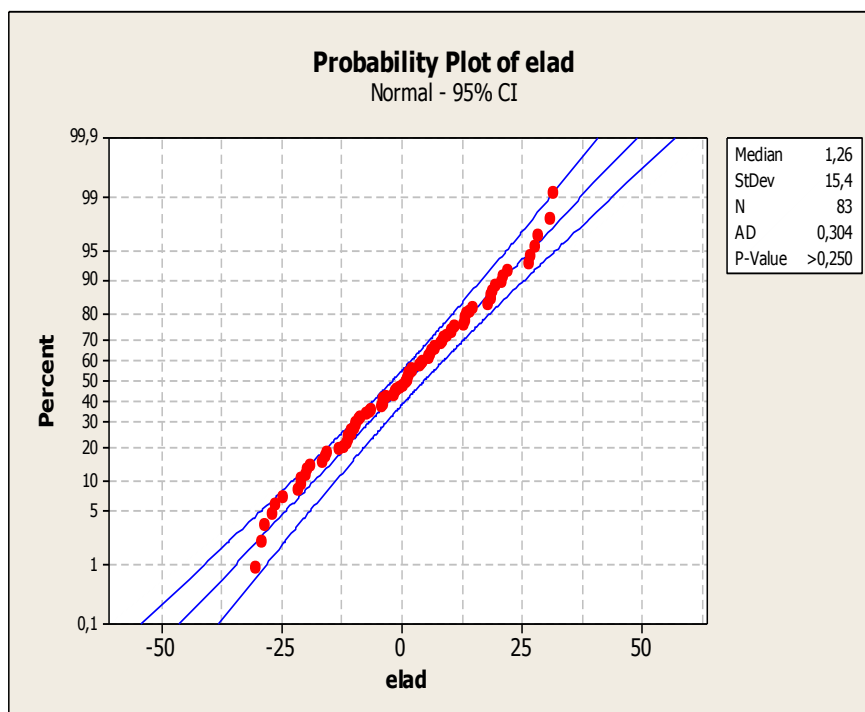


Figure 22: Diagramme des scores normaux (validation).

2-La Méthode LAD :



-Figure 23: Diagramme des scores normaux (Calibration).

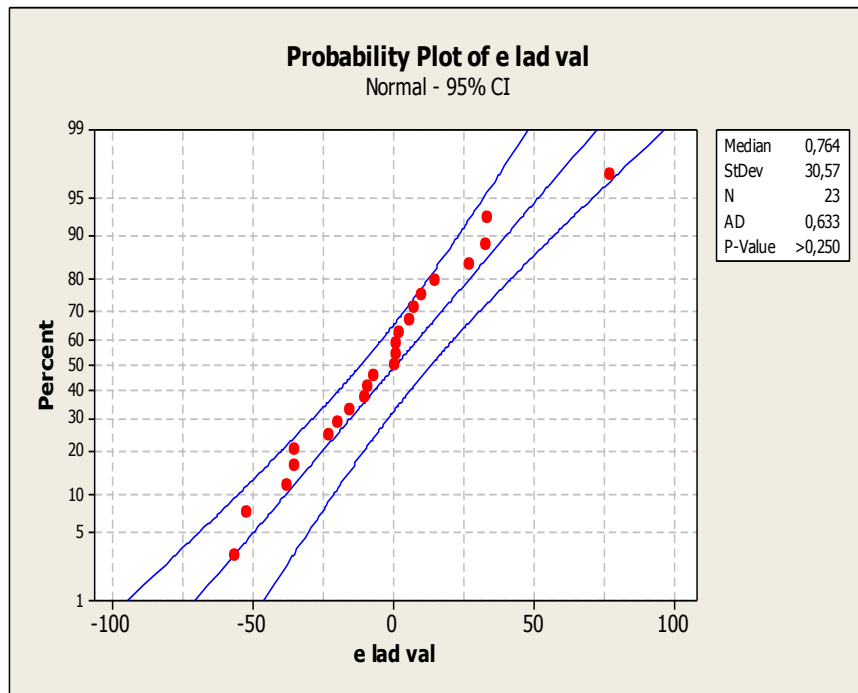


Figure 24: Diagramme des scores normaux (validation).

À partir d'une loi de probabilité Nous constatons que les points sont relativement alignés. Nous n'observons pas un écartement significatif, aucun point ne semble non plus se démarquer des autres (tous les point entre l'intervalle de confiance) (Figure 21, 22, 23,24)

2-2-Test statistiques:

Un autre critère lié au choix des tests statistiques est celui de la robustesse : la majorité des tests reposant sur un certain nombre d'hypothèses implicites, ou conditions d'application (normalité de la distribution des observations, etc.), la robustesse d'un test traduit sa tolérance à l'égard de la déviation par rapport à ces conditions d'applications.

Pour vérifier l'adéquation (la compatibilité) à la loi normale, nous présentons les tests statistiques de compatibilité à la loi normale, on choisit ce test si les résidus suivent une loi normale.

A tout test est associé un risque α dit de première espèce ; dans notre travail, nous adopterons le risque $\alpha = 5\%$.

1-Test d'Anderson-Darling: Dans notre travail, on trouve que La Méthode LAD : ($AD_{\text{Calibration}}=0,364$, $AD_{\text{validation}}=0,747$ et La Méthode MLR ($AD_{\text{Calibration}}=0,236$, $AD_{\text{validation}}=0,739$) $< AD_{\text{crit}}=0.752$ à 5%, l'hypothèse de normalité est compatible avec nos données pour les deux méthodes.

La différence est dans le calcul de la valeur p, Minitab se contente de spécifier une plage de la valeur p en comparant la statistique aux valeurs critiques relatifs aux niveaux de risque de la valeur $p > 0.10$, dans les données de la Calibration pour la méthode LAD ($p > 0.250$) et pour la méthode MLR ($p=0.783$) et dans les données de la validation pour la méthode LAD ($p > 0.250$) et pour la méthode MLR ($p=0.84$) en comparant les valeurs statistique aux valeur critique relatifs aux niveaux de risque de la valeur $p > 0.1$.

2-3-Intervalles de confiance

L'intervalle de confiance et le risque α constituent ainsi une approche complémentaire (une approche d'estimation) L'intervalle de confiance le plus utilisé est l'intervalle de confiance a $100(1 - \alpha) = 95\%$.

Les zones d'acceptation et de rejet pour une distribution normale, pour un seuil de décision $\alpha = 0:05$ sous l'hypothèse nulle H:

- dans le cas de la Calibration: l'estimation LAD :(-28.92, 31.44).

l'estimation MLR (-30.18, 30.18).

- dans le cas de la validation : l'estimation LAD (-59.15, 60.68).

l'estimation MLR (-58.82, 58.82).

Les données peuvent être compatibles avec hypothèse, qui exprime que la moyenne avec la méthode MLR (calibration, validation) et la médiane avec la méthode LAD (calibration,

validation) dans le centre de la zone d'acceptation de l'hypothèse nulle, vérifie la position à 95%.

-Remarque générale : On confirme exactement l'approche entre les deux méthodes soit statistiquement ou graphiquement par l'approche de la valeur de tous les tests statistiques et graphiques entre les deux méthodes.

I-2- Cas de la colonne polaire (CARBOWAX- 20M)

1-2-1-La régression avec la moindre carré :

Le modèle basé sur ces descripteurs avec la méthode MLR donne l'équation suivante :

$$Y = 853 + 513 \text{ RDCHI} - 636 \text{ GATS1p} + 32,7 \text{ Mor02m} \quad (53)$$

$S = 34,93$ $R\text{-Sq} = 98,08\%$, $R\text{-Sq}(\text{adj}) = 98,01\%$, $Q^2=97,86\%$, $Q^2_{\text{boo}}=97,72\%$, $Q^2_{\text{ext}}=77,02\%$,
 $F=1444,5$, $P=0,000$.

Le tableau XI résume les trois descripteurs utilisés pour la modélisation.

Tableau XI : Les descripteurs sélectionnés pour la modélisation d'IR.

Descripteur	Définition	Classe
RDCHI	Profils moléculaires Randic,	indice de connectivité
GATS 1p	l'autocorrélation Geary - lag 1 / pondéré par la pression .	Indices d'autocorrélation 2D
Mor 02m	3D-MoRSE - signal 02 / pondéré par la masse atomique.	3D-MoRSE (Morsw)

Les trois descripteurs ont été obtenus en utilisant le logiciel Dragon [6]. On trouvera plus d'informations concernant ces descripteurs dans le guide de l'utilisateur du logiciel Dragon (Todeschini R *et al.*, 2005) [8] et les références afférentes.

Les valeurs de $R^2=98.08\%$ et $R^2_{\text{adj}}=98.01\%$ montrent la qualité de l'ajustement, alors que la petite différence entre R^2 et $Q(=0.22)$ renseigne sur la robustesse du modèle qui est, en outre hautement significatif (grande valeur de la statistique F de Fisher= $F=1444,5, P=0.00$,

S= 34.93). La validation par $Q2_{EX} = 77.2\%$ confirme tout à la fois la bonne capacité de prédiction interne et la stabilité du modèle.

La matrice de corrélation, selon le tableau(XII) suggère que :

Les deux descripteurs importants ayant une corrélation positive avec les valeurs de IR sont RDCHI (R=0.893) Mor02m(R=0.896), ils présentent le plus grand impact sur l'indice de rétention (relation positive entre ce descripteur et l'indice de rétention), corrélation forte sur la phase CRW-20M.

Le descripteur important ayant une corrélation négative avec les valeurs de IR est GATS1p(R=-0.375), il présente le plus petit impact sur l'indice de rétention (relation négative entre ce descripteur et l'indice de rétention), corrélation moindre sur la phase CRW-20M.

Tableau XII : Matrice de corrélation

	IR (cw)	RDCHI	GATS1p
RDCHI	0,893 0,000		
GATS1p	-0,375 0,000	0,044 0,681	
Mor02m	0,896 0,000	0,930 0,000	-0,024 0,821

Le tableau XIII résume les résultats des caractéristiques des descripteurs sélectionnés par L'estimation MLR.

Tableau XIII : Évaluations de l'estimation MLR pour le model.

Predictor	Coef	SE coef	T	P	VIF
Constant	852,50	44,50	19,16	0	
RDCHI	512,40	33,40	15,34	0	7,625
GATS1p	-636,04	24,60	-25,85	0	1,035
Mor02m	32,685	4,612	7,09	0	7,614

- Les valeurs des probabilités de t pour les trois descripteurs sont nuls, ceci indique qu'ils sont hautement significatifs avec un risque d'erreur de première espèce $\alpha=0.05$, les paramètres sont tous significatifs parce que leurs estimations sont de l'ordre : $\beta_0 = 852.3$, $\beta_1 = 512.52$, $\beta_2 = -636.05$ et $\beta_3 = 32.671$.

La méthode MLR est la meilleure méthode de régression linéaire ; et la plus utilisée dans les études QSPR Sous certaines conditions :

1—la normalité des résidus :

-Pour le modèle de la colonne CRW-20M pour un niveau de $\alpha=0.05$, on a testé la normalité avec le test d'Anderson-Darling) (figure 25)($A_Squared = 0.270 < v_{cri} = 0.752$) les données d'observations des erreurs suivent une distribution normale mais quelques perturbations des points sont à signaler .

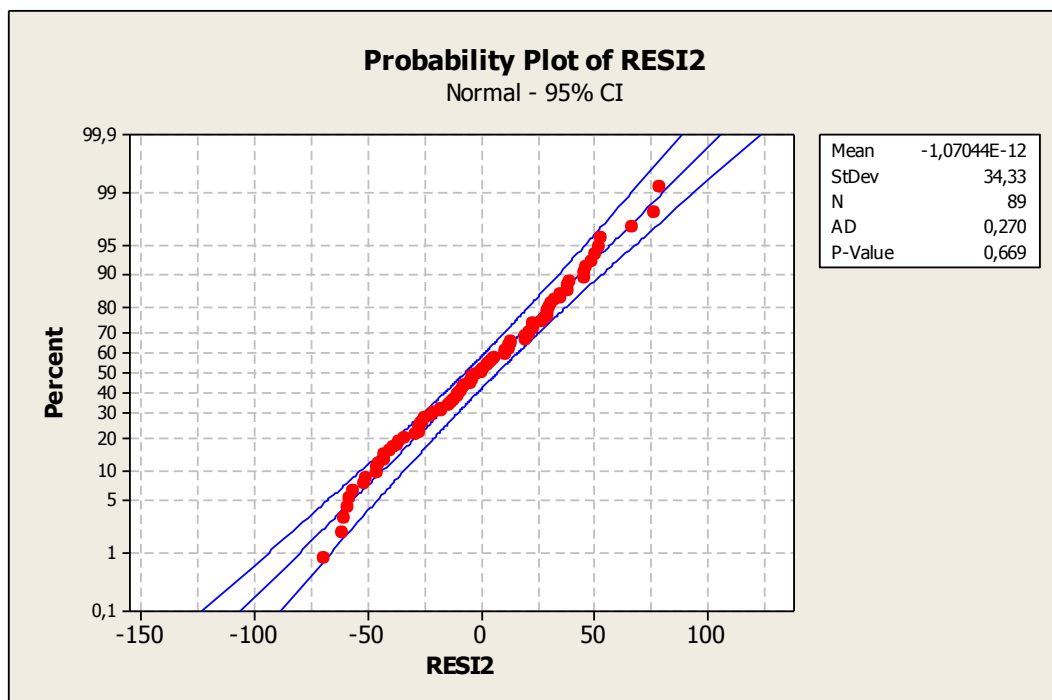


Figure 25: Diagramme des scores normaux.

2-auto corrélation des résidues :

Les valeurs des statistiques de Durbin-Watson (Durbin, et de Watson, 1951), pour la colonne CRW-20M : $dw=1.16$ sont inférieures aux valeurs données par les tables respectivement pour 3 régresseurs, pour le risque raisonnable $\alpha =0.05$ (valeur critique entre : ($d_L=1,55$, $d_U =1.72$) ce qui exprime l'auto corrélation positive des résidus (la présence de l'autocorrélation) (Figure 26) ;ceci indique le problème de la présence de valeurs aberrantes. les coefficients de régression ne sont plus efficaces dans ce cas ; pour palier on compare les coefficients de régression par la méthode MLR et les différentes méthodes d'estimation .

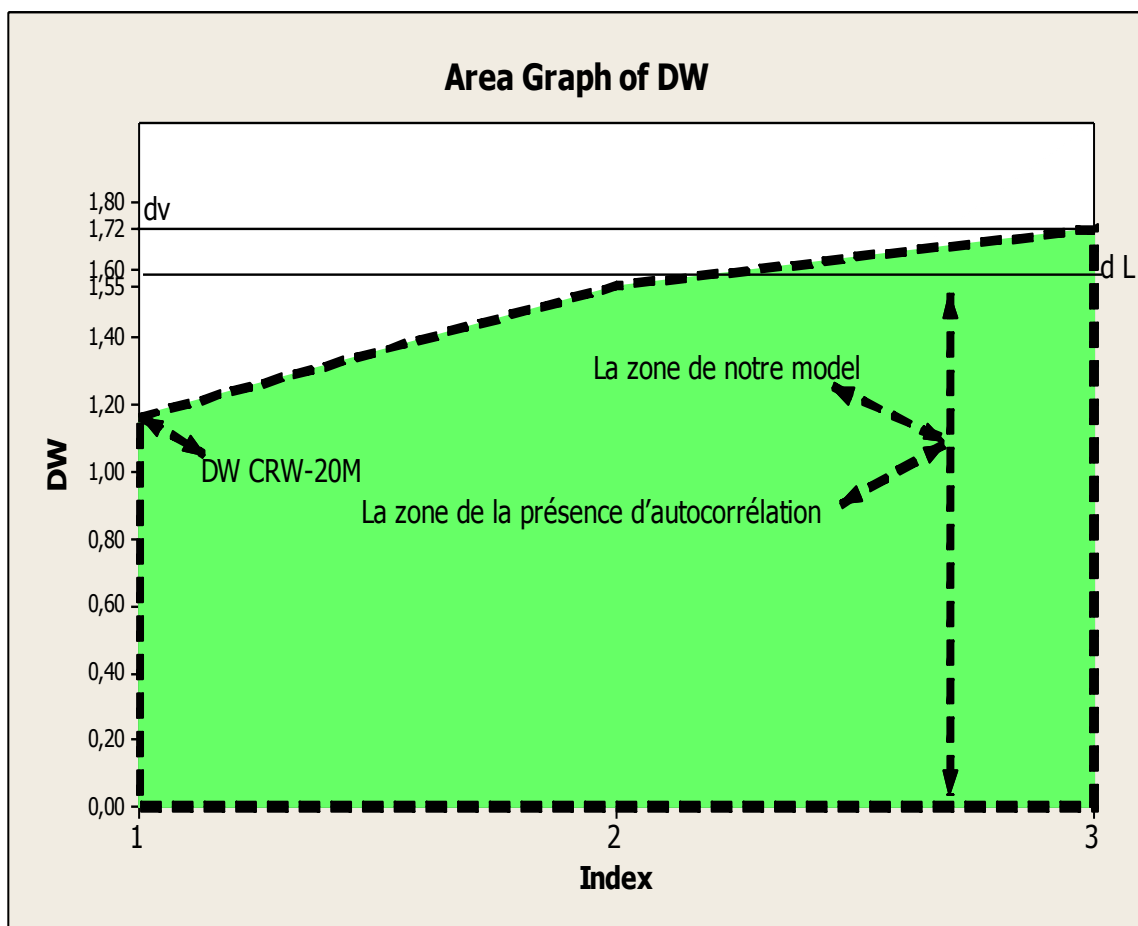


Figure 26 : Diagramme de valeur de Durbin Watson.

3-Test de Multicolinéarité :

-Le facteur d'inflation de la variance (FIV) permet de décrire l'importance de la multicolinéarité (la corrélation entre les prédicteurs) dans une analyse de régression (table XIII) et figure (27), Les valeurs des facteurs d'inflation de la variance (FIV) pour les deux descripteurs(RDCHI ,Mor02m) sont supérieures à 5 ceci indique qu'ils sont hautement corrélés donc le Coefficient de régression est mal estimé(en raison d'un problème de multicolinéarité à pourcentage faible lorsque cette valeur est de l'ordre de 5 à 10) on va tester ce problème après l'élimination des points aberrants,s'ils existent en utilisant la régression PLS.

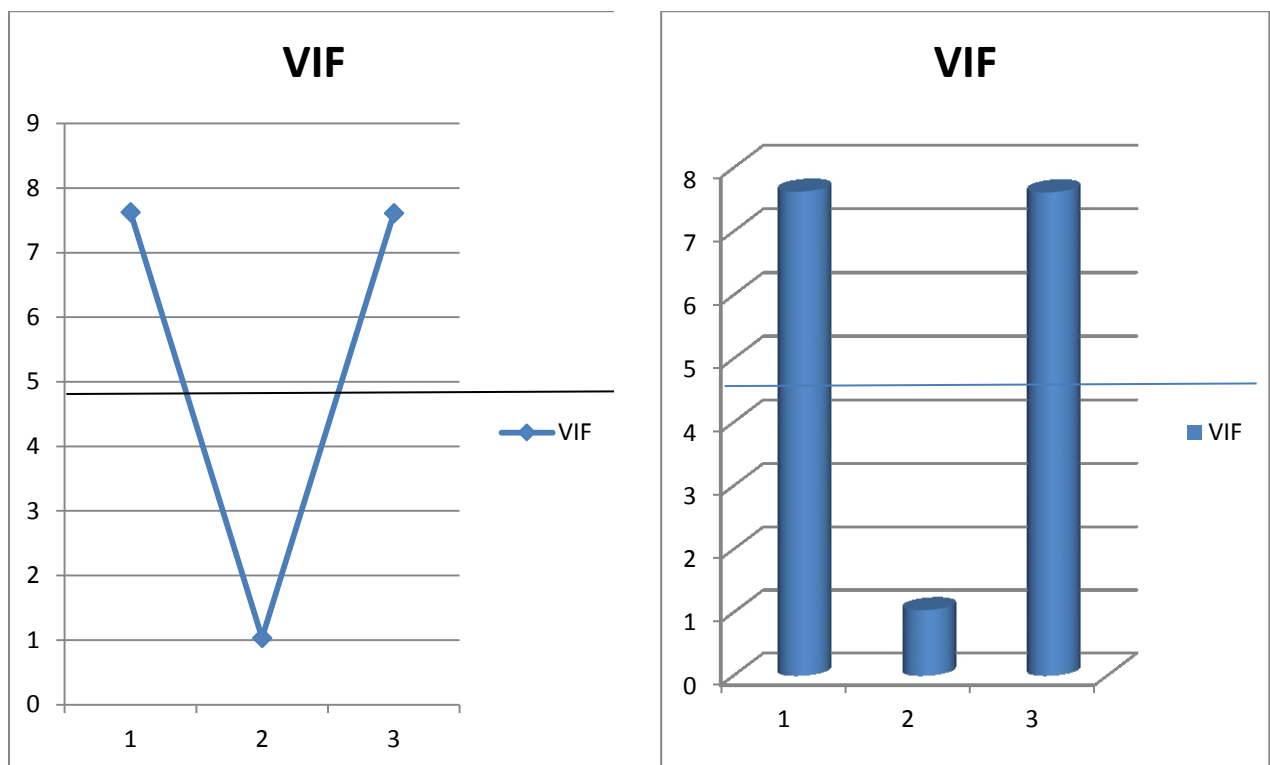


Figure 27 : Diagramme de VIF pour chaque descripteur.

1-2-2-Modèle de régression par la méthode LAD :**1-2-2-1-Programmes de calcul d'équation d'hyperplan par la méthode LAD :**

Les programmes de calculs de l'équation d'hyperplan de régression multiple LAD obtenus par le logiciel Matlab(R2009a) [7] compatible sur PC, écrits en langage Turbo Pascal (voir l'annexe 2). Conduit à l'équation de l'hyperplan par l'estimation LAD .

$$Y = 859.717 + 527.46 \text{ RDCHI} - 630.742 \text{ GATS1P} + 28.366 \text{ Mor02m} \quad (54)$$

Le tableau XIV résume les résultats des caractéristiques des descripteurs sélectionnés par L'estimation LAD.

Tableau XIV : Évaluations de l'estimation LAD pour le model.

Predictor	Coef	SE Coef	T	P
Constant	859,72	94,47	9,10	0,000
RDCHI	527,46	44,679	11,805	0,000
GATS1p	-630,74	20,68	-30,5	0,000
Mor02m	28.36	19,582	1,45	0,000

- Les valeurs des probabilités de t pour les trois descripteurs sont nuls, ceci indique qu'ils sont hautement significatifs avec un risque d'erreur de première espèce $\alpha=0.05$, les paramètres sont tous significatifs parce que leurs coefficients d'estimations sont de l'ordre: $\beta_0 = 859.72, \beta_1 = 527.46, \beta_2 = -630.74$ et $\beta_3 = 28.36$.

1-2-3-Comparaison de la Régression Robuste par la méthode MLR et la méthode LAD :**1-2-3-1- Comparaison des hyperplans par la méthode MLR et la méthode LAD :**

1-La méthode LAD: A partir de l'équation (54)

$$Y = 859,71 + 527,46 \text{ RDCHI} - 630,64 \text{ GATS1p} + 28,36 \text{ Mor02m}$$

2-La méthode MLR : A partir de l'équation (53)

$$Y = 852 + 513 \text{ RDCHI} - 636 \text{ GATS1p} + 32,7 \text{ Mor02m}$$

On remarque que les coefficients β calculés par la méthode MLR sont peu différents par rapport aux coefficients β calculés par la méthode LAD sur la Colonne CRW-20M (l'équation 53 et 54, Figure 8).

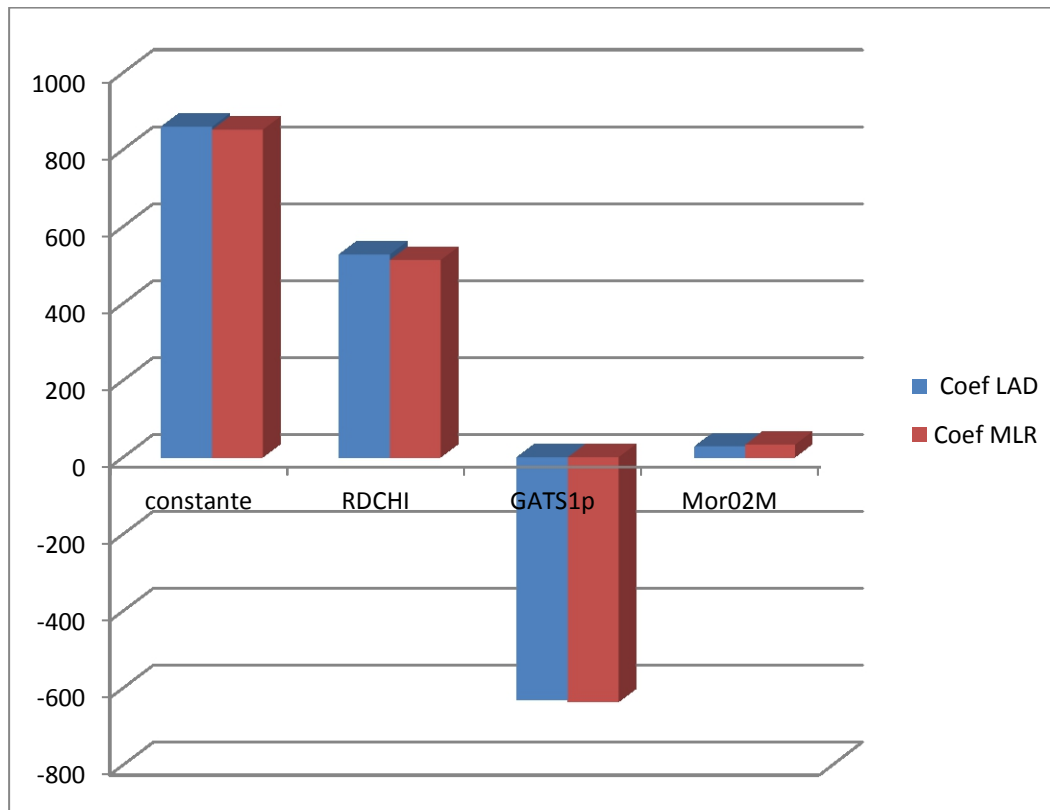


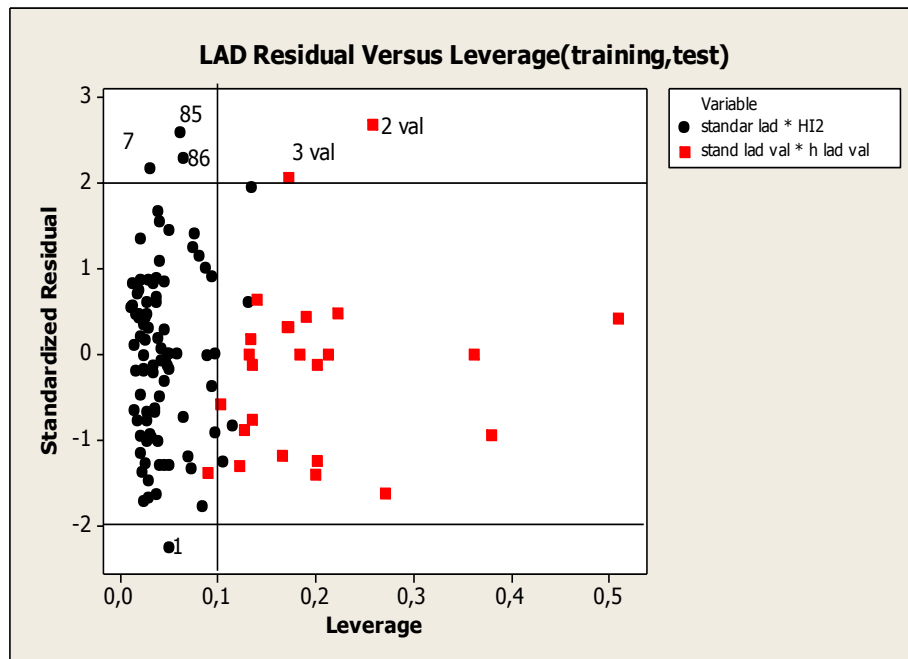
Figure 28 : Diagramme de comparaison des coefficients de régression entre les deux Méthodes.

Il est donc pertinent de refaire une vérification sur la présence des valeurs aberrantes, Puisque L'hyperplan de régression peut radicalement changer, avec le changement de coefficients de l'hyperplan

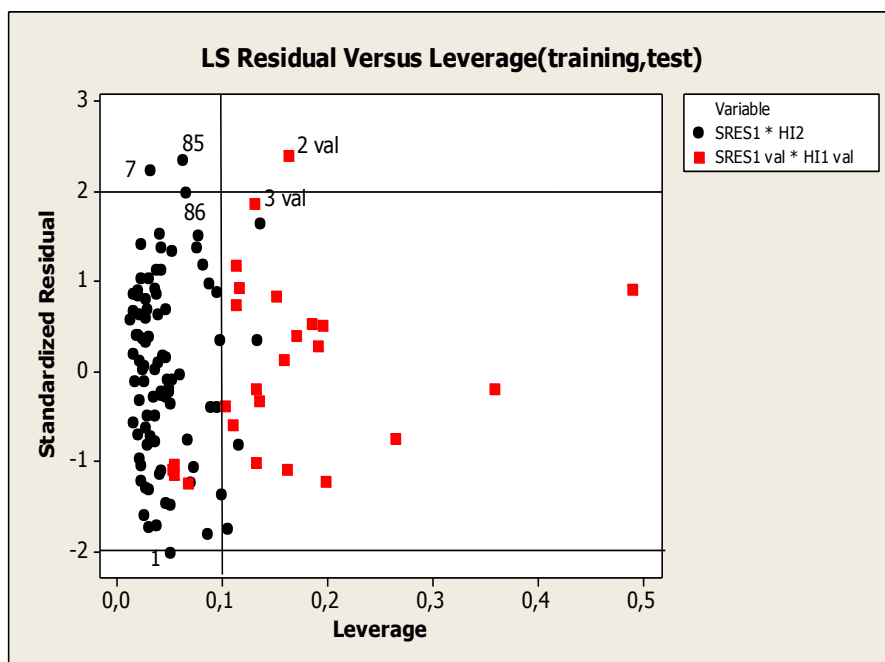
On trace le diagramme de Williams des LAD et MLR pour pouvoir les comparer :

1-2-3-2-Comparaisons graphiques des modèles alternatifs de régression :

1-domaine d'application :



La Méthode LAD (Calibration, validation)



La Méthode MLR (Calibration, validation)

Figure 29: Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes.

Dans ce diagramme (Figure 29) L'analyse des résidus pour les deux ensembles donne avec :

- l'estimation LAD :

- dans le cas de la Calibration : quatre points peuvent être considérés comme valeurs aberrantes:

1 -pyrazine

7 - Triméthylpyrazine

85- Phénoxypyrazine

86 - 2-méthyl-3-phénoxypyrazine

Car ils se situent en dessous de la ligne de référence horizontale.

- dans le cas de la validation : deux points peuvent être considérés comme valeurs aberrantes:

2 - (phénylthio)pyrazine

3 - 3-méthyl-2-(phénylthio)pyrazine.

car ils se situent en dessous de la ligne de référence horizontale et situés à droite de la ligne verticale ($h_{ii} > 0.1$) ; ceci influe sur l'ajustement du modèle.

- l'estimation MLR :

- dans le cas de la Calibration : trois points peuvent être considérés comme valeurs aberrantes:

1 - pyrazine.

7 - Triméthylpyrazine

85 - Phénoxypyrazine.

Car ils se situent en dessous de la ligne de référence horizontale.

- dans le cas de la validation : un point peut être considéré comme une valeur aberrante

2 - (phenylthio)pyrazine. Car il se situe en dessous de la ligne de référence horizontales et situés à droite de la ligne verticale ($h_{ii} > 0.1$) ; ceci influe sur l'ajustement du modèle.

2-La qualité de l'ajustement :

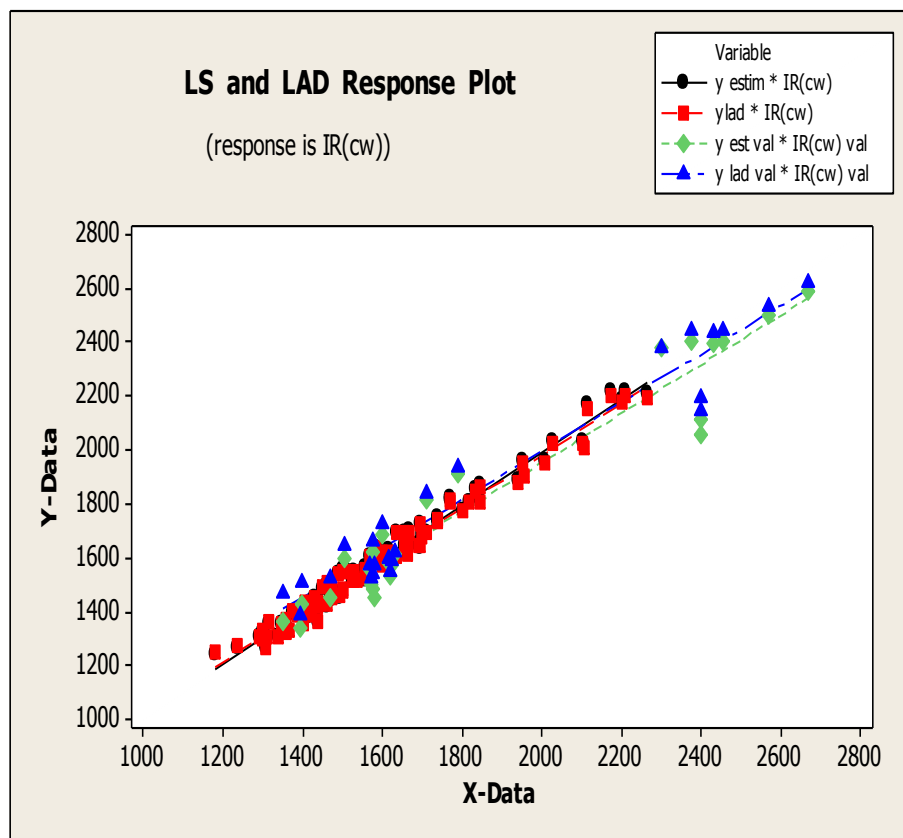


Figure 30: Valeurs estimées en fonctions de valeurs observées pour les deux méthodes.

On remarque une relation linéaire entre les valeurs estimées et les valeurs observées pour les deux méthodes la Figure (30) ceci indique que le modèle est correctement ajusté aux données pour les deux ensembles, Ce qui prouve l'approche et la proportionnalité entre les deux méthodes.

Il est donc pertinent de refaire une seconde étude après l'élimination des points aberrants communs entre les deux méthodes sur la colonne CRW -20M :

- dans le cas de la Calibration: on supprimé trois points communs :

1 - Pyrazine

7 - Tetramethylpyrazine

85 - Phenoxy pyrazine .

- dans le cas de la Validation : on supprimé un point commun:

2 -(phenylthio)pyrazine.

1-2-3-3-Comparaison des hyperplans par la méthode MLR et la méthode LAD:

1-La méthode LAD:

$$Y = 859,71 + 527,40 \text{ RDCHI} - 630,64 \text{ GATS1p} + 28.30 \text{ Mor02m} \quad (55)$$

2-La méthode MLR :

$$\text{IR(cw)} = 849 + 521 \text{ RDCHI} - 632 \text{ GATS1p} + 30,6 \text{ Mor02m} \quad (56)$$

On remarque que les coefficients β calculés par la méthode MLR se rapprochent des coefficients β calculés par la méthode LAD (l'équation 55 et 56, figure 31).

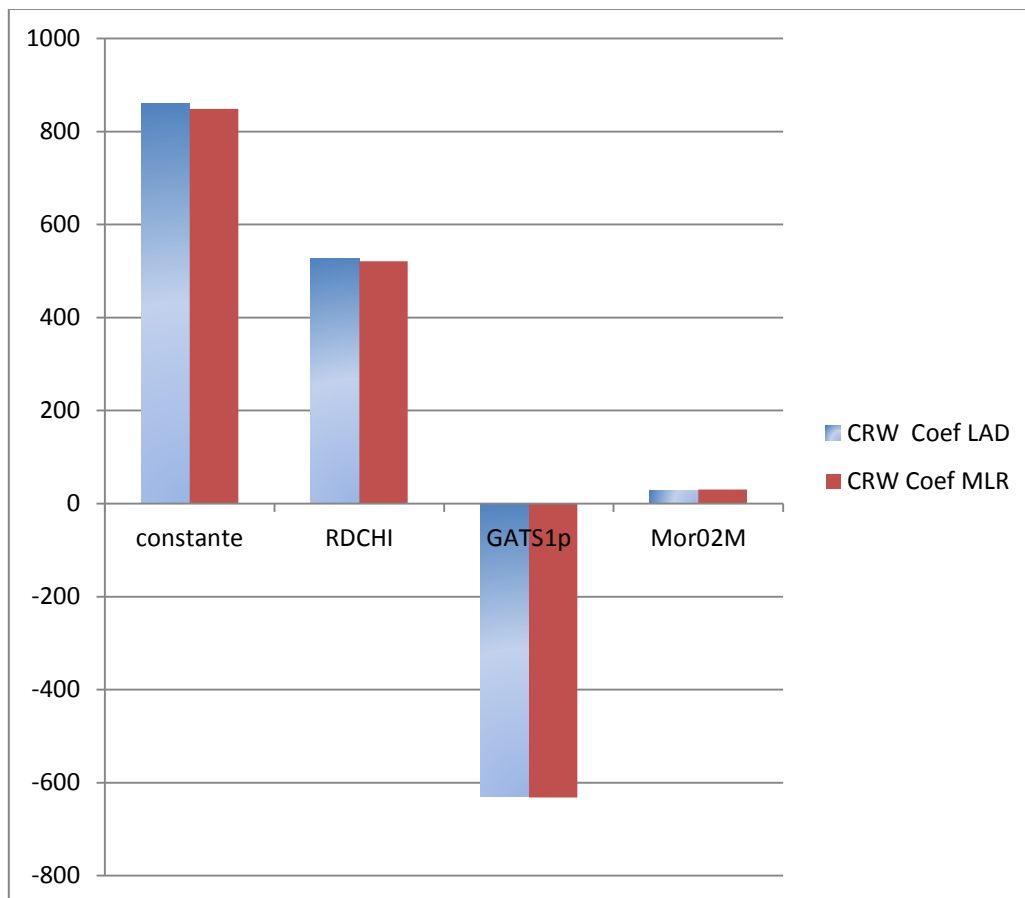


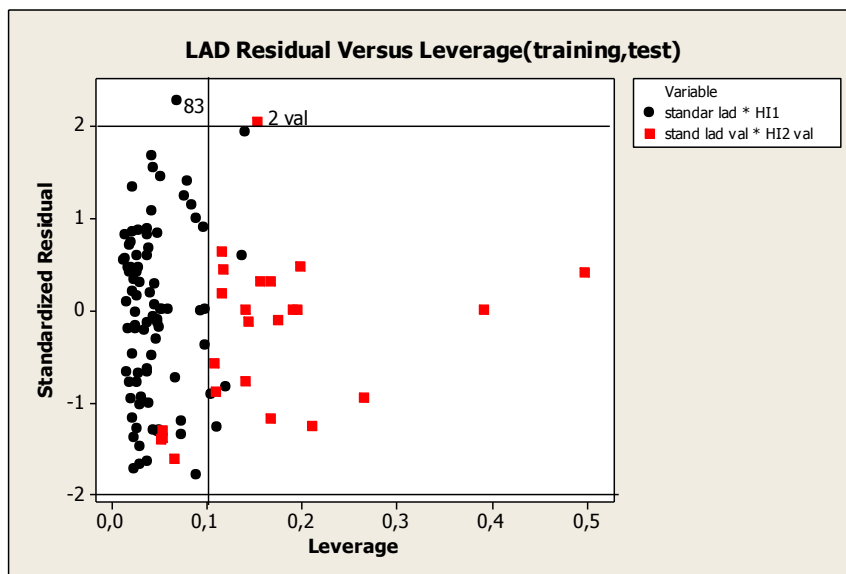
Figure 31: Diagramme de comparaison des coefficients de régression entre les deux Méthodes.

Il est donc pertinent de refaire une vérification de la présence de valeurs aberrantes.

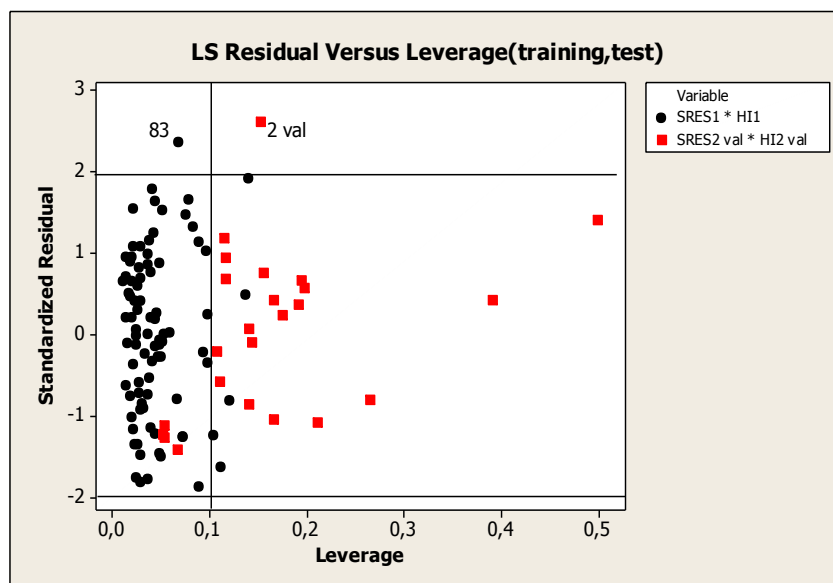
On trace le diagramme de Williams pour la méthode LAD et la méthode MLR pour pouvoir les comparer :

1-2-3-4--Comparaisons graphiques des modèles alternatifs de régression :

1-Domaine d'application :



La Méthode LAD (Calibration, validation).



La Méthode MLR (Calibration, validation)

Figure 32: Diagramme de Williams des valeurs Résidueles vers les valeurs de levier avec les deux méthodes

Dans ce diagramme (Figure 32) L'analyse des résidus pour les deux ensembles donne avec :

- l'estimation LAD :

- dans le cas de la Calibration : un point peut être considéré comme une valeur aberrante :

83 - 2-ethylthio-3-methyl-5-(2 methylbutyl)pyrazine).Car il se situe en dessous de la ligne de référence horizontale.

- dans le cas de la validation : un point peut être considéré comme une valeur aberrante :

2 - 3-methyl-2-(phenylthio)pyrazine. Car il se situe en dessous de ligne de référence horizontale et situés à droite de la ligne verticale ($h_{ii} > 0.1$) ; ceci influe sur l'ajustement du modèle.

- l'estimation MLR :

- dans le cas de la Calibration : un point peut être considéré comme une valeur aberrante

83 - 2-ethylthio-3-methyl-5-(2 methylbutyl)pyrazine).Car il se situe en dessous de la ligne de référence horizontale.

- dans le cas de la validation : un point peut être considéré comme une valeur aberrante

2 - 3-methyl-2-(phenylthio)pyrazine. Car il se situe en dessous de ligne de référence horizontale et situés à droite de la ligne verticale ($h_{ii} > 0.1$) ; ceci influe sur l'ajustement du modèle.

On remarque que les deux méthodes gardent les valeurs aberrantes, il est donc pertinent de refaire une seconde étude après l'élimination les points aberrants communs entre les deux méthodes sur la colonne CRW -20M :

-dans le cas de la Calibration : on supprime un point commun entre les deux méthodes

83 - Phenoxy pyrazine.

-dans le cas de la Validation : on supprime un point commun entre les deux méthodes

2 - 3-methyl-2-(phenylthio) pyrazine.

1-2-3-5-Comparaison des hyperplans par la méthode MLR et la méthode LAD:

1-La méthode LAD:

$$Y = 859,71 + 527,46 \text{ RDCHI} - 630,64 \text{ GATS1p} + 28,36 \text{ Mor02m} \quad (57)$$

2- La méthode MLR :

$$\text{IR}(cw) = 842 + 527 \text{ RDCHI} - 625 \text{ GATS1p} + 29,2 \text{ Mor02m} \quad (58)$$

On remarque que les coefficients β calculés par la méthode MLR sont plus proches des coefficients β calculés par la méthodes LAD (l'équation 57 et 58, figure 33).

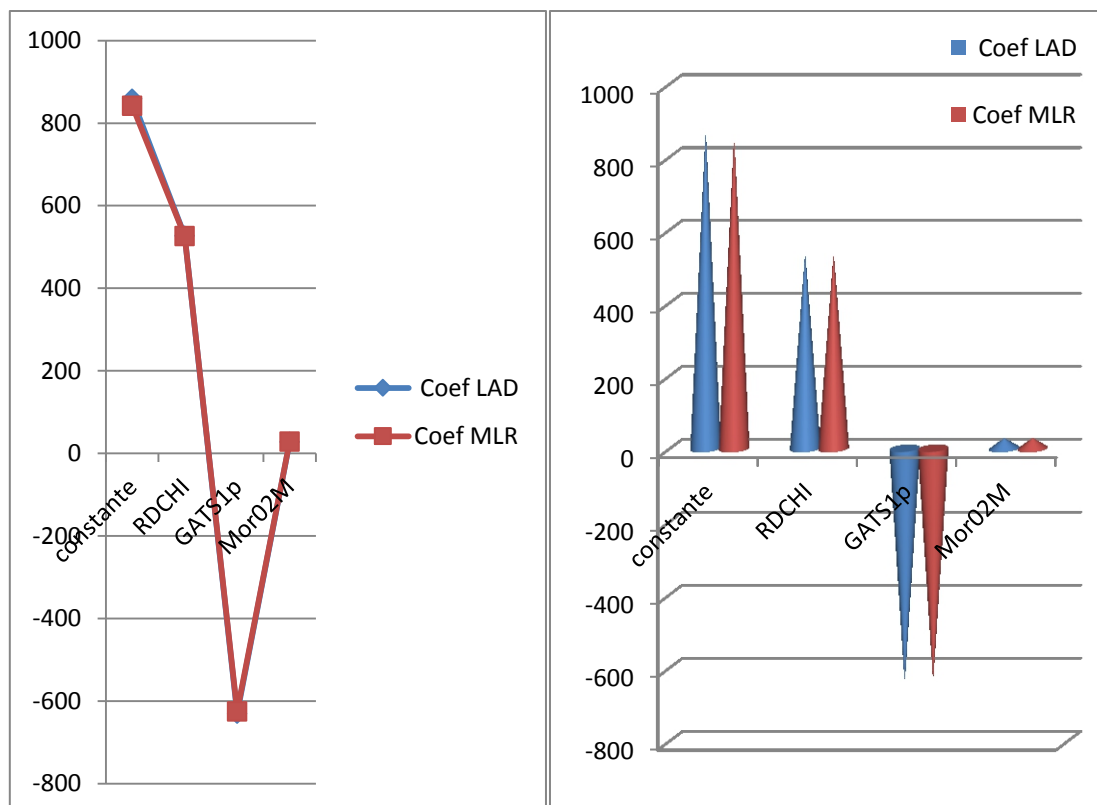


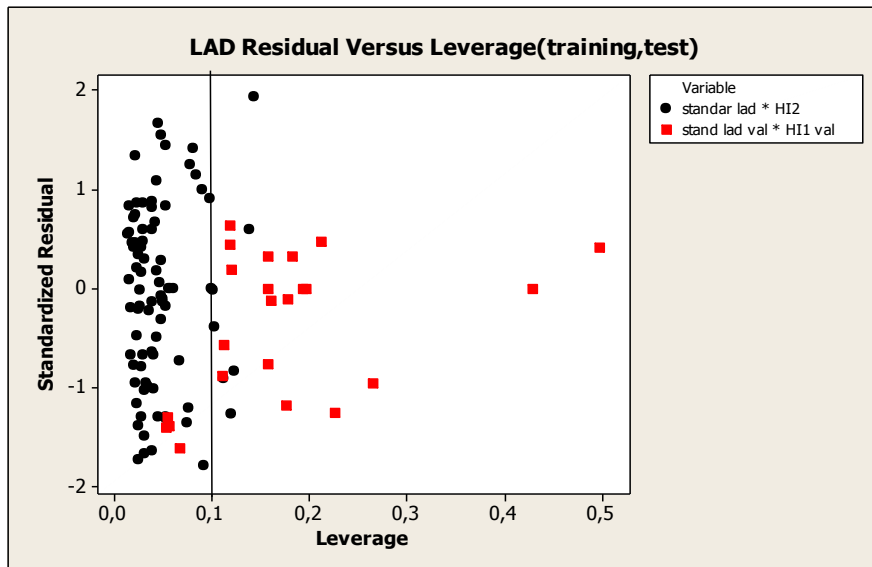
Figure 33: Diagramme de comparaison des coefficients de régression entre les deux Méthodes.

Il est donc pertinent de refaire un test graphique pour confirmer l'état proche et la vérification de la présence des valeurs aberrantes.

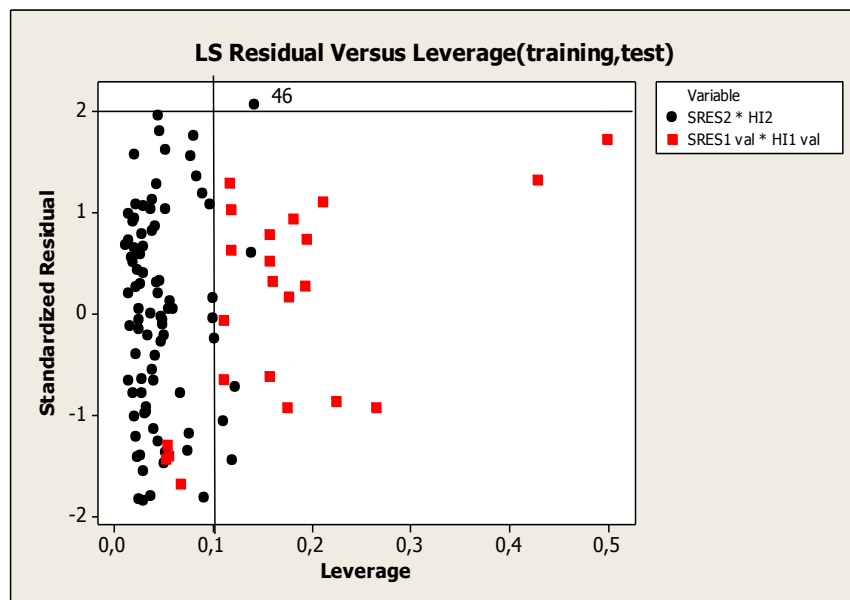
On trace le diagramme de Williams pour la méthode LAD et la méthode MLR pour pouvoir les comparer :

1-2-3-6-Comparaisons graphiques des modèles alternatifs de régression :

1-Domaine d'application :



La Méthode LAD (Calibration, validation).



La Méthode MLR (Calibration, validation).

Figure 34: Diagramme de Williams des valeurs Résiduelles vers les valeurs de levier avec les deux méthodes.

Dans cette étape l'analyse de résidus (Figure 34) pour les deux ensembles donne avec :

- **l'estimation LAD** : toutes les observations par la méthode LAD entre l'intervalle (-2,2) dans les deux ensembles (Calibration, Validation), ne présentent pas de points aberrants.

- **l'estimation MLR** :

- dans le cas de la Calibration : toutes les observations par la méthode MLR entre l'intervalle (-2,2), ne présentent pas de points aberrants.

- dans le cas de la validation : un point peut être considéré comme une valeur aberrante

-46 : 2-methyl-6-(2-methylbutyl)-3-octylpyrazine. Car il se situe en dessous de la ligne de référence horizontale et situés à droite de la ligne verticale ($h_{ii} > 0.1$) ; ceci influe sur l'ajustement du modèle.

Sur cette colonne, 108 dérivés sur 114 ont été séparés avec la méthode LAD (85 dérivés sur 89 ont été séparés dans l'ensemble de calibration et 23 dérivés sur 25 ont été séparés dans l'ensemble de validation) ,108 dérivés sur 114 avec la méthode MLR (85 dérivés sur 89 dans l'ensemble de calibration (reste un point aberrante) et 23 dérivés sur 25 ont été séparés dans l'ensemble de validation).

Nous constatons que la méthode LAD est la plus efficace pour cette séparation chromatographique, dans la stabilité et la Robustesse par rapport à la méthode des moindres carrée après élimination de points aberrants.

La Colonne CARBOWAX est plus efficace pour cette séparation chromatographique avec la méthode LAD par rapport à la méthode MLR avec minimisation de valeurs aberrantes.

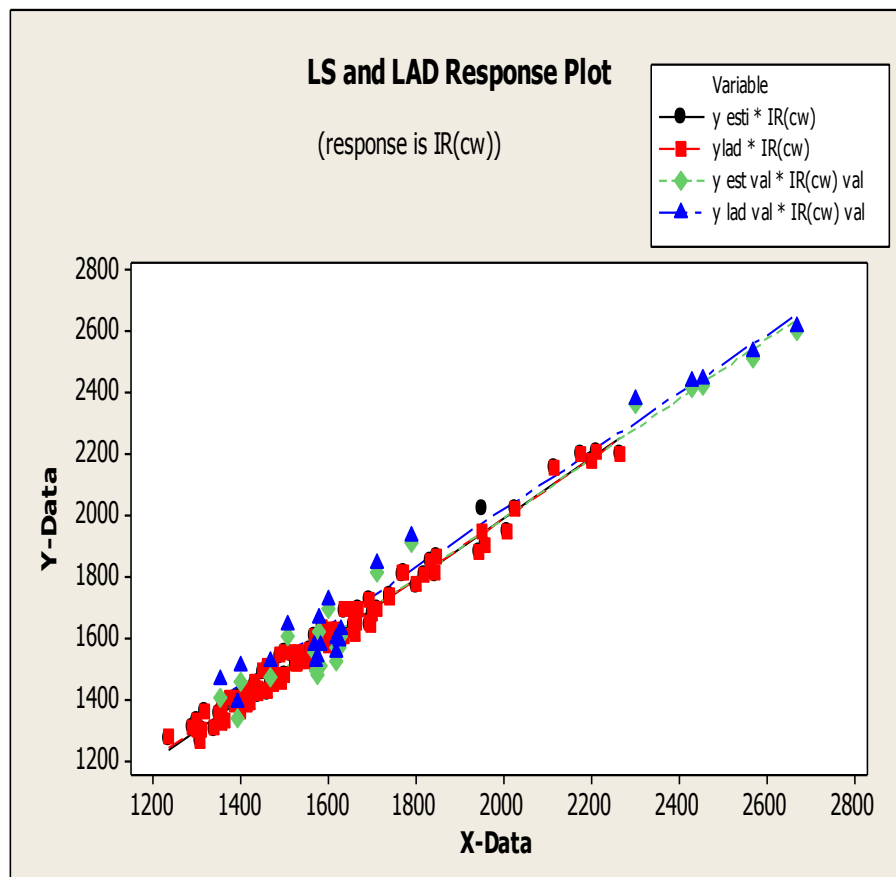
1-2-3-7- Confirmation de l'approche entre les deux méthodes : validation des résultats.**1-La qualité de l'ajustement :**

Figure 35 : valeur estimée en fonction des valeurs observées pour les deux méthodes.

On remarque une relation linéaire entre les valeurs estimées et les valeurs observées pour les deux méthodes, la Figure (35) montre que le modèle est correctement ajusté aux données pour les deux ensembles, Ce qui prouve l'approche et la proportionnalité entre les deux méthodes.

2-Tests de normalité des erreurs :

Il existe un paquet de tests de normalité En effet, grâce à la notion de robustesse, un test peut s'appliquer même si l'on s'écarte légèrement des conditions d'applications initiales. Dans ce point de vue, nous pouvons dès lors nous contenter de techniques simples (ex. statistique

descriptives, techniques graphiques) pour vérifier si la distribution des données est réellement inconciliable avec la distribution normale.

2-1-Tests graphiques :

2-1-1-Probabilité Plot de l'erreur :

Pour vérifier la normalité des erreurs d'un modèle de régression on effectue la Probabilité plot pour chaque méthode et pour les deux ensembles (Calibration, Validation):

1-La Méthode MLR :

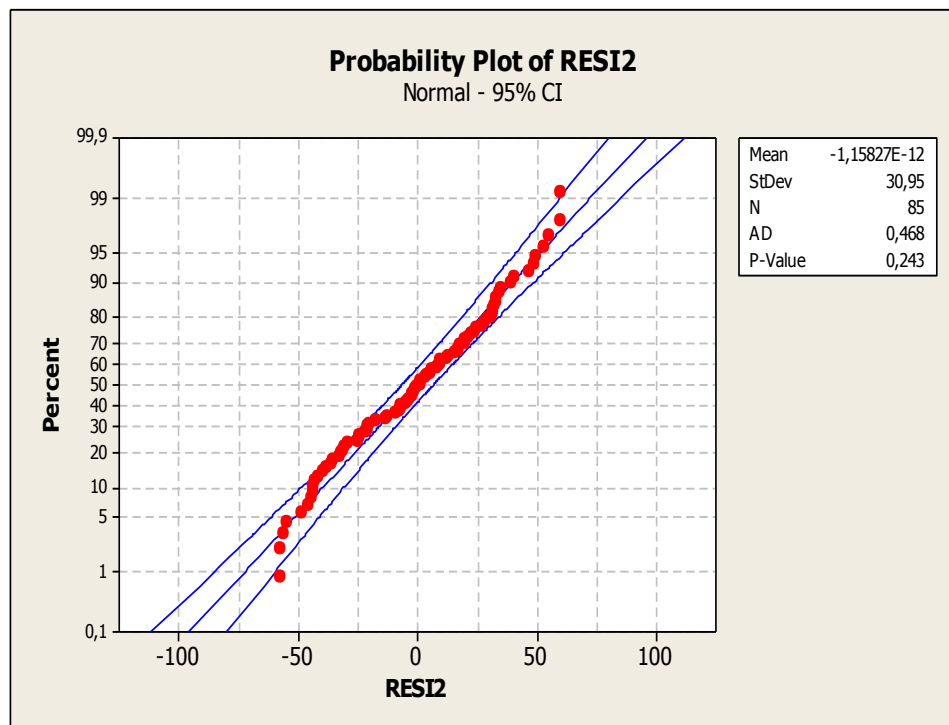


Figure 36: Diagramme des scores normaux (Calibration).

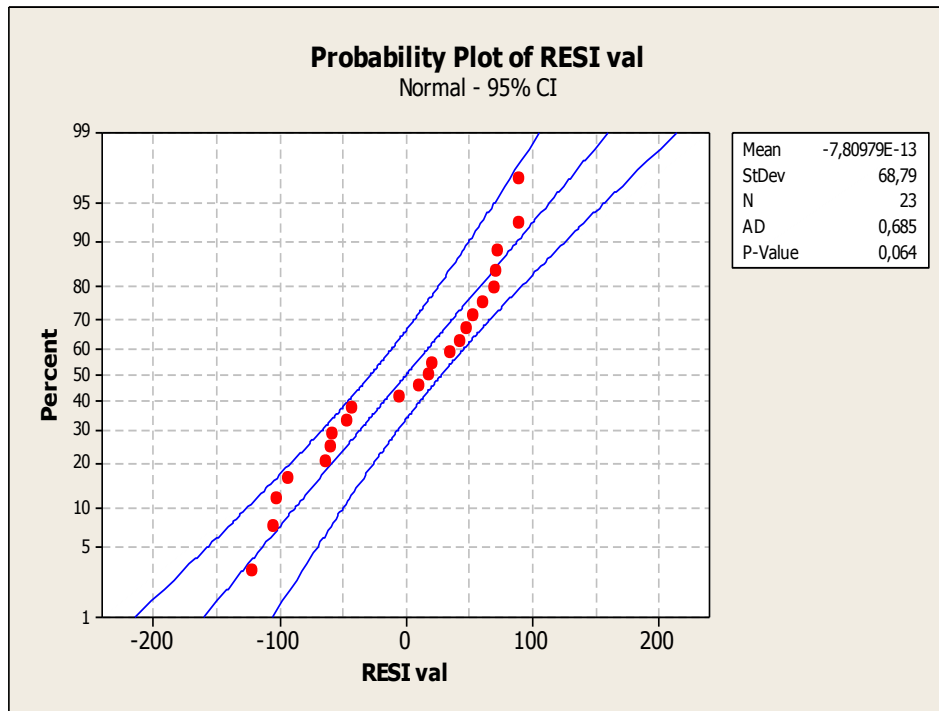


Figure 37: Diagramme des scores normaux (validation).

2-La Méthode LAD :

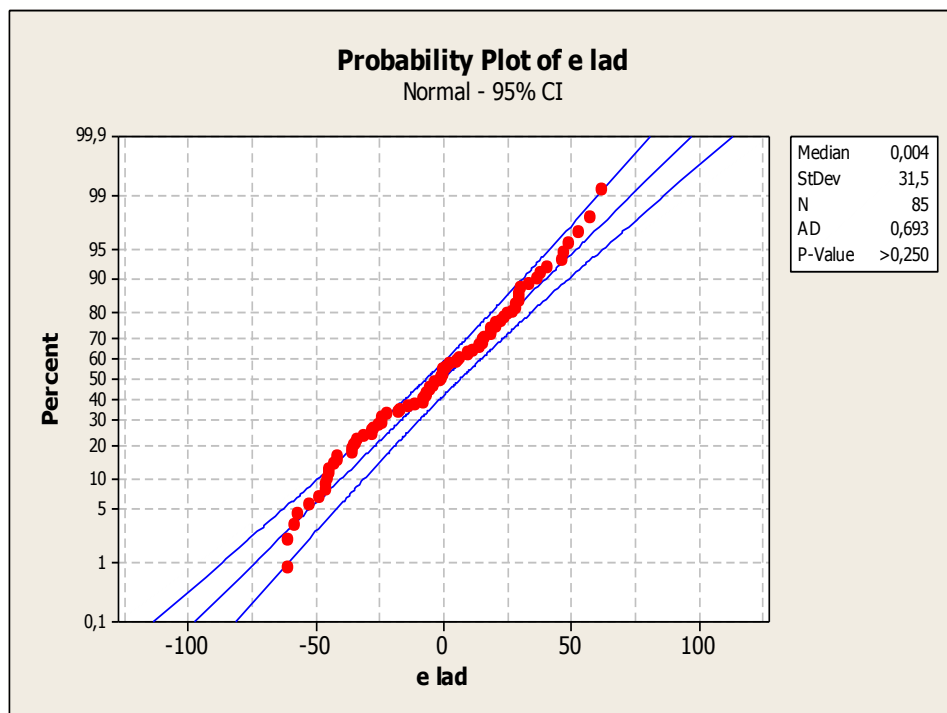


Figure 38: Diagramme des scores normaux (Calibration).

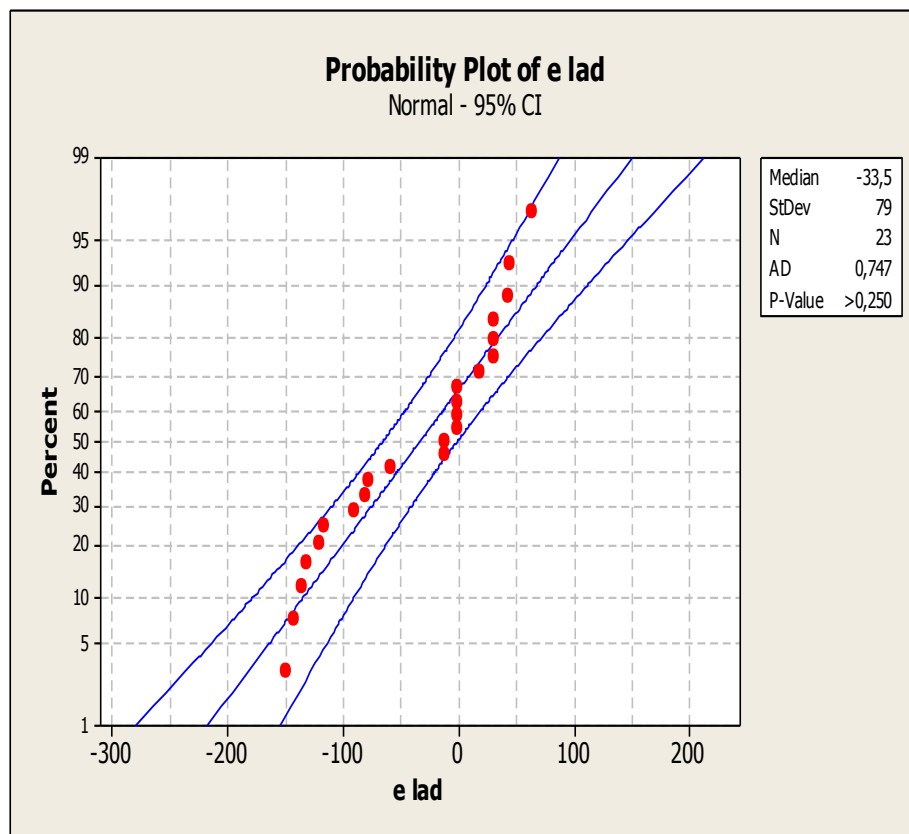


Figure 39: Diagramme des scores normaux (validation).

À partir d'une loi de probabilité Nous constatons que les points sont relativement alignés. Nous n'observons pas un écartement significatif, aucun point ne semble non plus se démarquer des autres (tous les points entre l'intervalle de confiance) (Figure 36, 37, 38,39).

2-2-Test statistiques:

Un autre critère lié au choix des tests statistiques est celui de la robustesse : la majorité des tests reposant sur un certain nombre d'hypothèses implicites, ou conditions d'application (normalité de la distribution des observations, etc.), la robustesse d'un test traduit sa tolérance à l'égard de la déviation par rapport à ces conditions d'application.

Pour vérifier l'adéquation (la compatibilité) à la loi normale, nous présentons le test statistique de compatibilité à la loi normale, on choisit ce test si les résidus suivent une loi normale.

A ce test est associé un risque α dit de première espèce; dans notre travail, nous adopterons le risque $\alpha = 5\%$.

1-Test d'Anderson-Darling:

Dans notre travail, on nous trouve que ($AD_{\text{Calibration}}=0,693$, et $AD_{\text{validation}}=0,747$ pour La Méthode LAD aussi $AD_{\text{Calibration}}=0,468$ et $AD_{\text{validation}}=0,685$ pour la Méthode MLR $< AD_{\text{crit}}=0.752$ à 5%, l'hypothèse de normalité est compatible avec nos données avec les deux Méthodes.

La différence est dans le calcul de la valeur p, Minitab se contente de spécifier une plage de la valeur p en comparant les valeurs statistiques aux valeurs critiques relatifs aux niveaux de risque la valeur $p > 0.10$, dans les données de Calibration pour la méthode LAD ($p > 0.250$) et pour la méthode MLR ($p = 0.243$) et dans les données de validation pour la méthode LAD ($p > 0.250$) sauf dans le donnée de validation pour la méthode MLR ($p = 0.046$) en comparant le valeur statistique aux valeur critique relatifs aux niveaux de risque de la valeur la valeur $p < 0.05$

2-3-Intervalles de confiance

L'intervalle de confiance et le risque α constituent ainsi une approche complémentaire (une approche d'estimation) L'intervalle de confiance le plus utilisé est l'intervalle de confiance à $100(1 - \alpha) = 95 \%$.

Les zones d'acceptation et de rejet pour une distribution normale, pour un seuil de décision $\alpha = 0:05$ sous l'hypothèse nulle H:

- dans le cas de la Calibration : l'estimation LAD : (-61.70, 61.78).

l'estimation MLR (-60.66, 60.66),

- dans le cas de la validation : l'estimation LAD (-188.3, 121.3).

l'estimation MLR (-135.0, 135.0) .

Les données peuvent être compatibles avec hypothèse qui exprime que la moyenne avec La méthode MLR (Calibration, validation) et la médiane avec La méthode LAD (Calibration, validation) dans le centre de la zone d'acceptation de l'hypothèse nulle, vérifie la position à 95°/°.

Remarque générale : On a confirmé exactement l'approche entre les deux méthodes soit statistiquement ou graphiquement par l'approche de la valeur de tous les tests statistiques et graphiques entre les deux méthodes.

Dans l'état proche on supprime 6 composés, donc on a 108 dérivée (85 Calibration, 23 validation) ; on va vérifier le problème de multicollinéarité.

1-2-4-La régression de Model avec MLR :

Le tableau XV résume les résultats des caractéristiques des descripteurs sélectionnés par L'estimation MLR

Tableau XV: Evaluations de l'estimation MLR pour le model.

Predictor	Coef	SE Coef	T	P	VIF
Constant	841,95	41,17	20,45	0,000	
RDCHI	526,58	30,55	17,24	0,000	7,210
GATS1p	-624,82	22,71	-27,51	0,000	1,027
Mor02m	29,204	4,238	6,89	0,000	7,170

S = 31, 5137 R-Sq = 98, 3% R-Sq(adj) = 98,2%, F=1541,79 =P 0,000

- Les valeurs des probabilités de T pour les trois descripteurs sont nuls, ceci indique qu'ils sont hautement significatifs avec un risque d'erreur de première espèce $\alpha=0.05$. Les paramètres sont tous significatifs parceque. Leurs estimations sont de l'ordre: $\beta_0 = -841.48, \beta_1 = 526.66, \beta_2 = -624.81$ et $\beta_3 = 29.204$.

La valeur de $R^2 = 98.2\%$ (ajustement), montre la qualité de l'ajustement, alors que la valeur très élevée du rapport de la variance expliquée par le modèle à la variance résiduelle ($F = 1541,79$; $p = 0,000$) montre que le modèle permet une très bonne prédiction des ($n = 85$) valeurs de IR de l'ensemble de calibration, (erreur standard $S = 31.51$) ; la valeur très élevée de Q^2_{LOO} , qui diffère très peu de celle de R^2 , renseigne sur la robustesse du modèle.

Test de multicollinéarité :

Les facteurs d'inflation de la variance (Tableau (XV)) détectent un problème de multicollinéarité à pourcentage faible, lorsque cette valeur est de l'ordre de 5 à 10 ; on parle de forte multicollinéarité pour des valeurs > 10 , on trouvera une explication en utilisant la régression PLS.

1-2-5-La Régression de modèle avec la méthode PLS

Le tableau XVI résume la méthode et les nombres des composants.

Tableau XVI: méthode et nombre des composants par l'estimation PLS.

Validation croisée	Leave-one-out
Composantes à évaluer	Ensemble
Nombre de composantes évaluées	3
Nombre de composantes sélectionnées	3

-D'après ces résultats (tableau XVI), la validation croisée a été utilisée la méthode et le modèle avec 3 composantes a été sélectionné.

Le tableau XVII résume le Choix du Modèle et la Validation pour IR(cw)

Tableau XVII: Choix du Modèle et Validation pour IR(cw) par l'estimation PLS.

Composants	X Variance	Error	R-Sq	PRESS	R-Sq(pred)
1	0,62362	299823	0,934405	308048	0,932606
2	0,97514	106381	0,976726	118202	0,974140
3	1,00000	84358	0,981544	94415	0,979344

-D'après ces résultats (tableau XVII) et le diagramme de sélection du modèle (Figure 40) ;

Minitab a sélectionné le modèle à 3 composantes qui a un $R^2_{(pred)}$ d'environ 97.93 % avec

$R^2=98.15\%$. D'après la variance X, le modèle à 3 composantes explique presque 100 % de la variance des prédicteurs.

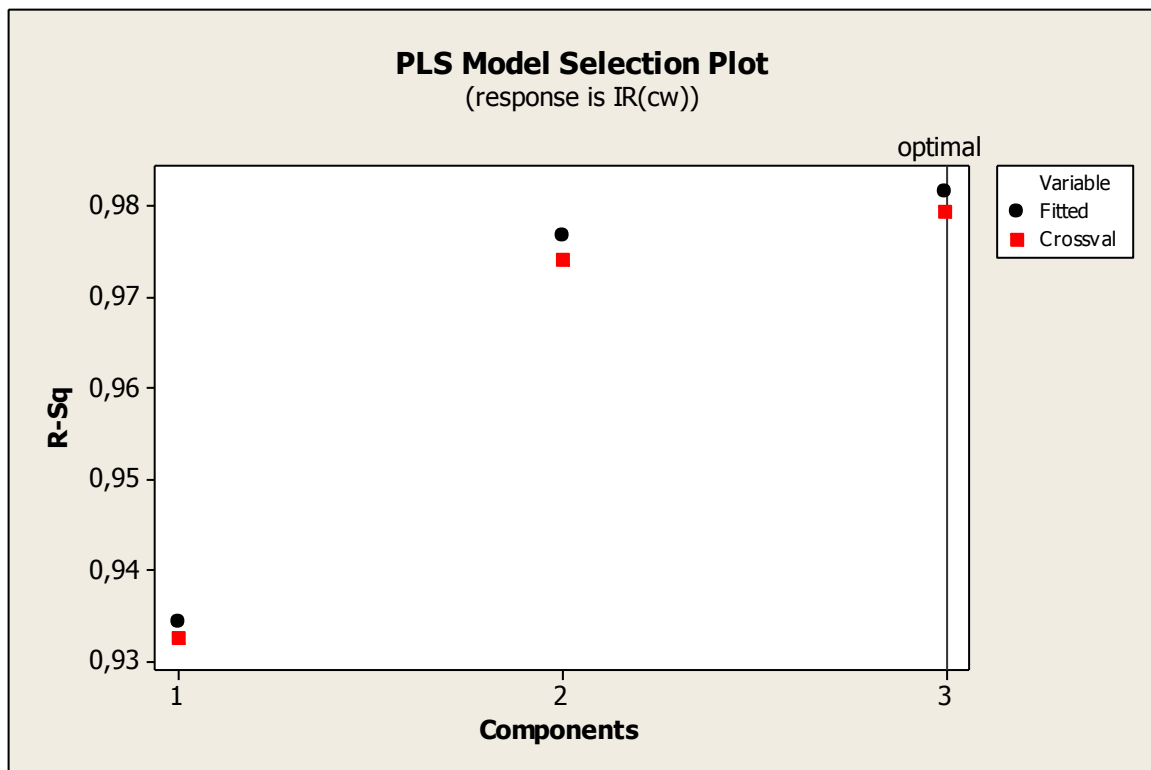


Figure 40 : Diagramme de sélection du modèle vers le nombre des composantes.

1-2-5-1-Analyse en composantes principales

Le tableau XVIII résume l'Analyse des valeurs propres de la matrice de corrélation pour IR (cw).

Tableau XVIII: Analyse des valeurs propres de la matrice de corrélation.

Eigenvalue	1,9309	0,9971	0,072
Proportion	0,644	0,332	0,024
Cumulative	0,644	0,976	1

D'après ces résultats (tableau XVIII) et le Diagramme des valeurs propres et les proportions (Figure 41):

La 1^{ère} composante principale avec une proportion de 0,644 explique 64,4 % de la variabilité des données ; par conséquent, cette composante doit être incluse et la 3^{ème}

composante a une proportion de 0,024 ceci explique uniquement que 2,4 % de la variabilité des données. Cette composante n'est sans doute pas suffisamment importante pour être incluse.

Les composantes principales n'expliquent que 97.6 % de la variance. On réalise d'autres analyses sur les données lorsque les composantes principales expliquent au moins 100 % de la variance.

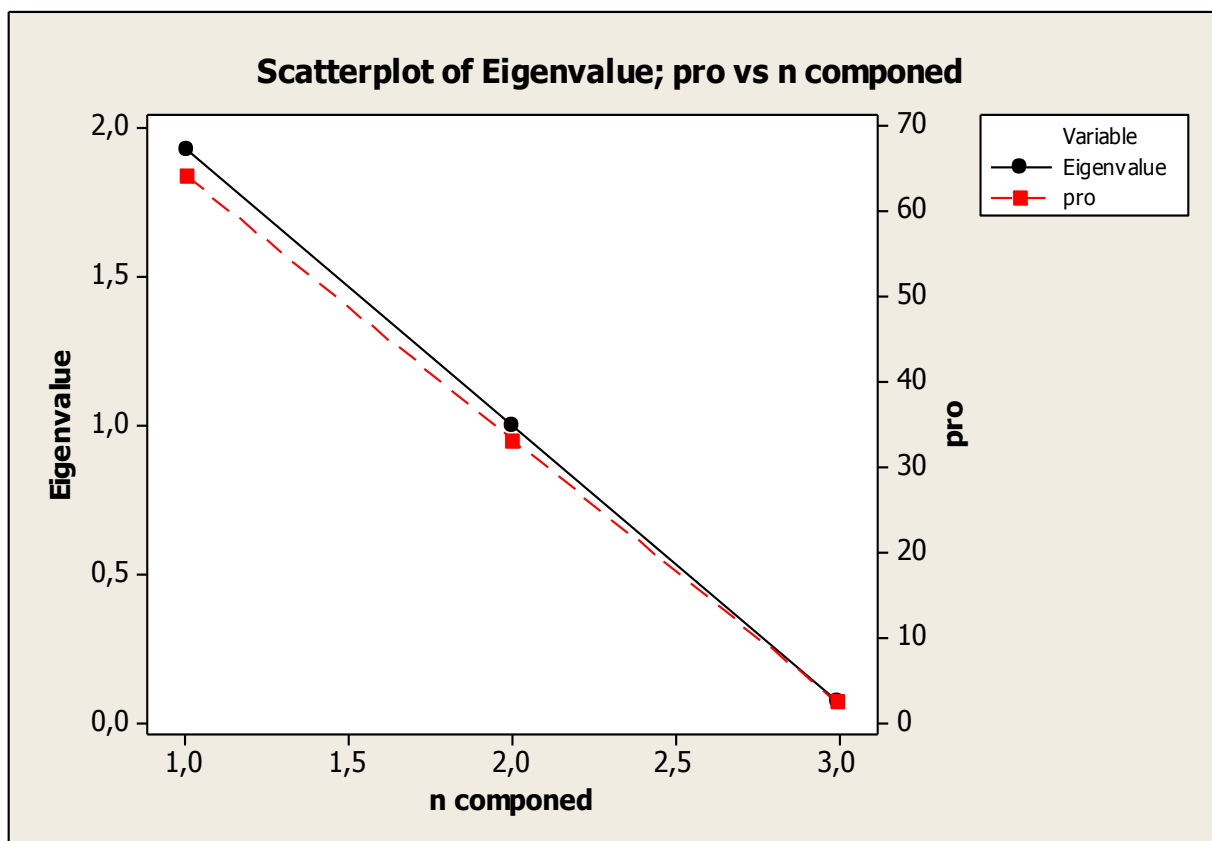


Figure 41 : Diagramme en cône des valeurs propres et des proportions vers le nombre de composantes.

Le tableau XIX résume les Analyses des vecteurs propres de la matrice de corrélation pour IR (cw).

Tableau XIX : Analyse des vecteurs propres de la matrice de corrélation.

Variable	PC1	PC2	PC3
RDCHI	0,706	0,019	-0,708
GATS1p	0,072	-0,996	0,045
Mor02m	0,704	0,083	0,705

1-2-5-2-Composantes principales (CP)

Les composantes principales sont les combinaisons linéaires des variables d'origine qui rendent compte de la variance des données. Le nombre maximal de composantes extraites est toujours égal au nombre de variables. Les vecteurs propres, constitués de coefficients correspondant à chaque variable, sont utilisés pour calculer les scores des composantes principales. Les coefficients indiquent la pondération relative de chaque variable dans la composante.

D'après ces résultats (Tableau XIX), la première composante principale présente une forte association positive avec les descripteurs Mor02m et RDCHI. La composante est principalement une mesure de la relation entre la dispersion théorique et la connectivité avec l'indice de rétention. La deuxième composante présente une forte association négative avec GATS1p, et mesure principalement l'autocorrélation avec l'indice de rétention. La troisième composante ; présente une forte association positive avec Mor02m ; et elle présente une forte association négative avec RDCHI et mesure principalement la relation d'équilibre entre la dispersion théorique et la connectivité avec l'indice de rétention.

D'après ces résultats, le score pour la première composante principale peut être calculé à partir des données normalisées, à l'aide des coefficients fournis sous PC1, PC2, PC3 :

$$PC1 = 0.706 RDCHI + 0.072 GATS1p + 0.704 Mor02m \quad (59)$$

$$PC2 = 0.019 RDCHI - 0.996 GATS1p + 0.083 Mor02m \quad (60)$$

$$PC3 = -0.708 RDCHI + 0.045 GATS1p + 0.705 Mor02m \quad (61)$$

1-2-5-3-Diagramme des contributions

Le diagramme des contributions (Figure 42) représente deux prédicteurs fortement corrélés car les angles entre les lignes (Mor02m RDCHI) sont faibles et, l'autre faiblement corrélés car l'angle entre la ligne (GATS1p) est grands . Les lignes sont d'une taille équivalente, ce qui indique que les prédicteurs sont d'importance égale. Sur la première composante, les deux premiers prédicteurs ont des contributions absolues plus grandes que l'autre. Sur la seconde composante, les prédicteurs possèdent des contributions négatives similaires, indiquant qu'ils sont d'importance égale.

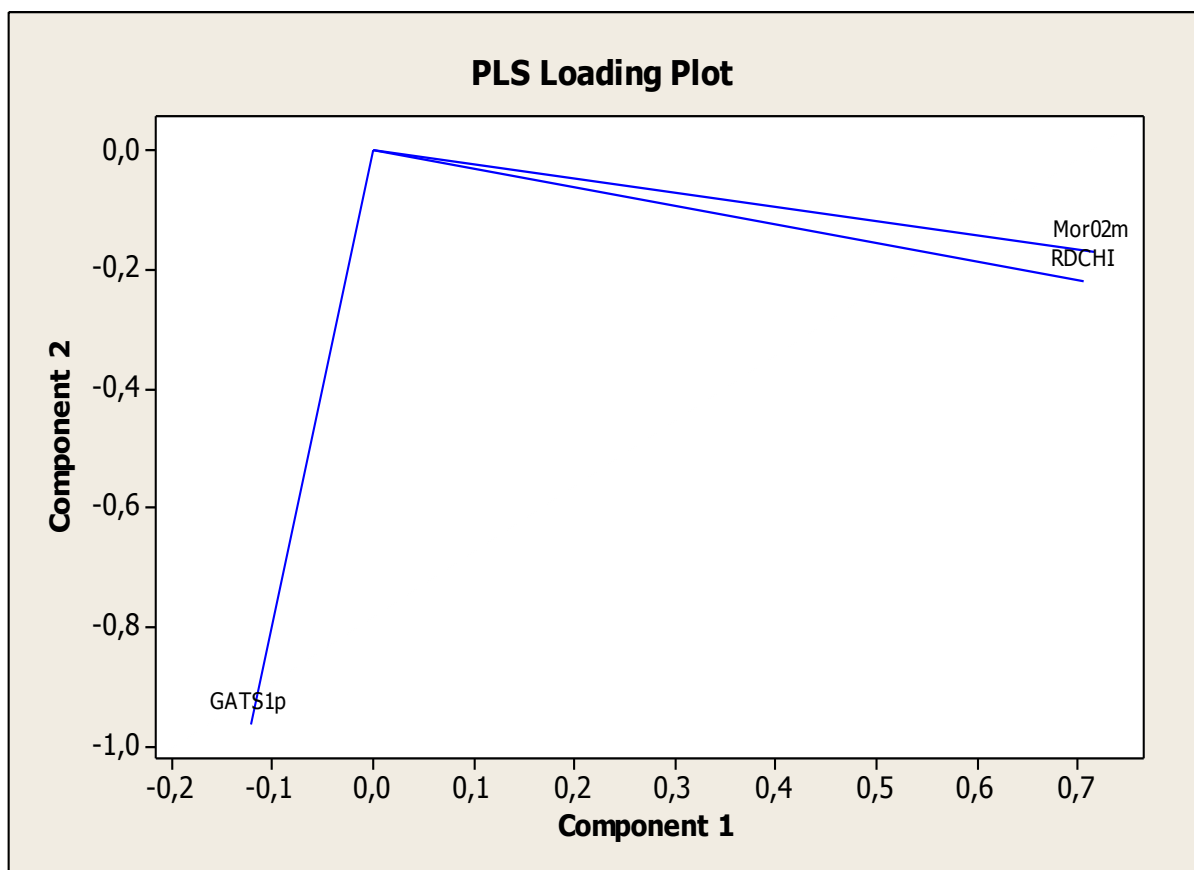


Figure 42: Diagramme des contributions de corrélation pour les descripteurs.

1-2-5-4-Le diagramme de coefficients normalisés :

Le graphique ci-dessous (Figure 43) correspond aux coefficients normalisés pour le modèle avec 2 composantes qui sont significatifs et la troisième composante qui n'est pas significative on la supprime.

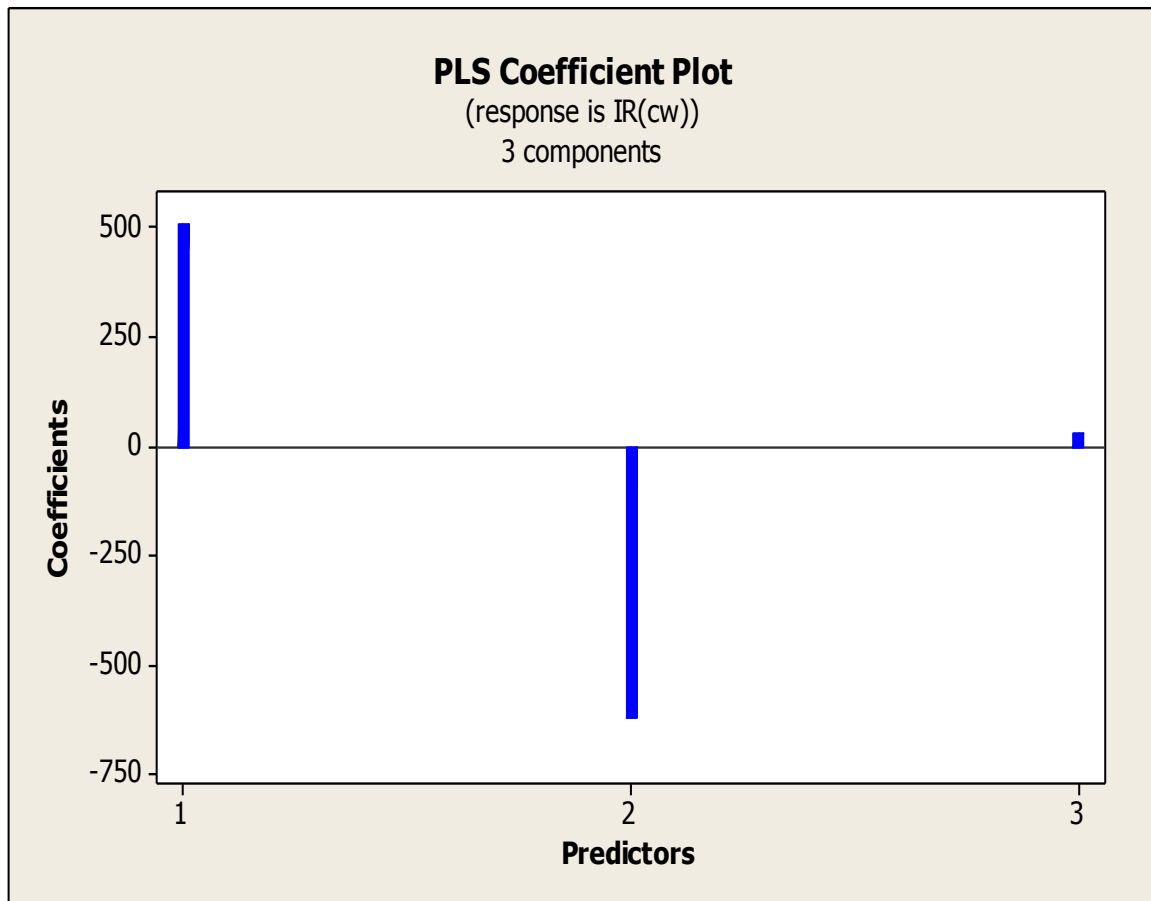


Figure 43 : Diagramme de coefficient vers le nombre de composants.

1-2-5-5-Le diagramme des réponses :

Dans ce diagramme (Figure 44), les points suivent généralement un schéma linéaire, indiquant que le modèle est correctement ajusté aux données. Les points figurant dans le diagramme des valeurs résiduelles en fonction de l'effet de levier ci-dessus ne semblent pas poser problème dans ce diagramme.

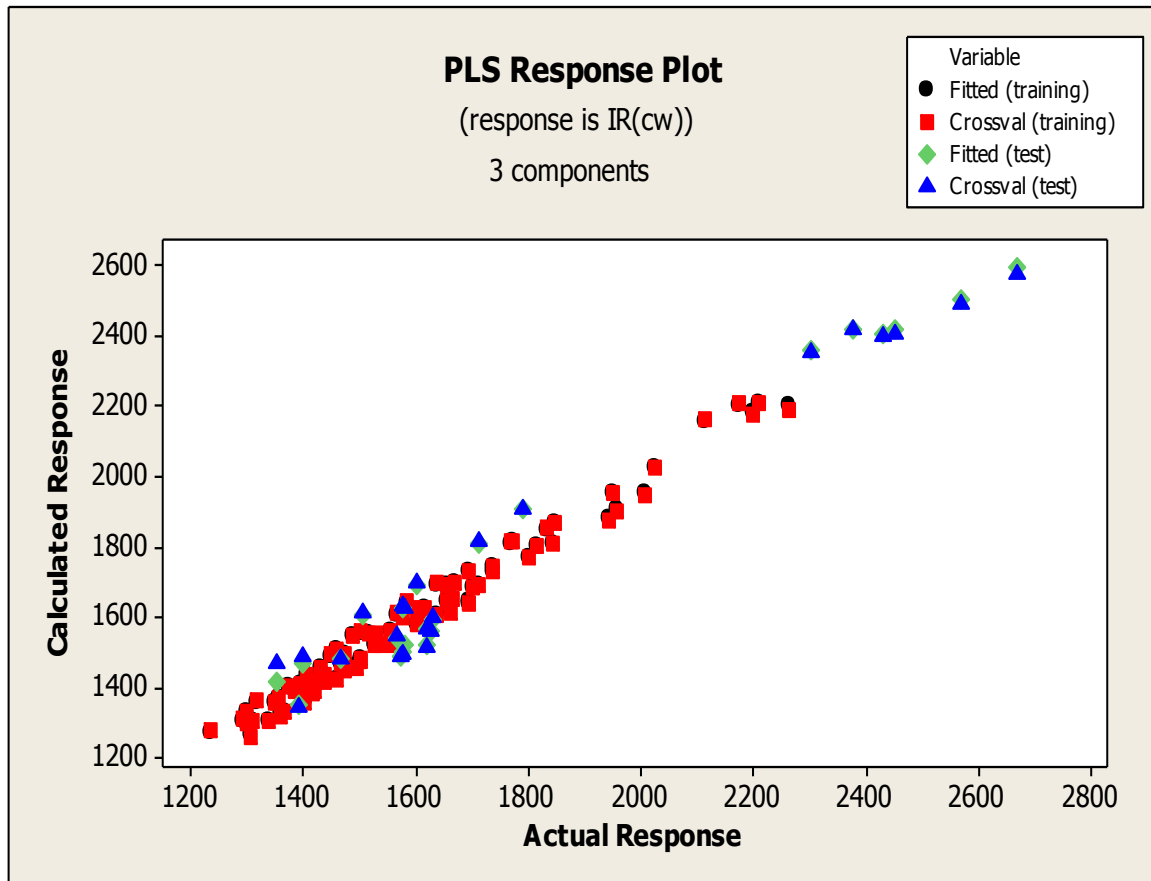


Figure 44 : Diagramme des réponses pour les valeurs calculées vers les valeurs observées.

1-2-5-6-Diagrammes de valeurs aberrantes

Dans ce diagramme (Figure 45) les deux ensembles montre:

- dans le cas de la Calibration : l'observation 46 (2-méthyl-6-(2-méthylbutyl)-3-octylpyrazine) peut être considéré comme une valeur aberrante, car elle se situe en dessous de ligne de référence horizontale et 15 points peuvent être considérés comme de points d'effet de levier car ils sont situés à droite de la ligne verticale ($h_{ii} > 0.1$).

- dans le cas de la Validation : 9 points peuvent être considérés comme de points d'effet de levier, car ils sont situés à droite de la ligne verticale ($h_{ii} > 0.071$).

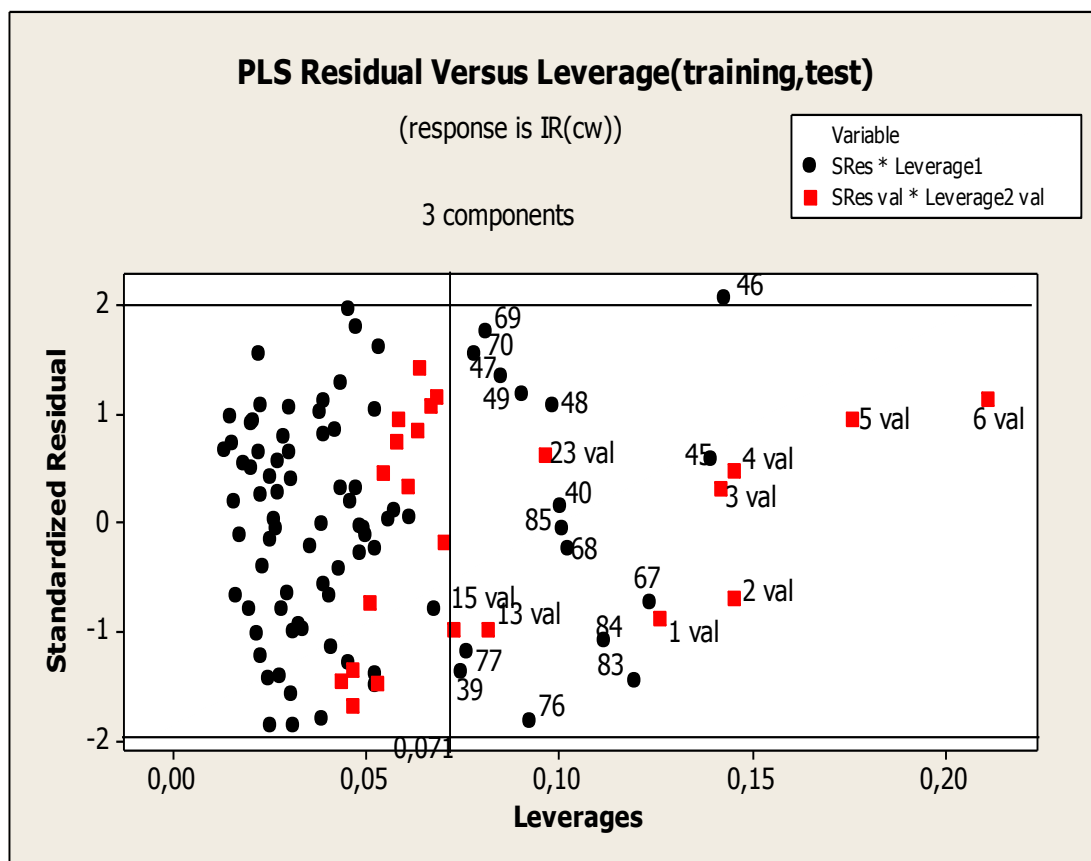


Figure 45: Diagramme des valeurs Résiduelles vers les valeurs de levier.

Après la mise en place du modèle avec la régression PLS, la régression MLR a été utilisée pour développer un modèle sur les composés de l'ensemble de calibration, sur la base de deux composants après l'élimination du troisième composant qui n'est pas significatif.

Le modèle basé sur ces descripteurs avec MLR donne l'équation suivante :

$$Y=1159.74+63,022*PC1+561.43*PC2 \quad (62)$$

$$S = 62, 2583 \quad R\text{-Sq} = 93, 2\% \quad R\text{-Sq}(\text{adj}) = 93,0\% \quad F = 323,54, \quad P=0,000$$

La valeur de $R^2 = 93.2\%$ montre la qualité de l'ajustement, alors que la valeur très élevée du rapport de la variance expliquée par le modèle à la variance résiduelle ($F = 323.54; p = 0,000$) montre que le modèle permet une très bonne prédiction des $n (=85)$ valeurs de IR de l'ensemble de calibration, (erreur standard $s= 62.25$).

Le tableau XX résume les résultats des caractéristiques des deux composants sélectionnés par l'estimation MLR

Tableau XX : Évaluations de l'estimation MLR pour le modèle

Predictor	Coef	SE Coef	T	P	VIF
Constant	1159,74	70,47	16,46	0,000	
pc 1	63,022	6,033	10,45	0,000	2,359
pc 2	561,43	43,82	12,81	0,000	2,359

- Les valeurs des probabilités de T pour les deux composants sont nulles, ceci indique qu'ils sont hautement significatifs avec un risque d'erreur de première espèce $\alpha=0.05$ les paramètres sont tous significatifs parce que. Leurs estimations sont de l'ordre : $\beta_0 = 1159.74, \beta_1 = 63.022$ et $\beta_2 = 561.43$.

1-Test de la multicolinéarité :

-Les valeurs des facteurs d'inflation de la variance (FIV) pour les deux descripteurs sont $> 1 < 5$ (Tableau XX et figure 46), ceci indique qu'ils ne sont pas corrélés donc le problème de multicolinéarité n'existe pas.

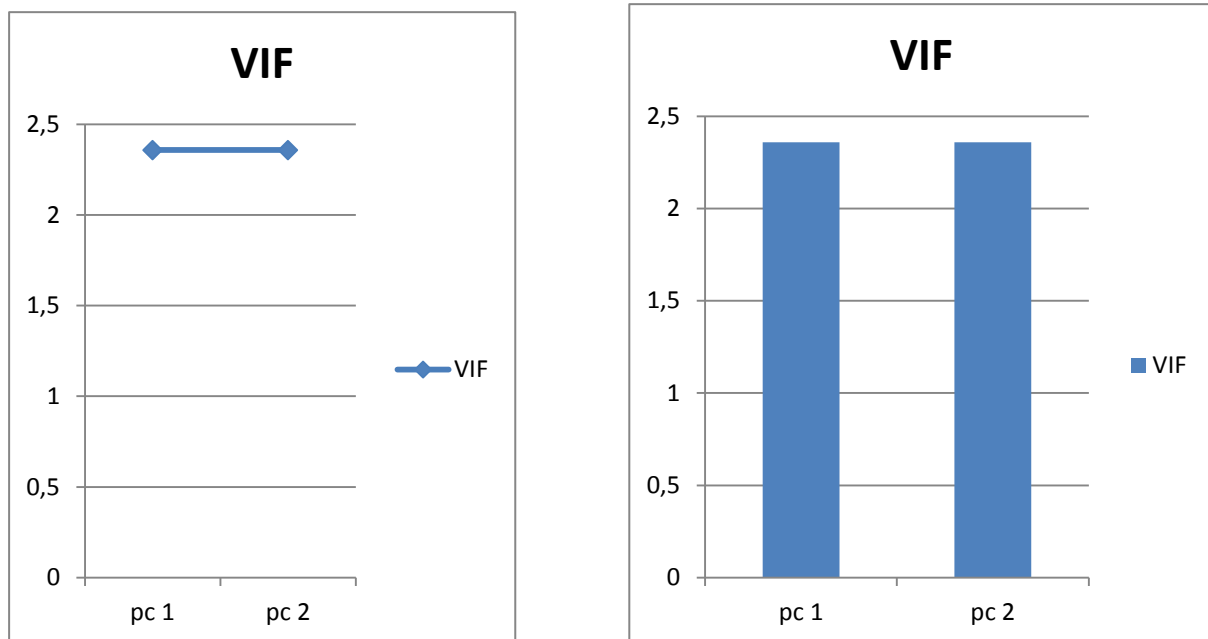


Figure 46: Digramme de VIF pour chaque descripteur.

1-2-7-La qualité de l'ajustement :

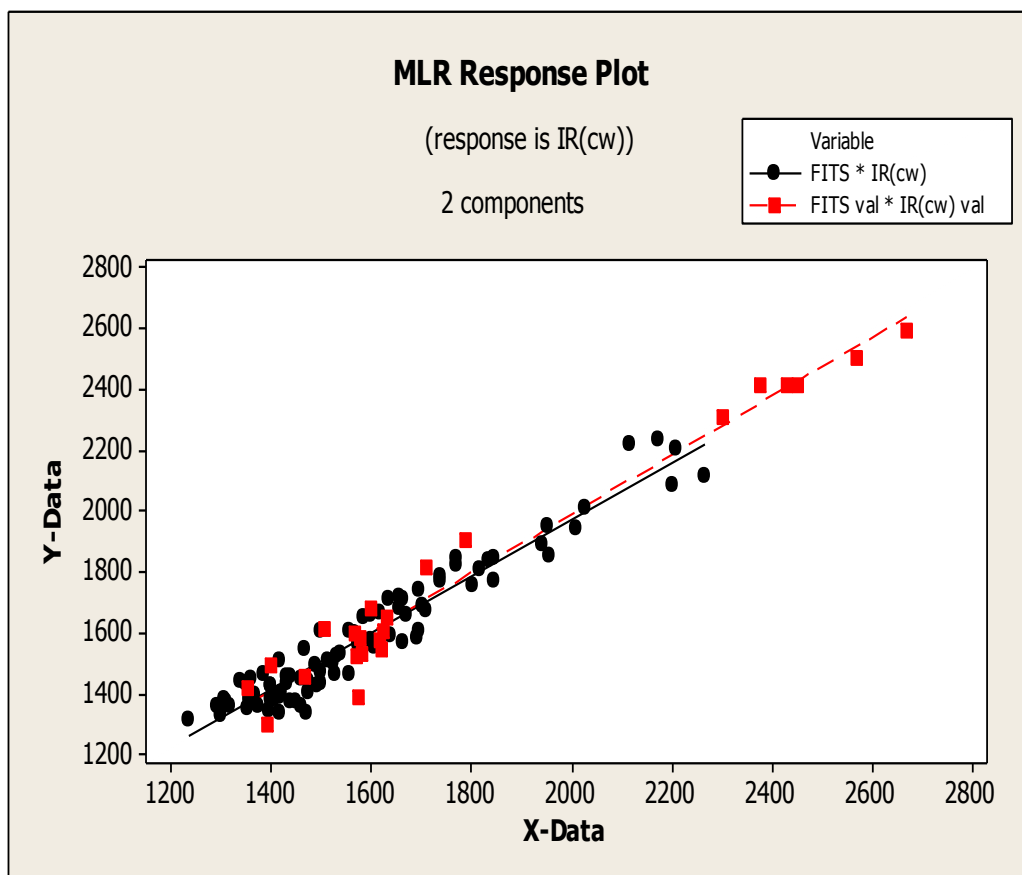


Figure 47 : Valeurs estimées en fonctions des valeurs observées.

On remarque, relativement que les points suivent généralement un schéma linéaire, indiquant que le modèle est correctement ajusté aux données, Ce qui prouve l'approche et la proportionnalité dans les deux ensembles (Calibration, Validation) (Figure 49).

1-2-8--domaine d'application :

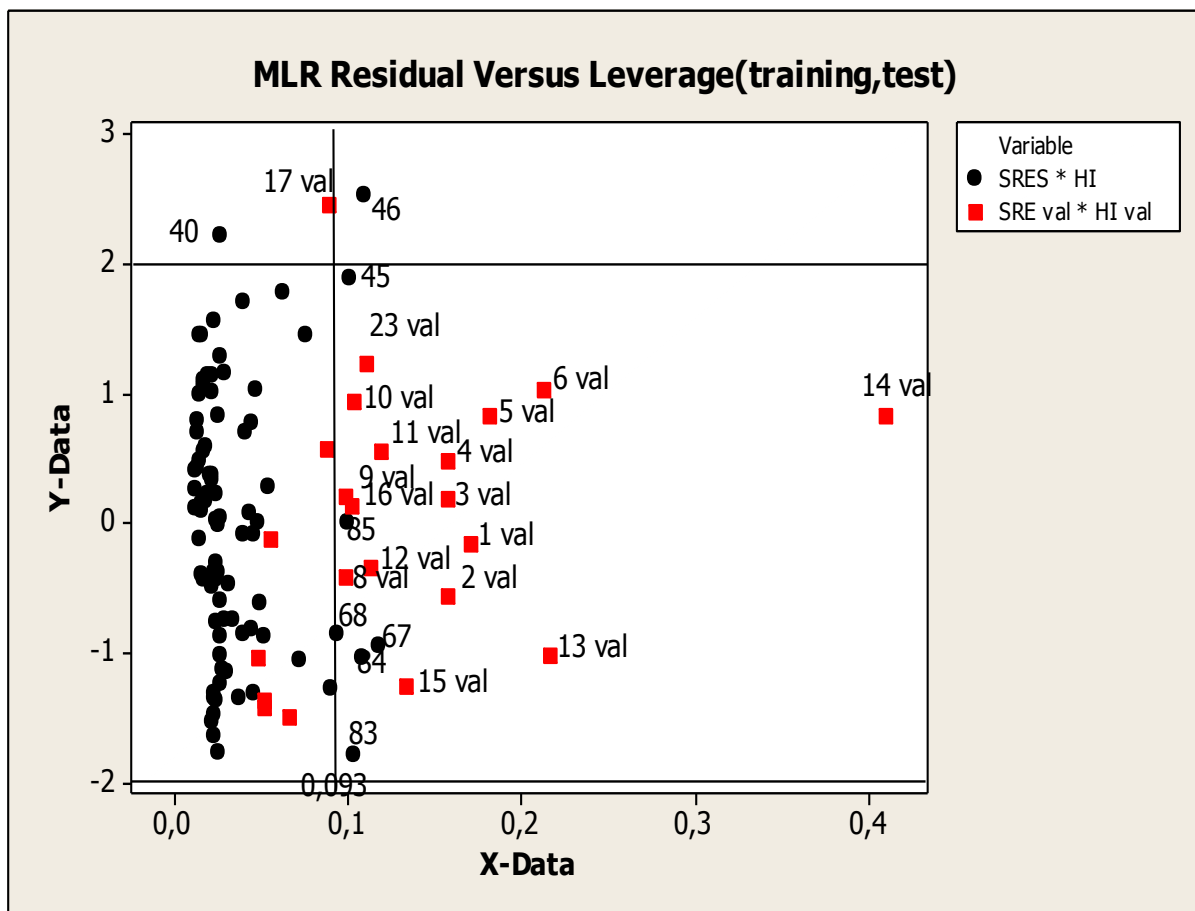


Figure 48: Diagramme de Williams des valeurs Résidueles vers les valeurs de levier.

Dans ce diagramme (Figure 48) l'analyse des résidus dans les deux ensembles montre que :

- dans le cas de la Calibration : six points peuvent être considérés comme des points d'effet de levier (45, 67, 68, 69, 83, 84,85), car ils sont situés à droite de la ligne verticale ($h_{ii} > 0.093$).

Deux points peuvent être considérés comme de valeurs aberrants :

-40 : 1-methylbutyl)pyrazine

-46 : 2-methyl-6-(2-methylbutyl)-3-octylpyrazine,.

Car ils se situent en dessous de la ligne de référence horizontale supérieure à l'intervalle ± 2 ainsi que le point 46, situés à droite de la ligne verticale ($h_{ii} > 0.093$) ceci influe sur l'ajustement du modèle.

- dans le cas de la validation : 16 points peuvent être considérés comme des points d'effet de levier, car ils sont situés à droite de la ligne verticale ($h_{ii} > 0.093$), un point peut être considéré comme une valeur aberrante 17 (2,5-diméthyl-3-propylpyrazine), car il se situe en dessous de la ligne de référence horizontale supérieure à l'intervalle ± 2 .



REFERENCES BIBLIOGRAPHIQUES

- [1] Stanton, D.T., Jurs, P.C. 1989. Computer-assisted prediction of gas chromatographic retention Indexes of pyrazines. *Anal. Chem.*, 61: 1328-1332.
- [2] Hansch, C., Fujita, T. 1964. ρ - σ - π analysis. A method for the correlation of biological activity and chemical structure.. American Chemistry Society., 86: 1616-1662.
- [3] Todeschini R, Consonni V, Pavan, M .2006. Dragon Software for the Calculation of Molecular Descriptors, Release 5.3 for windows, Milano.
- [4] Moby Digs 1.1, <http://www.disat.unimib.it>.
- [5] Minitab, Release 16.1, Statistical Software, 2003.
- [6] Dragon 5.4, <http://www.disat.unimib.it>
- [7] MATLAB, Version 7,0,0,19920 (Release 14), The Language of Technical Computing, The Math Works, Inc, May 06 (2004).
- [8] Todeschini R, Consonni V, Pavan M.2005, DRAGON, Software for the Calculation of Molecular Descriptors, Release 5.3 for windows, Milano.
- [9] Berlin, G .B. The Pyrazine; Wiley-Interscience: New York, 1982.
- [10] Imen Touhami, Karima Mokrani et Djelloul Messadi .2012. Modèles QSRR Hybrides Algorithme Génétique-Régression Linéaire Multiple des Indices de Rétention de Pyrazines en Chromathographie Gazeuse. *Lebanese Science Journal*, Vol. 13, No. 1.
- [11] Parliment, T.H., Epstein, M.F. 1973. Organoleptic properties of some alkyl-substituted Alkoxy- and alkylthiopyrazines. *J. Agric. Food Chem.* 21: 714-716.
- [12] Kaliszan, R. 1986. Quantitative relationships between molecular structure and Chromatographic retention. *CRC Crit. Rev. Anal. Chem.*, 16: 323-383.

- [13] Kaliszan, R. 1987. Quantitative structure-chromatographic retention relationships. J. Wiley, New York.
- [14] Pynnönen, Seppo and Timo Salmi .1994. A Report on Least Absolute Deviation Regression with Ordinary Linear Programming. Finnish Journal of Business Economics 43:1, 33-49.
- [15] Tiffany Machabert .2014 .Modèles en très grande dimension avec des outliers. Théorie, simulations, applications" paris
- [16] Dodge Y ,Rousson V .2004 .Analyses de régression appliquée.Paris
- [17] By Kani Chen, Zhiliang,Hong Zhang,and Lincheng Zhao..Analysis of least absolute deviation.
1. [18] Faria, S. and Melfi, G. 2006. Lad regression and nonparametric methods for detecting outliers and leverage points. Student, 5 :265– 272.
- [19] Gabriela Ciuperca.2009.Estimation robuste dans un modèle paramétrique avec rupture. Bordeaux .
- [20] Gilbert Saporta .2012. Régression robuste .
- [21] Ndèye Niang- Gilbert Saporta .2014.Régression robuste Régression non-paramétrique Mars .
- [22] Soumaya REKAIA. Indicateurs de la sensibilité de l'estimateur Least Absolute Deviation Assas Paris.
- [23] Dodge, Y. 2004. Statistique : Dictionnaire encyclopédique. Springer-Verlag France Paris.

- [24] Dodge, Y. and Jureckova, J. 2000. Adaptive Regression. Springer-Verlag New York.
- [25] Hyperchem 6.03, (Hypercube), <http://www.hyper.com>.
- [26] Kaliszan, R. 1987. Quantitative structure-chromatographic retention relationships. J. Wiley, New York.
- [27] Lee, Seung Ki., Polyakova, Yulia. Row, Kyung Ho. 2004. Evaluation of predictive retention factors for phenolic compounds with QSPR equations. *J. Liq. Chromatogr and Rel. Tech.*, 27(4): 629-639.
- [28] Levine, I.N. 2000. Quantum chemistry. 5th ed., New Jersey, Prentice-Hall.
- [29] Magnuson, V.R., Harriss, D.K., Basak, S.C. 1983. Topological indices based on neighbor
- [30] Symmetry: chemical and biological applications In: Chemical Applications of Topology and Graph Theory. R.B. King, ed., Elsevier, Amsterdam. 178-191.
- [31] Masuda, H., Misaku, Y., Shibamoto, T. 1981. Synthesis of new pyrazines for flavor use. *J. Agric. Food Chem.*, 29: 944-947.
- [32] Masuda, H., Mihara, S. 1986. Use of modified molecular connectivity indices to predict retention indices of monosubstituted alkyl, alkoxy, alkylthio, phenoxy and (phenylthio) pyrazines. *J. Chromatogr.*, 366: 373-377.
- [33] Mihara, S., Enomoto, N. 1985. Calculation of retention indices of pyrazines on the basis of molecular structure. *J. Chromatogr.*, 324: 428-430.
- [34] Mihara, S., Masuda, H. 1987. Correlation between molecular structures and retention indices of pyrazines. *J. Chromatogr.*, 402:309-317.

- [35] Buchbauer, G. 2000. Threshold-based structure-activity relationships of pyrazines with bellpepper Flavor.
- [36] Leardi, R., Boggia, R. and Terrile, M. 1992. Genetic algorithms as a strategy for feature selection, *J Chemometrics*, Vol. 6, pp. 267-281.
- [37] Todeschini, R. 1997. Data correlation, number of significant principal components and shape of molecules. The K correlation index. *Anal. Chim. Acta*, 348: 419-430.
- [38] Dewar M J S, Zoebisch E G, Ealy E F, Stewart J J P, 1985. AMI: A New General Purpose Quantum Mechanical Model . *J Am, Chem, Soc*, Vol, 107, pp, 3902-3909.
- [39] Holder, A.J. 1998. AM1, *Encyclopedia of Computational Chemistry*, P.V.R. Scheleyer, N.L. Allinger, T. Clarck, J. Gasteiger, P.A. Kollman, H.F Schaefer, III and P.R.Schreiner (Eds), Wiley, Chichester, 1, 8.
- [40] Durbin, J., Watson, G.S. 1951. Testing for serial correlation in least squares regression. II. *Biometrika*, 38(1-2): 159-178.
- [41] Umaporn Chantasorn.2011. Efficiency Comparisons of Normality Test Using Statistical Packages. *J. Sc. Tech.*, Vol. 16, No. 3.
- [42] Nornadiah Mohd Razali ,Yab Bee Yah .2011. Power Comparaisons of shapiro-wilk,Kolmogorov-smornov,lillieffors and Anderson-Darling tests,journal of statistique Modelling and analytics .vol 2 No 1:21-33.
- [43] Gilbert Colletaz. 2004 *Statistique non paramétrique élémentaire* .Université d'Orléans.
- [44] Ricco Rakotomalala. 2011. Tests de normalité Techniques empiriques et tests statistiques. Université Lumière Lyon 2.
- [45] Cludio araujo.2013 .Micro économétrie stratégie d'estimation méthode de base.

- [46] Damodar N. Gujarati, Dawn C. Porter.2009.Basic Econometrics Fifth Edition.
- [47] J. Paul Tsasa Vangu.2011.Econometrie.German.
- [48] Oya Can Mutan , Birdal Şenoğlu.2009.A Monte Carlo Comparison of Regression Estimators When the Error Distribution is Long-Tailed Symmetric. Journal of Modern Applied Statistical Methods.Volume 8 | Issue 1.
- [49] Tron Foss, Ingunn Myrtveit, Erik Stensrud. Yinbo Li, Gonzalo R. Arce.2004. AMaximum Likelihood Approach to Least Absolute Deviation Regression. Journal on Applied Signal Processing, 12, 1762–1769.
- [50] Richard J. Butler,James B.McDonald,Ray.Nelson,Steven B.White .1990.Robust and Partially Adaptive Estimation of Regression Models.The Review of Economics and Statistics.Volume 72 ,Issue 2,321-327.
- [51] Jolliffe I T, (1986) "Principal Component Analysis", Springer- Verlag, Berlin.
- [52] Escofier B, Pages J, (1998),"Analyses factorielles simples et multiples: Objectifs, méthodes et interprétation", 3ème ed,Dunod, Paris.
- [53] Escofier B, Pages J..Analyses factorielles simples et multiples: Objectifs, méthodes et interprétation", 3ème ed,Dunod, Paris.
- [54] Gauchi J P, (1995) "Utilisation de la régression PLS pour l'analyse des plans d'expériences en chimie de formulation", *Rev, Stat, Appl*, Vol, 43, pp, 65-89.
- [55] Gelada P, Kowalski B, R, (1986) " Partial least- squares regression: tutorial", *Anal,Chim,Acta*, Vol, 185, pp, 1- 17.
- [56] Vancolen, S.2004. la régression PLS, groupe de statistique, université de Neuchâtel, Suisse.

- [57] Tenenhaus M.1998. la régression PLS, théorie et pratique Paris : Technip.
- [58] Pagès J, Tenenhaus M.2001. Multiple factor analysis combined with PLS path modeling, applications to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgments., *Chemometrics and Intelligent Laboratory Systems*, Vol, 58, pp, 261- 273.
- [59] Gauchi J P.1995. "Utilisation de la régression PLS pour l'analyse des plans d'expériences en chimie de formulation", *Rev, Stat, Appl*, Vol, 43, pp, 65-89.
- [60] Mesquita D P O, Dias A M A, Dias A L, Amaral E C, Ferreira .2009. "Correlation between sludge settling ability and image analysis information using partial least squares", *Anal, Chim,Acta*, Vol, 642, pp, 94-101.
- [61] Wold S, Ruhe A, Wold H, Dunn W.1984. "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses", *SIAMJ, Sci, Stat,Comput*, Vol, 5, pp, 735- 743.
- [62] Tenenhaus M.1998. la régression PLS, théorie et pratique Paris : Technip.



CONCLUSION GENERALE

CONCLUSION GENERALE

Les pyrazines ; molécules présentes de façon naturelle dans notre environnement, ont un intérêt dans de multiples domaines, notamment dans l'alimentation. L'un des intérêts des chercheurs pour ces molécules vient de leur pouvoir odorant.

La modélisation des indices de rétention de 114 pyrazines élués sur deux colonnes différents OV -101 par deux méthodes MLR et LAD et la colonne CRW-20M par trois méthodes MLR, LAD et PLS.

Sur La colonne OV-101 :

Le problème des observations aberrantes pour le model à 3 descripteurs nous conduit à :

-une comparaison des équations des hyperplans pour les 2 techniques MLR, LAD

-Une Comparaison graphique

On a constaté que la méthode LAD est plus efficace pour cette séparation chromatographique (stabilité, robustesse) par rapport à la méthode des moindres carrées avec un minimum de valeurs aberrantes.

L'application du test d'Anderson-Darling de la compatibilité d'une distribution avec la loi normale à l'aide de la valeur de p. Est acceptable pour $p > 0.1$.

On remarque que toutes les valeurs observées sur tous les tests (statistique et graphique) se rapprochent pour les deux méthodes dans l'état de l'approche.

Sur la colonne CRW-20M :

Le problème des observations aberrantes pour le model a 3 descripteurs nous conduit à :

-une comparaison des équations des hyperplans pour les 2 techniques MLR, LAD

-Une Comparaison graphique

On a constaté que la méthode LAD est plus efficace pour cette séparation chromatographique (stabilité, robustesse) par rapport à la méthode des moindres carrées avec un minimum de valeurs aberrante.

L'application du test d'Anderson-Darling de la compatibilité d'une distribution avec la loi normale à l'aide de la valeur de p. Est acceptable pour $p > 0.1$.

On remarque que toutes les valeurs observées sur tous les tests (statistique et graphique) se rapprochent pour les deux méthodes.

CONCLUSION GENERALE

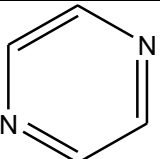
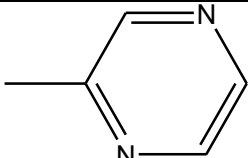
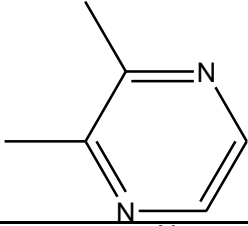
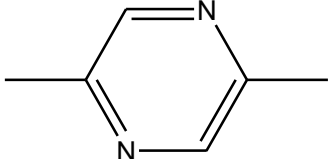
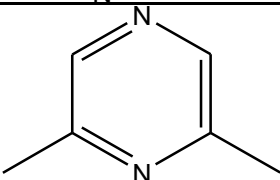
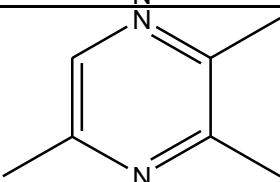
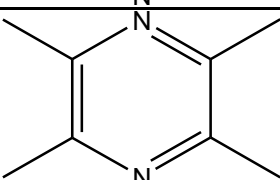
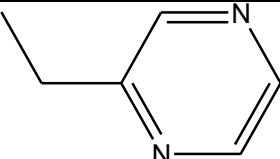
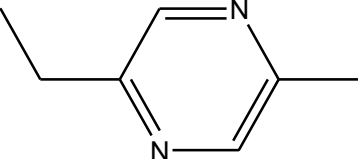
Enfin on utilise la Méthode PLS après l'élimination des points aberrants à cause de la détection du problème de la multicollinéarité à pourcentage faible.

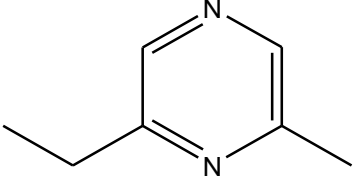
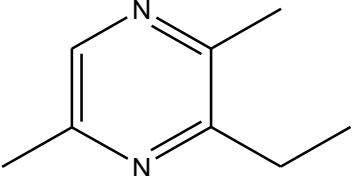
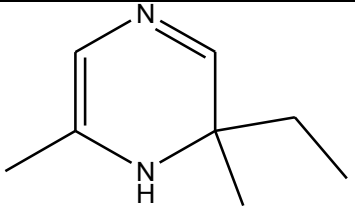
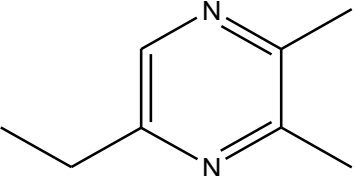
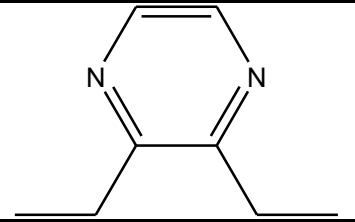
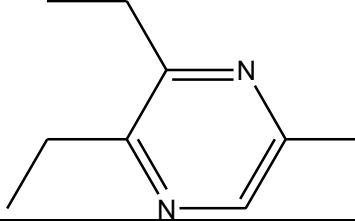
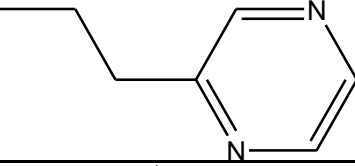
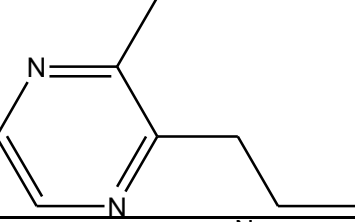
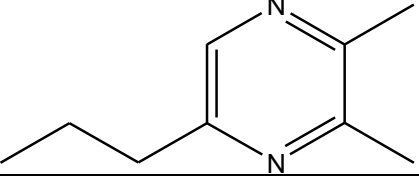
Pour cette étude la colonne CRW-20M est la plus adaptée pour la modélisation des indices de rétention des composés pyrazines.

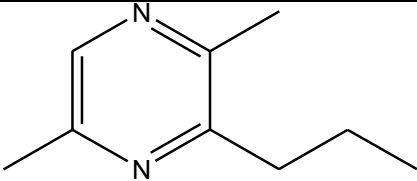
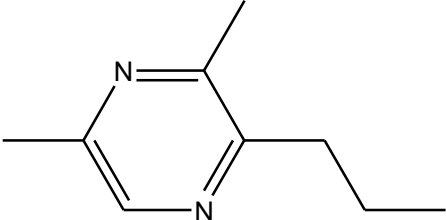
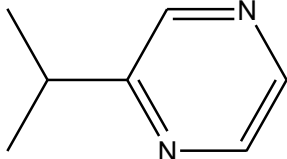
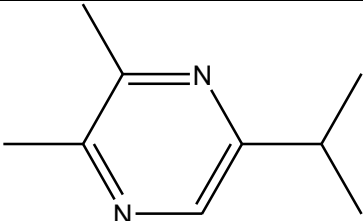
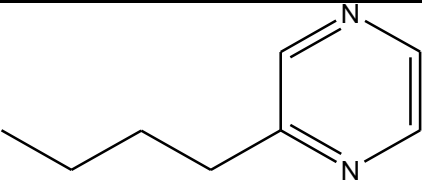
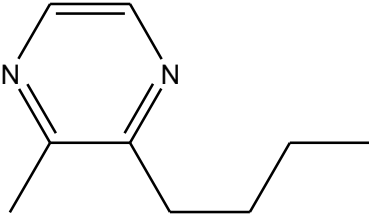
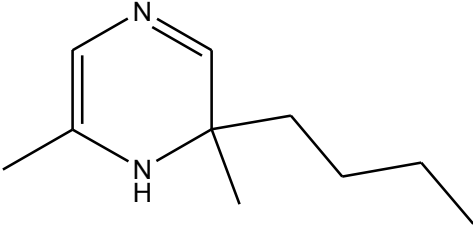
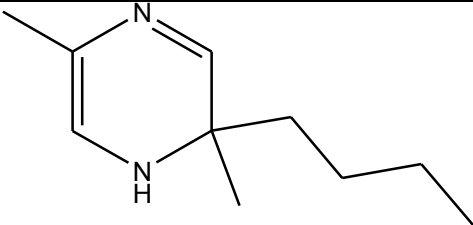
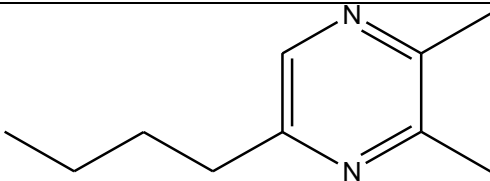


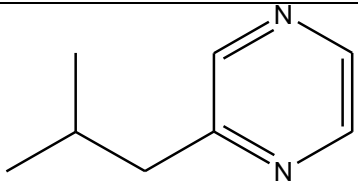
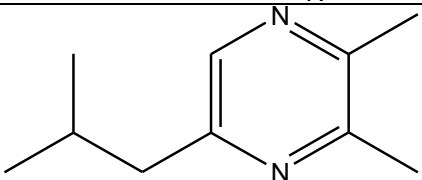
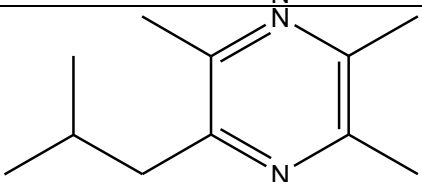
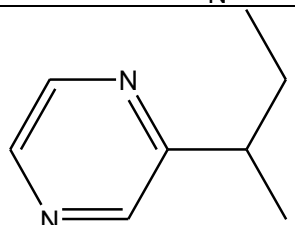
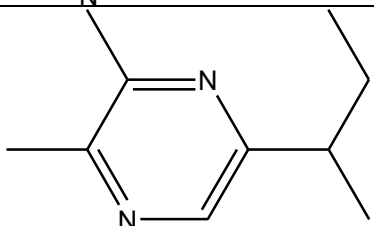
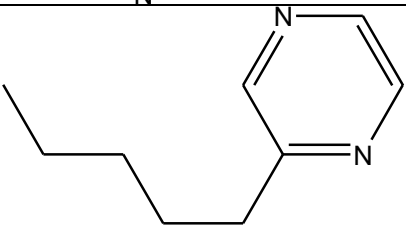
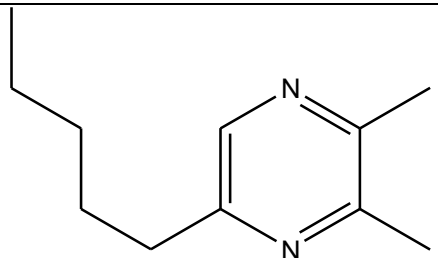
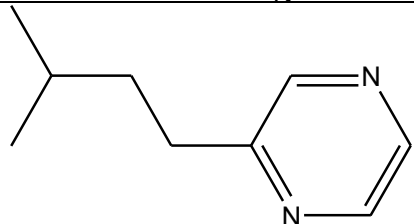
***ANNEXE* : Présentation des données.**

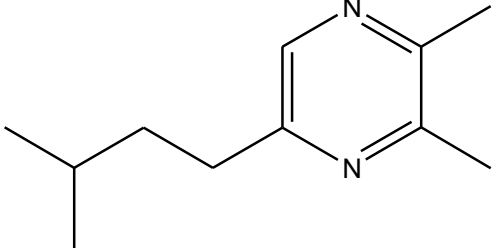
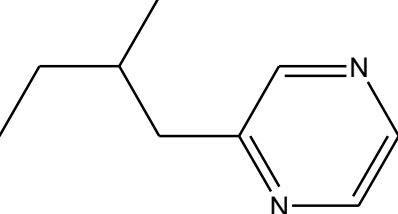
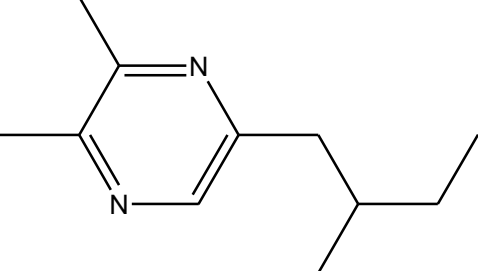
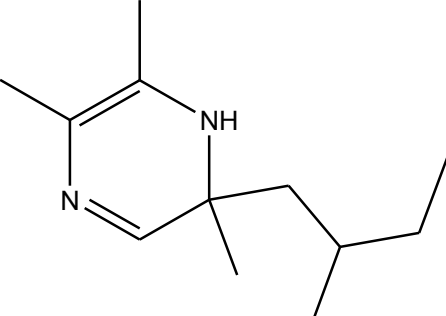
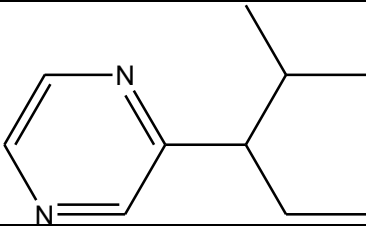
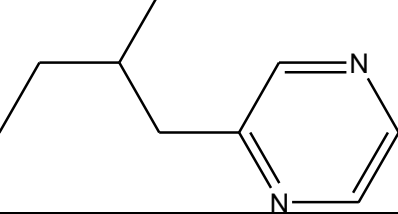
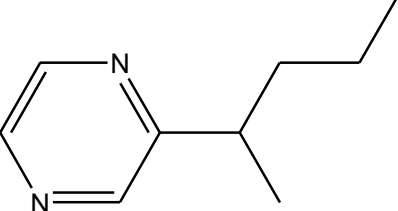
Tableau I –l'indice de retention de Pyrazine Pour OV-101 et CRW-20M.

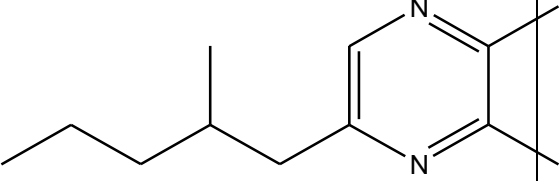
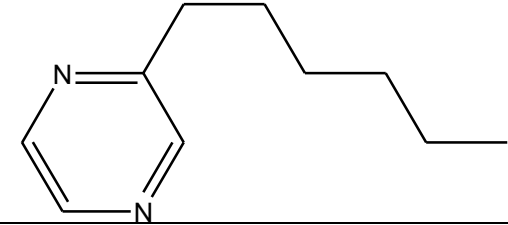
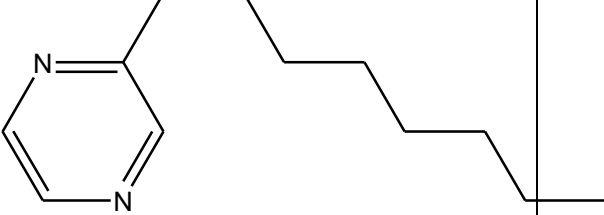
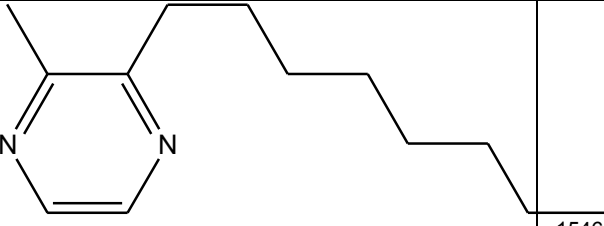
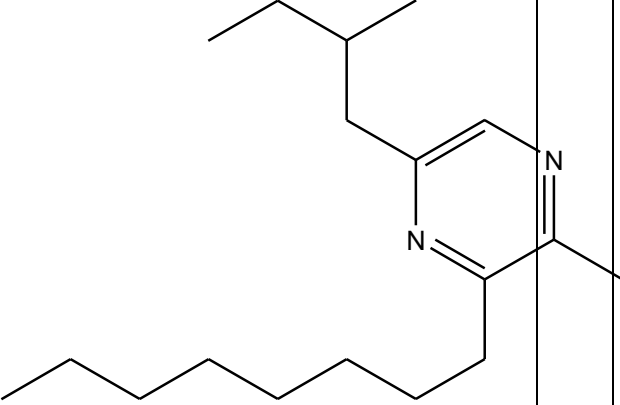
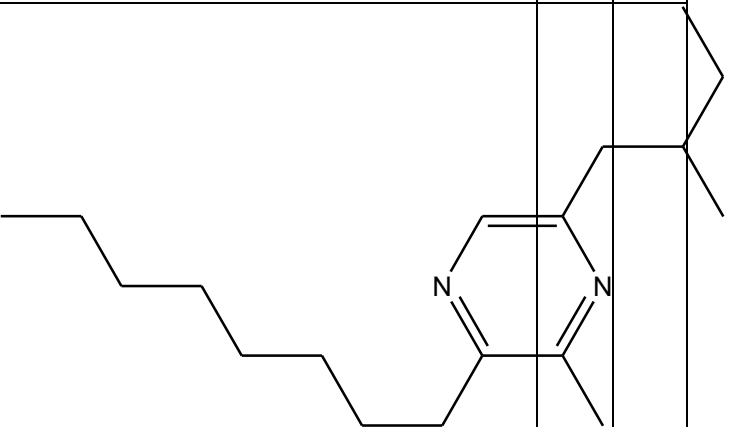
n°	Compounds	structure	ov-101	IR(cw)
1	Pyrazine		710	1179
2	Methylpyrazine		801	1235
3	2,3-dimethylpyrazine		897	1309
4	2,5-dimethylpyrazine		889	1290
5	2,6-dimethylpyrazine		889	1300
6	Trimethylpyrazine		981	1365
7	Tetramethylpyrazine		1067	1439
8	Ethylpyrazine		894	1300
9	2-ethyl-5-methylpyrazine		980	1357

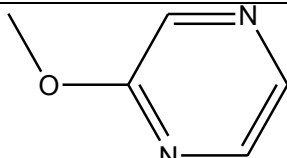
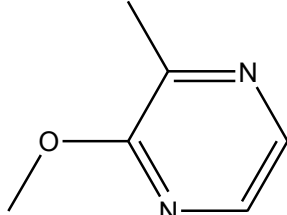
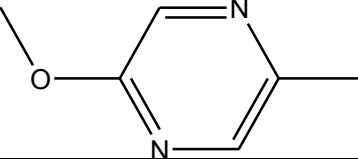
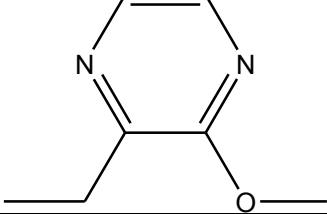
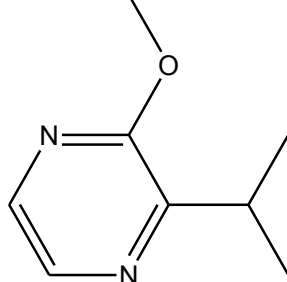
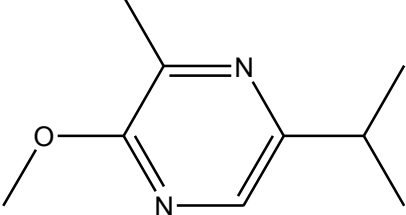
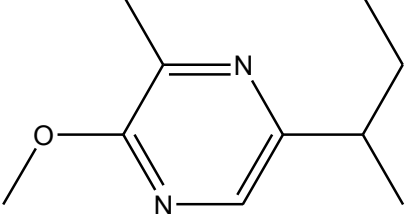
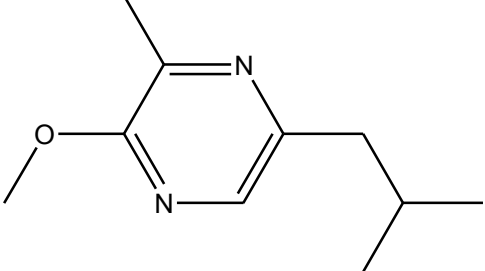
10	2-ethyl-6-methylpyrazine		977	1353
11	2,5-dimethyl-3-ethylpyrazine		1059	1400
12	2,6-dimethyl-6-ethylpyrazine		1064	1415
13	2,3-dimethyl-5-ethylpyrazine		1066	1421
14	2,3-diethylpyrazine		1065	1417
15	2,3-diethyl-5-methylpyrazine		1137	1459
16	Propylpyrazine		986	1374
17	2-methyl-3-propylpyrazine		1072	1438
18	2,3-dimethyl-5-propylpyrazine		1154	1500

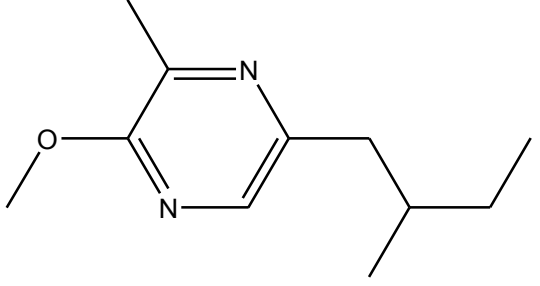
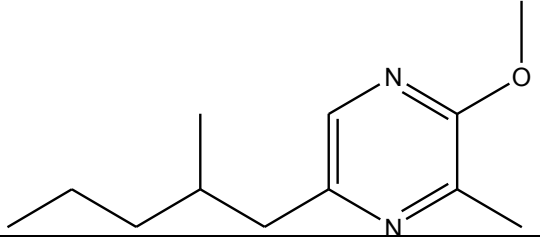
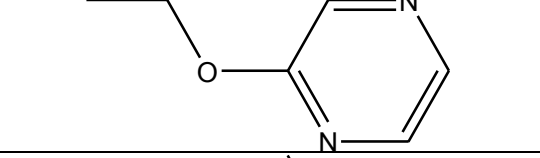
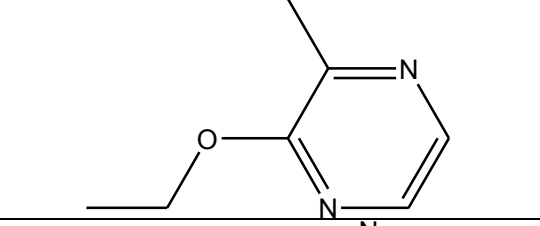
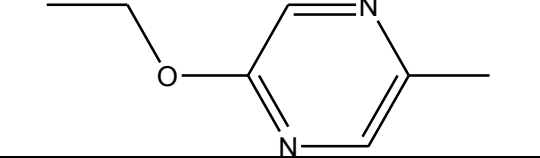
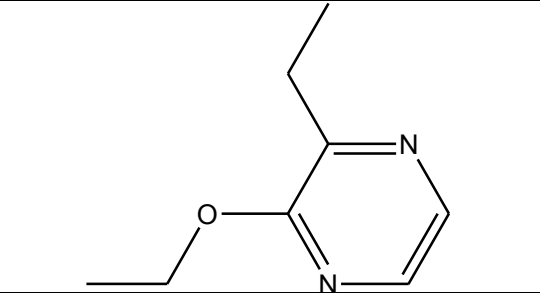
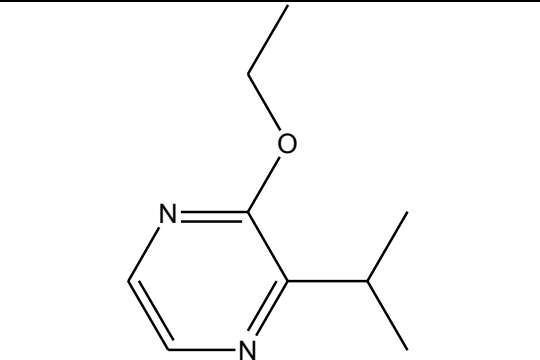
19	2,5-dimethyl-3-propylpyrazine		1142	1474
20	2,6-dimethyl-3-propylpyrazine		1151	1493
21	Isopropylpyrazine		949	1316
22	2,3-dimethyl-5-isopropylpyrazine		1112	1431
23	Butylpyrazine		1088	1474
24	2-butyl-3-methylpyrazine		1121	1459
25	3-butyl-3,5-dimethylpyrazine		1184	1487
26	3-butyl-3,6-dimethylpyrazine		1196	1514
27	5-butyl-2,3-dimethylpyrazine		1254	1600

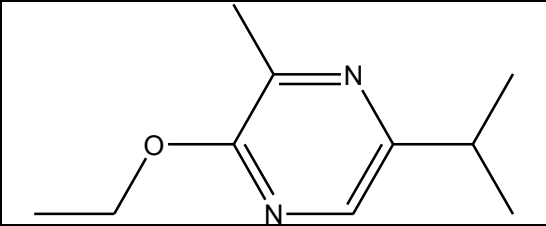
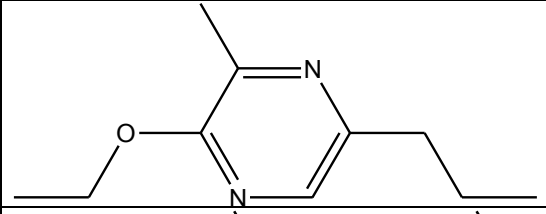
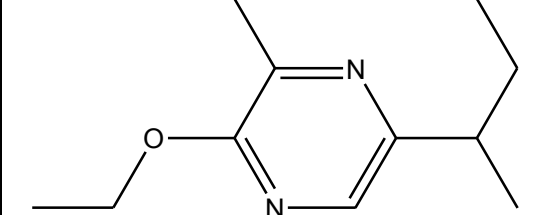
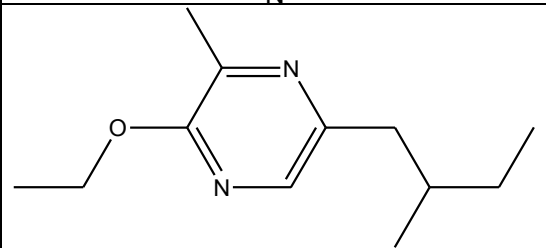
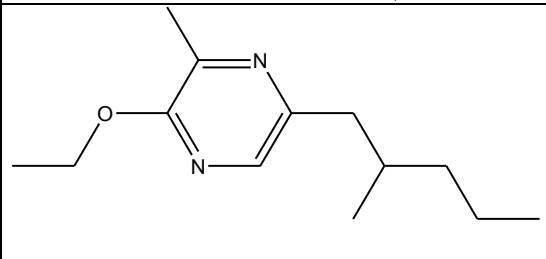
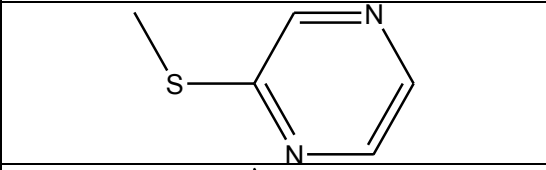
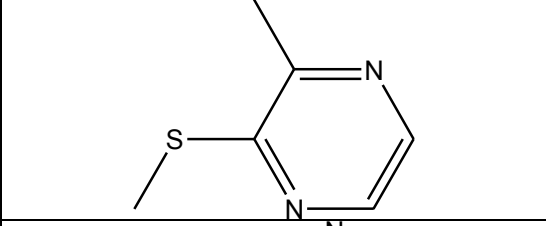
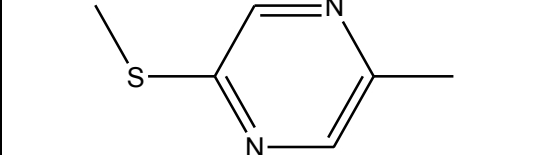
28	Isobutylpyrazine		1043	1406
29	2,3-dimethyl-5-isobutylpyrazine		1200	1525
30	2-isobutyl-3,5,6-trimethylpyrazine		1263	1556
31	sec-butylpyrazine		1040	1394
32	5-sec-butyl-2,3-dimethylpyrazine		1194	1500
33	Pentylpyrazine		1192	1575
34	2,3-dimethyl-5-pentylpyrazine		1352	1700
35	Isopentylpyrazine		1157	1530

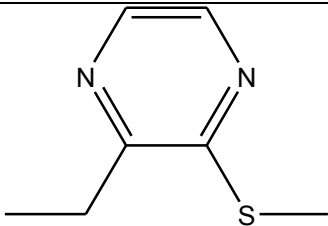
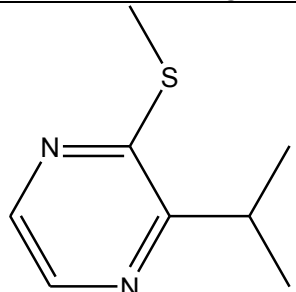
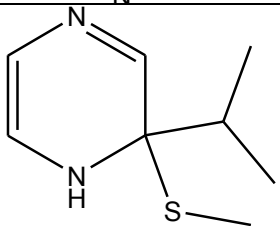
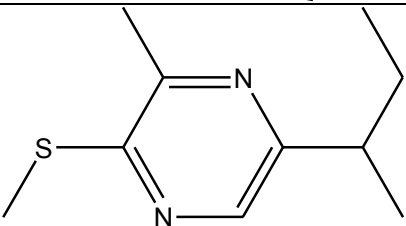
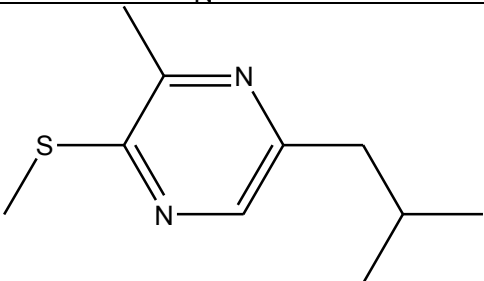
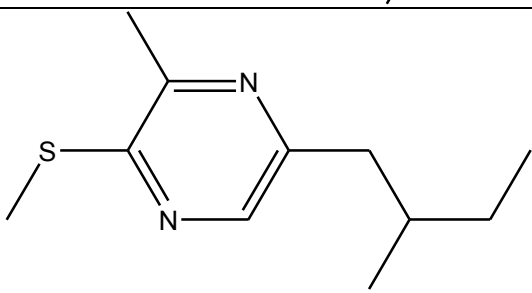
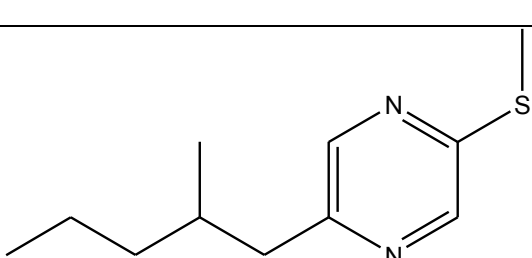
36	2,3-dimethyl-5-isopentylpyrazine		1317	1655
37	(2-methylbutyl)pyrazine		1151	1527
38	2,3-dimethyl-5-(2-methylbutyl)pyrazine		1306	1636
39	2-(2-methylbutyl)-2,5,6-trimethylpyrazine		1363	1661
40	(2-methyl-3-pentyl)pyrazine		1240	1606
41	(2-ethylpropyl)pyrazine		1121	1449
42	(1-methylbutyl)pyrazine		1133	1471

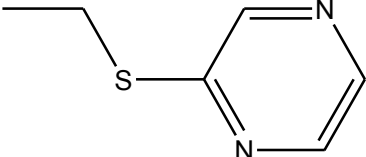
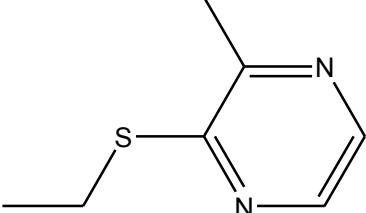
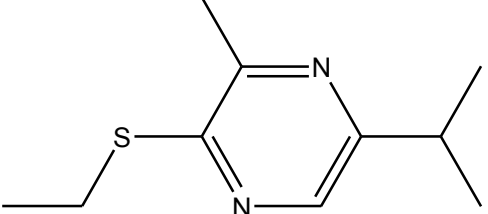
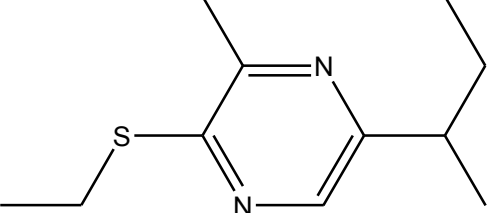
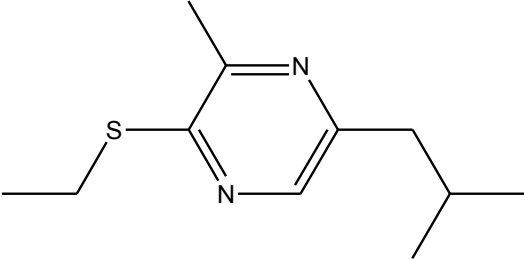
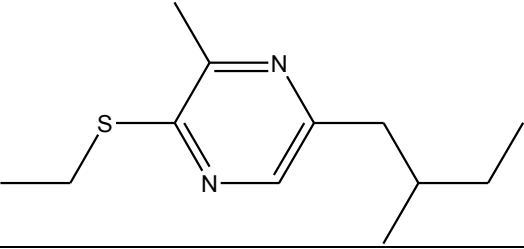
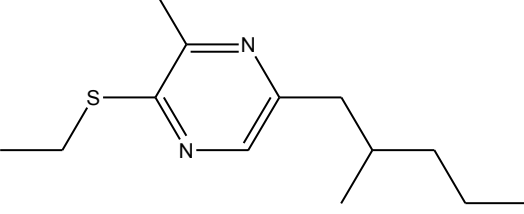
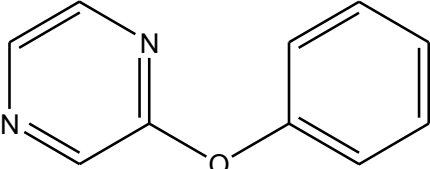
43	2,3-dimethyl-5-(2-methylpentyl)pyrazine		1377	1710
44	Hexylpyrazine		1293	1668
45	Octylpyrazine		1495	1845
46	2-methyl-3-octylpyrazine		1546	1956
47	2-methyl-5-(2-methylbutyl)-3-octylpyrazine		1923	2200
48	2-methyl-6-(2-methylbutyl)-3-octylpyrazine		1962	2264

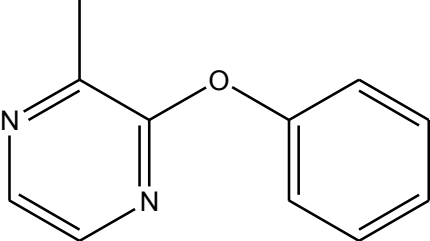
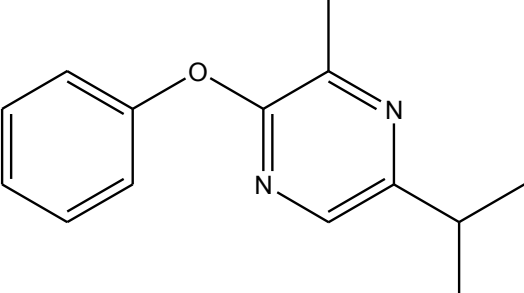
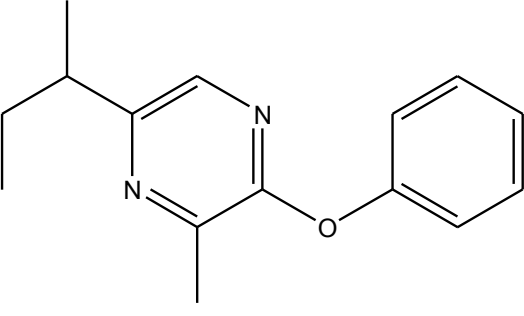
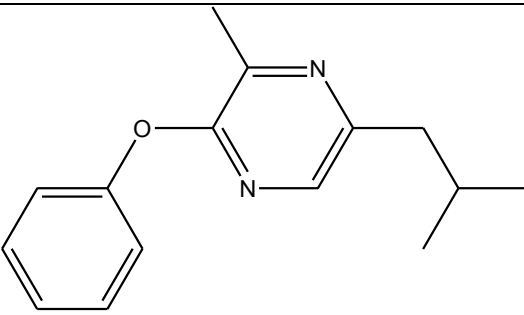
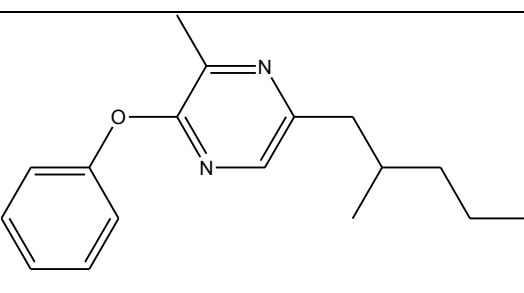
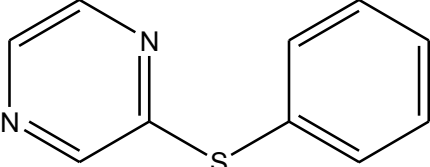
49	Methoxypyrazine		877	1306
50	2-methoxy-3-methylpyrazine		954	1339
51	2-methoxy-5-methylpyrazine		969	1358
52	3-ethyl-2-methoxypyrazine		1037	1400
53	3-isopropyl-2-methoxypyrazine		1078	1400
54	5-isopropyl-3-methyl-2-methoxypyrazine		1170	1467
55	5-sec-butyl-3-methyl-2-methoxypyrazine		1250	1536
56	5-isobutyl-3-methyl-2-methoxypyrazine		1257	1556

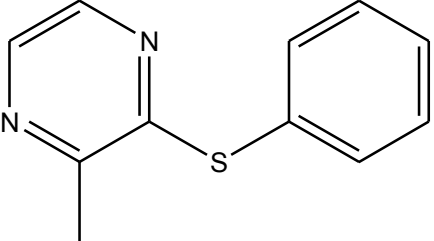
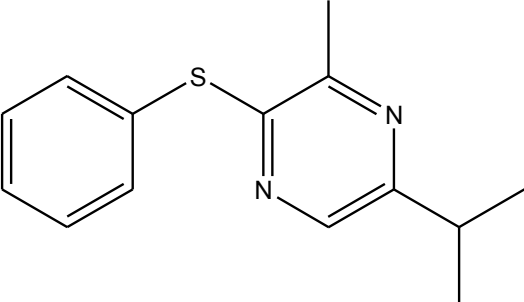
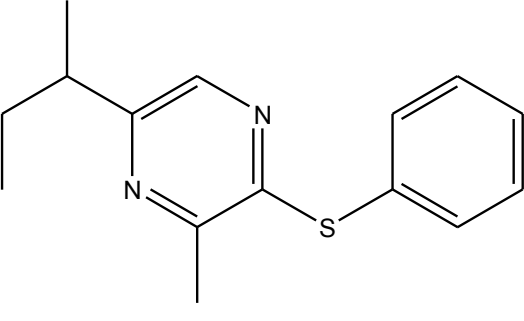
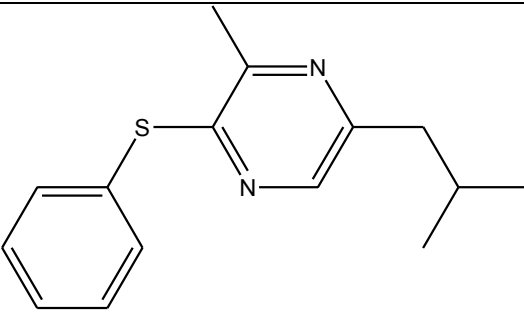
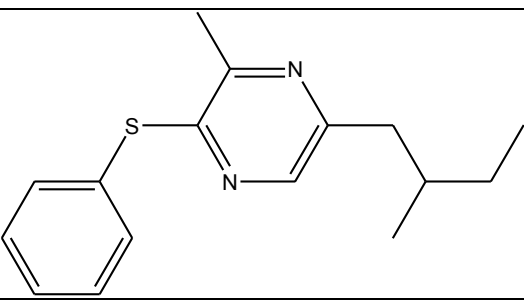
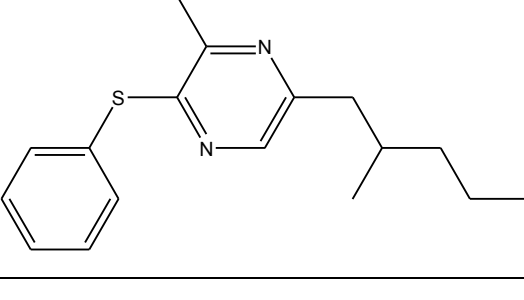
57	3-methyl-2-methoxy-5-(2-methylbutyl)pyrazine		1362	1664
58	3-methyl-2-methoxy-5-(2-methylpentyl)pyrazine		1444	1737
59	Ethoxypyrazine		959	1348
60	2-ethoxy-3-methylpyrazine		1029	1385
61	2-ethoxy-5-methylpyrazine		1047	1418
62	2-ethoxy-3-ethylpyrazine		1101	1439
63	2-ethoxy-3-isopropylpyrazine		1143	1431

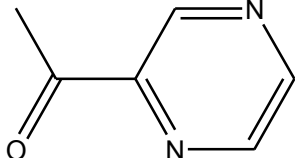
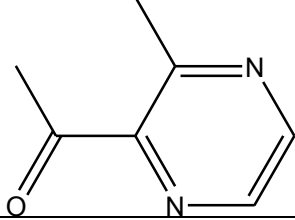
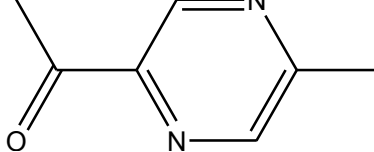
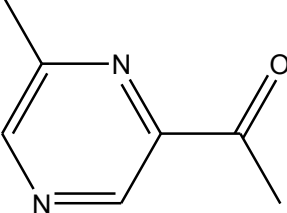
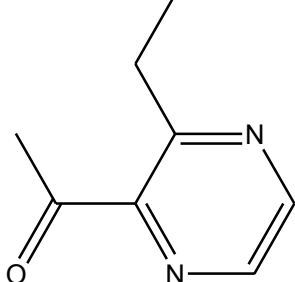
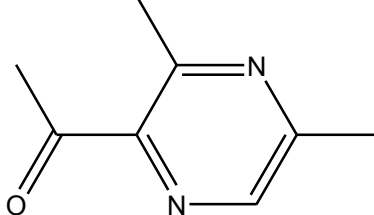
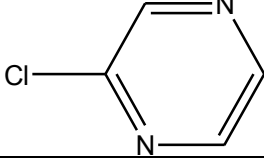
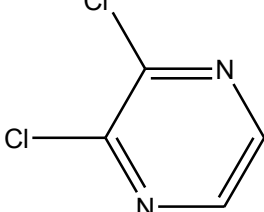
64	2-ethoxy-5-isopropyl-3-methylpyrazine		1230	1500
65	2-ethoxy-5-isobutyl-3-methylpyrazine		1314	1584
66	5-sec-butyl-2-ethoxy-3-methylpyrazine		1306	1566
67	2-ethoxy-3-methyl-5-(2-methylbutyl)pyrazine		1415	1693
68	2-ethoxy-3-methyl-5-(2-methylpentyl)pyrazine			1771
69	(methylthio)pyrazine		1076	1600
70	3-methyl-2-(methylthio)pyrazine		1151	1616
71	5-methyl-2-(methylthio)pyrazine		1163	

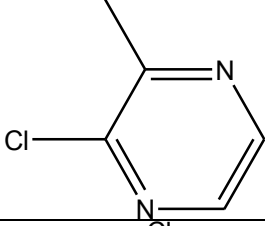
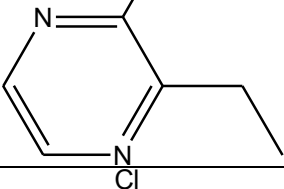
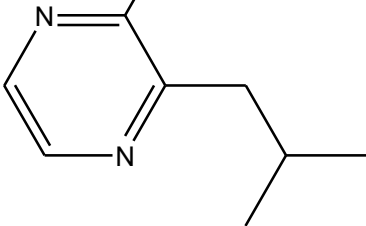
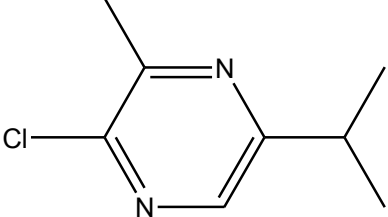
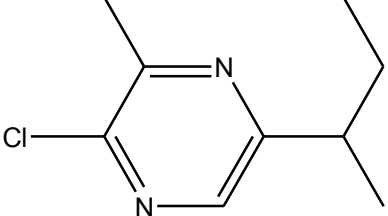
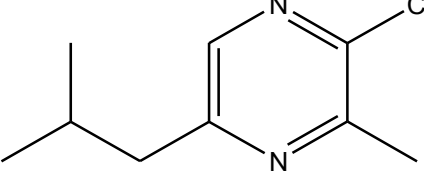
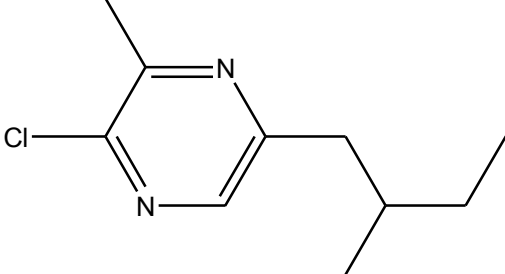
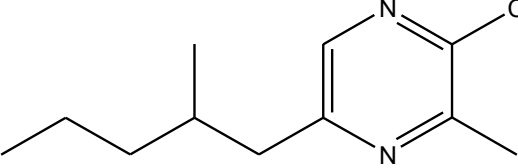
72	3-ethyl-2-(methylthio)pyrazine		1237	1695
73	3-isopropyl-2-(methylthio)pyrazine		1273	1692
74	3-isopropyl-3-(methylthio)pyrazine		1362	1737
75	5-sec-butyl-3-methyl-2-(methylthio)pyrazine		1441	1800
76	5-isobutyl-3-methyl-2-(methylthio)pyrazine		1446	1816
77	3-methyl-5-(2-methylbutyl)-2-(methylthio)pyrazine		1552	1941
78	3-methyl-5-(2-methylpentyl)-2-(methylthio)pyrazine		1638	2008

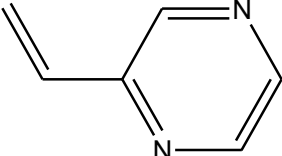
79	(ethylthio)pyrazine		1148	1635
80	2-ethylthio-3-methylpyrazine		1215	1655
81	2-ethylthio-5-isopropyl-3-methylpyrazine		1418	1769
82	5-sec-butyl-2-ethylthio-3-methylpyrazine		1494	1832
83	2-ethylthio-5-isobutyl-3-methylpyrazine		1496	1843
84	2-ethylthio-3-methyl-5-(2-methylbutyl)pyrazine		1602	1951
85	2-ethylthio-3-methyl-5-(2-methylpentyl)pyrazine		1686	2026
86	Phenoxy pyrazine		1415	2104

87	2-methyl-3-phenoxy pyrazine		1465	2103
88	5-isopropyl-3-methyl-2-phenoxy pyrazine		1620	2114
89	5-sec-butyl-3-methyl-2-phenoxy pyrazine		1694	2173
90	5-isobutyl-3-methyl-2-phenoxy pyrazine		1706	2209
91*	3-methyl-5-(2-methylpentyl)-2-phenoxy pyrazine		1807	2301
92*	(phenylthio)pyrazine		1606	2400

93*	3-methyl-2-(phenylthio)pyrazine		1658	2399
94*	5-isopropyl-3-methyl-2-(phenylthio)pyrazine		1806	2375
95*	5-sec-butyl-3-methyl-2-(phenylthio)pyrazine		1874	2430
96*	5-isobutyl-3-methyl-2-(phenylthio)pyrazine		1882	2452
97*	3-methyl-5-(2-methylbutyl)-2-(phenylthio)pyrazine		1985	2569
98*	3-methyl-5-(2-methylpentyl)-2-(phenylthio)pyrazine		2064	2669

99*	Acetylpyrazine		993	1571
100*	2-acetyl-3-methylpyrazine		1061	1567
101*	2-acetyl-5-methylpyrazine		1093	1625
102*	2-acetyl-6-methylpyrazine		1089	1618
103*	2-acetyl-3-ethylpyrazine		1138	1617
104*	2-acetyl-3,5-dimethylpyrazine		1153	1629
105*	Chloropyrazine		861	1351
106*	2,3-dichloropyrazine		1032	1581

107 *	2-chloro-3-methylpyrazine		951	1399
108 *	2-chloro-3-ethylpyrazine		1044	1467
109 *	2-chloro-3-isobutylpyrazine		1187	1575
110 *	2-chloro-5-isopropyl-3-methylpyrazine		1173	1505
111 *	5-sec-butyl-2-chloro-3-methylpyrazine		1256	1577
112 *	2-chloro-5-isobutyl-3-methylpyrazine		1264	1600
113 *	2-chloro-3-methyl-5-(2-methylbutyl)pyrazine		1371	1710
114 *	2-chloro-3-methyl-5-(2-methylpentyl)pyrazine		1456	1789

115 *	2-Vinylpyrazine		907	1392
----------	-----------------	--	-----	------

(*) les Composés de validation



ANNEXE : Programmes de calculs

Annexe

1-Calibrage :

Avec n=89 et 3 descripteurs :

```

load -ascii des.dat
load -ascii y.dat
X=des;
Y=y;

minp=0;
minN=0;
for v=1:1 %while ((minp*minN)<0)

c=Y(1:4);
a=X(1:4,1:4);

ainv=inv(a);
%*
beta=a\c;%inv(a)*c
z = zeros(size(Y));
w = zeros(size(Y));
W = zeros(size(c));
Wop = zeros(size(c));

for j=1:4
    for i=1:89

        z(i)=Y(i)-X(i,1:4)*beta;
        w(i)=X(i,1:4)* ainv(:,j);
    end
    wminus=0;wplus=0;
    for i=1:89
        if ((z(i)/w(i))<=0)
            wminus=wminus+abs(w(i));
        else wplus=wplus+abs(w(i));
        end
    end
    W(j)= wminus-wplus;
    Wop(j)=- (W(j)-1)+1;
    Wop;
end
%trouver le min
[minp,iminp]=min(W);
[minN,iminN]=min(Wop);
if (minN < minp)
    r=iminN;
    ainv(:,r)=-ainv(:,r);

else r=iminp;
end
%ainv(:,r);
%%%calcul de zi/wi pour le vecteur le plus negatif
zdw= zeros(size(y));
absw= zeros(size(y));
for i=1:89
    z(i)=Y(i)-X(i,1:4)*beta;

```



```

w(i)=X(i,1:4)*ainv(:,r);
zdw(i)=z(i)/w(i);
absw(i)=abs(w(i));
end

s=cat(2,zdw,absw);
ss=s(5:89,:);% de P+1 jusqu'à N

%sortrows(cat(zdw,des,y),1);
d=sortrows(ss,1); % tri valeur absolut!!! de w selon zi/wi

T=0;
Tdemi=0;
for i=1:85
    T=T+d(i,2);
end

k=0;
for i=1:85
    Tdemi=Tdemi+d(i,2);
    if (Tdemi < T/2)
        if (Tdemi + d(i+1,2)> T/2)
            k=i+1;
            break;
        end
    end

end

end

%il faut trier les xi yi selon z/w
desy=cat(2,X,Y);
desr=cat(2,zdw,desy);
desr
    cc=desr(5:89,1:6);
    clas=sortrows(cc,1);

    desSorted=clas(1:85,2:6);
    recons=cat(1,desy(1:4,1:5),desSorted);

%remplacement des observations
%temp=0;
for j=1:5
    temp=recons(r,j);
    recons(r,j)=recons(k+4,j);
    recons(k+4,j)=temp;%
end
X=recons(1:89,1:4);
Y=recons(1:89,5);
k
r

end

minp
minN
c=Y(1:4);

```

```
a=X(1:4,1:4);
c
a
```

2-Validation :

```
load -ascii des.dat
load -ascii y.dat
X=des;
Y=y;

minp=0;
minN=0;
for v=1:1 %while ((minp*minN)<0)

c=Y(1:4);
a=X(1:4,1:4);

ainv=inv(a);
%*
beta=a\c;%inv(a)*c

z = zeros(size(Y));
w = zeros(size(Y));
W = zeros(size(c));
Wop = zeros(size(c));

for j=1:4
    for i=1:25

        z(i)=Y(i)-X(i,1:4)*beta;
        w(i)=X(i,1:4) * ainv(:,j);
    end
    wminus=0;wplus=0;
    for i=1:25
        if ((z(i)/w(i))<=0)
            wminus=wminus+abs(w(i));
        else wplus=wplus+abs(w(i));
        end
    end
    W(j)= wminus-wplus;
    Wop(j)=- (W(j)-1)+1;
    Wop;
end
%trouver le min
[minp, iminp]=min(W);
[minN, iminN]=min(Wop);
if (minN < minp)
    r=iminN;
    ainv(:,r)=-ainv(:,r);

else r=iminp;
end
%ainv(:,r);
%%%calcul de zi/wi pour le vecteur le plus negatif
zdw= zeros(size(y));
absw= zeros(size(y));
for i=1:25
```

```

    z(i)=Y(i)-X(i,1:4)*beta;
    w(i)=X(i,1:4)*ainv(:,r);
    zdw(i)=z(i)/w(i);
    absw(i)=abs(w(i));
end

s=cat(2,zdw,absw);
ss=s(5:25,:);% de P+1 jusqu'à N

%sortrows(cat(zdw,des,y),1);
d=sortrows(ss,1); % tri valeur absolut!!! de w selon zi/wi

T=0;
Tdemi=0;
for i=1:21
    T=T+d(i,2);
end

k=0;
for i=1:21
    Tdemi=Tdemi+d(i,2);
    if (Tdemi < T/2)
        if (Tdemi + d(i+1,2)> T/2)
            k=i+1;
            break;
        end
    end

end

end

%il faut trier les xi yi selon z/w
    desy=cat(2,X,Y);
    desr=cat(2,zdw,desy);
    desr
        cc=desr(5:25,1:6);
        clas=sortrows(cc,1);

        desSorted=clas(1:21,2:6);
        recons=cat(1,desy(1:4,1:5),desSorted);

%remplacement des observations
    %temp=0;
    for j=1:5
        temp=recons(r,j);
        recons(r,j)=recons(k+4,j);
        recons(k+4,j)=temp;%
    end
    X=recons(1:25,1:4);
    Y=recons(1:25,5);
    k
    r

end

minp
minN

```

```
c=Y(1:4);  
a=X(1:4,1:4);  
c  
a
```



ANNEXE : publications

Original Research Paper

Least Absolute Deviation Regression and Least Squares for Modeling Retention Indices of Set Compounds Food and Pollutants of the Environment

Fatiha Mebarki, Khadija Amirat, Salima Ali Mokhnach and Djellol Messadi

Department of Chemistry, Laboratory of Environmental security and Food,
Badji Mokhtar Annaba University, Annaba, Algeria

Article history

Received: 22-12-2016

Revised: 04-06-2017

Accepted: 05-06-2017

Corresponding Author:

Khadija Amirat

Department of Chemistry,
Laboratory of Environmental
security and Food, Badji
Mokhtar Annaba University,
Annaba, Algeria
kadijatoumi@yahoo.com

Abstract: Considering the importance of the statistical analysis of regression in modeling based separately on study for Quantitative structure retention indices on Carbowax 20 M ($I^{C_{w20M}}$) and OV-101 columns (I^{OV-101}) relationships (QSRR) are determined for 114 pyrazines. The detection of influential observations for the standard least squares regression model is a problem which has been extensively studied. Least Absolute Deviation regression diagnostics offers alternative approaches whose main feature is the robustness. Here a nonparametric method for detecting influential observations is presented and compared with other classical diagnostics methods. With have been applied for modeling separately retention indices of the same set of (89 pyrazines of Training and 25 of Test) eluted on Columns OV-101 and Carbowax-20M, using theoretical molecular descriptors derived from DRAGON Software and validating the results in the state approached graphically by Probability plot of the error and approached tests statistics of Anderson-Darling, in finished by the confidence interval thanks to robustness concept to check if errors distribution is really approximate.

Keywords: Least Absolute Deviation Regression, Robustness, Outliers, Leverage Points, Tests Statistics, Environmental

Introduction

Since the 1970 the environment term is used to indicate the global Ecologic context, i.e., the whole of the conditions physical, chemical, biological climatic and geographic conditions, in which are developed living conditions and humans being in particular. Air, earth, water, natural resources, flora, fauna, people and their social interactions are included.

The volatile heterocyclic constitute a significant family of odorous molecules, particularly interesting in the field of chemistry of the flavours and the odor can be regarded as a local pollution and a limited harmful effect to the bordering population of the potential sources. They represent more than one quarter of the 5 000 volatile compounds characterized up to now in our food

Pyrazines are heterocycles very present in our food. More than 80 derived from pyrazines are identified in a great number of cooked food, as bread, meat, torrefied coffee, the cocoa or hazel nuts; they are aromatizing compounds (Li *et al.*, 2014; Buchbauer, 2000).

Stanton and Jurs (1989), have used QSRR methodology to develop Models to link structural features of 107 pyrazines differently substituted, to their retention indices obtained up on two different polarities columns (OV-101 and Carbowax-20M). The equations have been calculated with the help of multilinear regression, the choice of the explanatory variables (topological, electronic and physical properties) being achieved by progressive elimination (Small and Jurs, 1983), among the 85 individual Molecular descriptors obtained for each whole molecule. The retention Indices (IR) obtained on each column are treated separately, while by drawing from the same sets of descriptors. The calculated models with 6 explanatory variables provide high standards errors ($S = 23$ units of index - u.i. - on OV-101 and $S = 36.33$ u.i. up on Carbowax- (20 M) which do not predict good predictive capacities for these models, which let to suppose nonlinear relations between descriptors and property (IR) studied (Mebarki *et al.*, 2016).

A large number of other estimation methods aimed at achieving robustness have been suggested and a considerable body of literature has developed. See for example, Gonin and Money (1989; Dodge, 1987) and the references therein. Generally the robust estimators in the literature can be classified as M-estimators, L-estimators, or R-estimators. Probably most attention has been paid to the L-estimators, for other type estimators, Judge *et al.* (1985).

The robustness of Least Absolute Deviation method in relation with influential observations and its susceptibility to leverage point which are largely studied in literature (Dodge, 1987; 1997). We propose non parametric method Least Absolute Deviation (LAD) to detect the influential observations (aberrant and affect leverage) in comparison with least squares method.

The tests of normality as whereas theory-driven methods include the normality test such Anderson Darling test. However, seier classified the test of normality into major categories test, empirical and normality distribution of *the observed data*.

The Durbin-Watson statistic is conditioned on the order of the observations (rows). Minitab assumes that the observations are in a meaningful order, such as time order. The Durbin-Watson statistic determines whether or not the correlation between adjacent error terms is zero. To reach a conclusion from the test, you will need to compare the displayed statistic with lower and upper bounds in a table. If $D >$ upper bound, no correlation exists; if $D <$ lower bound, positive correlation exists; if D is in between the two bounds, the test is inconclusive.

The objective of this work aims at using QSRR methodology, in the approach Method Least Absolute Deviation/Least Square (LAD/OLS), to model retention indices of (114) pyrazines (113 taken from Stanton and Jurs (1979) (1) and one compound (2-VinylPyrazine) taken from Mihara and Enomoto (1985), the molecular descriptors are only calculated starting from the chemical structure of the compounds.

The linear statistical model for fixed effects will be examined relationships between retention index and different descriptors for two columns [(between retention indices of non polar column (OV- 101) and descriptor of Connectivity indices (are among the most popular topological indices (it is a descriptor of Structure-Activity Analysis), descriptor of Geometrical descriptors (representation of a molecule involves the knowledge of the relative positions of the atoms in 3D space) and descriptor of 3D-Molecule Representation of Structures based on Electron diffraction (3D-MoRSE); for relationships between retention index of polar column (CRW-20M) and descriptor of Connectivity indices (are among the most popular topological indices), descriptor of 2D autocorrelations (are molecular descriptors which describe how a

considered property is distributed along a topological molecular structure) and descriptor of 3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction)] by two robust methods for the evaluation of regression parameters starting from robust coefficients of regression most popular by the appendices. We have based ourselves on comparison between the two methods, application field (DA) will be discussed using Williams diagram which presents residues of standardized prediction according to the levers values (hi) (Eriksson *et al.*, 2003; Tropsha *et al.*, 2003). We present the state approached graphically by Probability plot of the error and approached statistics tests (Anderson-Darling), in finished by the confidence interval of compatibility at normal law to validated results of approached state between two methods for a risk $\alpha = 5\%$ (Nornadiah and Yah, 2011; Damodar *et al.*, 2009).

Methodology

The Data Set

Molecular software Hyperchem 6.03 (AL-Noor and Asmaa, 2013) is used to represent the molecules, by employing semi-empirical method AM1 (Dewar *et al.*, 1985; Holder, 1998) to obtain final geometries. The implied compounds in this study have the general structure 1.

The retention data for the 114 compounds chromatographed on stationary phases OV-101 and CRW-20M have been taken from (113 taken from Stanton and Jurs (1979) (1) and 1 compound (2-VinylPyrazine) taken from (Mihara and Enomoto, 1985) and are enumerated in Table 1.

Descriptor Generation

The optimized geometries are transferred in software dragon from data-processing software version 5.4, for calculation of 1320 descriptors while operating on 89 pyrazines of test; subsets of descriptors are chosen by genetic algorithm, these descriptors can be separate in four categories: Topological, geometrical, physical and electronic descriptors have accounts of way and molecular indices of connectivity included. The geometrical descriptors included sectors of shade, the length with the reports/ratios of width, volumes of van der Waals, the surface and principal moments of inertia. The calculated descriptors of physical property included the molecular refringency of polariz ability and molar. The electronic descriptors included most positive and most negative described by Kaliszan.

By employing the software Mobydigs (Todeschini *et al.*, 2009) and by maximizing the coefficient of prediction Q^2 and minimal R^2 of S (the error).

Table 1. Experimentally determined Retention Indices for pyrazines on OV-101 and Carbowax-20 M

n°	Compounds	ov-101	Compounds	IR(cw)
1	Pyrazine	710	Pyrazine	1179
2	Methylpyrazine	801	Methylpyrazine	1235
3	2,3-dimethylpyrazine	897	2,3-dimethylpyrazine	1309
4	2,5-dimethylpyrazine	889	2,5-dimethylpyrazine	1290
5	2,6-dimethylpyrazine	889	2,6-dimethylpyrazine	1300
6	Trimethylpyrazine	981	Trimethylpyrazine	1365
7	Trimethylpyrazine	1067	Trimethylpyrazine	1439
8	Ethylpyrazine	894	Ethylpyrazine	1300
9	2-ethyl-5-methylpyrazine	980	2-ethyl-5-methylpyrazine	1357
10	2-ethyl-6-methylpyrazine	977	2-ethyl-6-methylpyrazine	1353
11	2,5-dimethyl-3-ethylpyrazine	1059	2,5-dimethyl-3-ethylpyrazine	1400
12	2,6-dimethyl-6-ethylpyrazine	1064	2,6-dimethyl-6-ethylpyrazine	1415
13	2,3-dimethyl-5-ethylpyrazine	1066	2,3-dimethyl-5-ethylpyrazine	1421
14	2,3-diethylpyrazine	1065	2,3-diethylpyrazine	1417
15	2,3-diethyl-5-methylpyrazine	1137	2,3-diethyl-5-methylpyrazine	1459
16	Propylpyrazine	986	Propylpyrazine	1374
17	2-methyl-3-propylpyrazine	1072	2-methyl-3-propylpyrazine	1438
18	2,3-dimethyl-5-propylpyrazine	1154	2,3-dimethyl-5-propylpyrazine	1500
19	2,5-dimethyl-3-propylpyrazine	1142	2,5-dimethyl-3-propylpyrazine	1474
20	2,6-methyl-3-propylpyrazine	1151	2,6-methyl-3-propylpyrazine	1493
21	Isopropyl pyrazine	949	Isopropylpyrazine	1316
22	2,3-dimethyl-5-isopropylpyrazine	1112	2,3-dimethyl-5-isopropylpyrazine	1431
23	Butylpyrazine	1088	Butylpyrazine	1474
24	2-butyl-3-methylpyrazine	1121	2-butyl-3-methylpyrazine	1459
25	3-butyl-3,5-dimethylpyrazine	1184	3-butyl-3,5-dimethylpyrazine	1487
26	3-butyl-3,6-dimethylpyrazine	1196	3-butyl-3,6-dimethylpyrazine	1514
27	5-butyl-2,3-dimethylpyrazine	1254	5-butyl-2,3-dimethylpyrazine	1600
28	Isobutyl pyrazine	1043	Isobutylpyrazine	1406
29	2,3-dimethyl-5-isobutylpyrazine	1200	2,3-dimethyl-5-isobutylpyrazine	1525
30	2,3-isobutyl-3,5,6-trimethylpyrazine	1263	2-isobutyl-3,5,6-trimethylpyrazine	1556
31	sec-butylpyrazine	1040	sec-butylpyrazine	1394
32	5-sec-butyl-2,3-dimethylpyrazine	1194	5-sec-butyl-2,3-dimethylpyrazine	1500
33	Pentylpyrazine	1192	Pentylpyrazine	1575
34	2,3-dimethyl-5-pentylpyrazine	1352	2,3-dimethyl-5-pentylpyrazine	1700
35	Isopentylpyrazine	1157	Isopentylpyrazine	1530
36	2,3-dimethyl-5-isopentylpyrazine	1317	2,3-dimethyl-5-isopentylpyrazine	1655
37	(2-methylbutyl) pyrazine	1151	(2-methylbutyl) pyrazine	1527
38	2,3-dimethyl-5-(2-methylbutyl) pyrazine	1306	2,3-dimethyl-5-(2-methylbutyl) pyrazine	1636
39	2-(2-methylbutyl)-2,5,6-trimethylpyrazine	1363	2-(2-methylbutyl)-2,5,6-trimethylpyrazine	1661
40	(2-methyl-3-pentyl) pyrazine	1240	(2-methyl-3-pentyl) pyrazine	1606
41	(2-ethylpropyl) pyrazine	1121	(2-ethylpropyl) pyrazine	1449
42	(1-methylbutyl) pyrazine	1133	(1-methylbutyl) pyrazine	1471
43	2,3-dimethyl-5-(2-methylpentyl) pyrazine	1377	2,3-dimethyl-5-(2-methylpentyl) pyrazine	1710
44	Hexylpyrazine	1293	Hexylpyrazine	1668
45	Octylpyrazine	1495	Octylpyrazine	1845
46	2-methyl-3-octylpyrazine	1546	2-methyl-3-octylpyrazine	1956
47	2-methyl-5-(2-methylbutyl)-3-octylpyrazine	1923	2-methyl-5-(2-methylbutyl)-3-octylpyrazine	2200
48	2-methyl-6-(2-methylbutyl)-3-octylpyrazine	1962	2-methyl-6-(2-methylbutyl)-3-octylpyrazine	2264
49	Methoxypyrazine	877	Methoxypyrazine	1306
50	2-methoxy-3-methylpyrazine	954	2-methoxy-3-methylpyrazine	1339
51	2-methoxy-5-methylpyrazine	969	2-methoxy-5-methylpyrazine	1358
52	3-ethyl-2-methoxypyrazine	1037	3-ethyl-2-methoxypyrazine	1400
53	3-isopropyl-2-methoxypyrazine	1078	3-isopropyl-2-methoxypyrazine	1400
54	5-isopropyl-3-methyl-2-methoxypyrazine	1170	5-isopropyl-3-methyl-2-methoxypyrazine	1467
55	5-sec-butyl-3-methyl-2-methoxypyrazine	1250	5-sec-butyl-3-methyl-2-methoxypyrazine	1536
56	5-isobutyl-3-methyl-2-methoxypyrazine	1257	5-isobutyl-3-methyl-2-methoxypyrazine	1556
57	3-methyl-2-methoxy-5-(2-methylbutyl) pyrazine	1362	3-methyl-2-methoxy-5-(2-methylbutyl)pyrazine	1664
58	3-methyl-2-methoxy-5-(2-methylpentyl) pyrazine	1444	3-methyl-2-methoxy-5-(2-methylpentyl)pyrazine	1737
59	Ethoxypyrazine	959	Ethoxypyrazine	1348
60	2-ethoxy-3-methylpyrazine	1029	2-ethoxy-3-methylpyrazine	1385
61	2-ethoxy-5-methylpyrazine	1047	2-ethoxy-5-methylpyrazine	1418
62	2-ethoxy-3-ethylpyrazine	1101	2-ethoxy-3-ethylpyrazine	1439
63	2-ethoxy-3-isopropylpyrazine	1143	2-ethoxy-3-isopropylpyrazine	1431

Table 1. Continuo

64	2-ethoxy-5-isopropyl-3-methylpyrazine	1230	2-ethoxy-5-isopropyl-3-methylpyrazine	1500
65	2-ethoxy-5-isobutyl-3-methylpyrazine	1314	2-ethoxy-5-isobutyl-3-methylpyrazine	1584
66	5-sec-butyl-2-ethoxy-3-methylpyrazine	1306	5-sec-butyl-2-ethoxy-3-methylpyrazine	1566
67	2-ethoxy-3-methyl-5-(2-methylbutyl) pyrazine	1415	2-ethoxy-3-methyl-5-(2-methylbutyl) pyrazine	1693
68	(methylthio) pyrazine	1076	2-ethoxy-3-methyl-5-(2-methylpentyl) pyrazine	1771
69	3-methyl-2-(methylthio) pyrazine	1151	(methylthio) pyrazine	1600
70	5-methyl-2-(methylthio) pyrazine	1163	3-methyl-2-(methylthio) pyrazine	1616
71	3-ethyl-2-(methylthio) pyrazine	1237	3-ethyl-2-(methylthio) pyrazine	1695
72	3-isopropyl-2-(methylthio) pyrazine	1273	3-isopropyl-2-(methylthio) pyrazine	1692
73	3-isopropyl-3-(methylthio) pyrazine	1362	3-isopropyl-3-(methylthio) pyrazine	1737
4	5-sec-butyl-3-methyl-2-(methylthio) pyrazine	1441	5-sec-butyl-3-methyl-2-(methylthio) pyrazine	1800
75	5-isobutyl-3-methyl-2-(methylthio) pyrazine	1446	5-isobutyl-3-methyl-2-(methylthio) pyrazine	1816
76	3-methyl-5-(2-methylbutyl)-2-(methylthio) pyrazine	1552	3-methyl-5-(2-methylbutyl)-2-(methylthio) pyrazine	1941
77	3-methyl-5-(2-methylpentyl)-2-(methylthio) pyrazine	1638	3-methyl-5-(2-methylpentyl)-2-(methylthio) pyrazine	2008
78	(ethylthio) pyrazine	1148	(ethylthio) pyrazine	1635
79	2-ethylthio-3-methylpyrazine	1215	2-ethylthio-3-methylpyrazine	1655
80	2-ethylthio-5-isopropyl-3-methylpyrazine	1418	2-	2-
	hylthio-5-isopropyl-3-methylpyrazine	1769		
81	5-sec-butyl-2-ethylthio-3-methylpyrazine	1494	5-sec-butyl-2-ethylthio-3-methylpyrazine	1832
82	2-ethylthio-5-isobutyl-3-methylpyrazine	1496	2-ethylthio-5-isobutyl-3-methylpyrazine	1843
83	2-ethylthio-3-methyl-5-(2-methylbutyl) pyrazine	1602	2-ethylthio-3-methyl-5-(2-methylbutyl) pyrazine	1951
84	2-ethylthio-3-methyl-5-(2-methylpentyl) pyrazine	1686	2-ethylthio-3-methyl-5-(2-methylpentyl) pyrazine	2026
85	Phenoxy pyrazine	1415	Phenoxy pyrazine	2104
86	2-methyl-3-phenoxy pyrazine	1465	2-methyl-3-phenoxy pyrazine	2103
87	5-isopropyl-3-methyl-2-phenoxy pyrazine	1620	5-isopropyl-3-methyl-2-phenoxy pyrazine	2114
88	5-sec-butyl-3-methyl-2-phenoxy pyrazine	1694	5-sec-butyl-3-methyl-2-phenoxy pyrazine	2173
89	5-isobutyl-3-methyl-2-phenoxy pyrazine	1706	5-isobutyl-3-methyl-2-phenoxy pyrazine	2209
90	3-methyl-5-(2-methylpentyl)-2-phenoxy pyrazine	1807	3-methyl-5-(2-methylpentyl)-2-phenoxy pyrazine	2301
91	(phenylthio) pyrazine	1606	(phenylthio) pyrazine	2400
92	3-methyl-2-(phenylthio) pyrazine	1658	3-methyl-2-(phenylthio) pyrazine	2399
93	5-isopropyl-3-methyl-2-(phenylthio) pyrazine	1806	5-isopropyl-3-methyl-2-(phenylthio) pyrazine	2375
94	5-sec-butyl-3-methyl-2-(phenylthio) pyrazine	1874	5-sec-butyl-3-methyl-2-(phenylthio) pyrazine	2430
95	5-isobutyl-3-methyl-2-(phenylthio) pyrazine	1882	5-isobutyl-3-methyl-2-(phenylthio) pyrazine	2452
96	3-methyl-5-(2-methylbutyl)-2-(phenylthio) pyrazine	1985	3-methyl-5-(2-methylbutyl)-2-(phenylthio) pyrazine	2569
97	3-methyl-5-(2-methylpentyl)-2-(phenylthio) pyrazine	2064	3-methyl-5-(2-methylpentyl)-2-(phenylthio) pyrazine	2669
98	Acetylpyrazine	993	Acetylpyrazine	1571
99	2-acetyl-3-methylpyrazine	1061	2-acetyl-3-methylpyrazine	1567
100	2-acetyl-5-methylpyrazine	1093	2-acetyl-5-methylpyrazine	1625
101	2-acetyl-6-methylpyrazine	1089	2-acetyl-6-methylpyrazine	1618
102	2-acetyl-3-ethylpyrazine	1138	2-acetyl-3-ethylpyrazine	1617
103	2-acetyl-3,5-dimethylpyrazine	1153	2-acetyl-3,5-dimethylpyrazine	1629
104	Chloropyrazine	861	Chloropyrazine	1351
105	2,3-dichloropyrazine	1032	2,3-dichloropyrazine	1581
106	2-chloro-3-methylpyrazine	951	2-chloro-3-methylpyrazine	1399
107	2-chloro-3-ethylpyrazine	1044	2-chloro-3-ethylpyrazine	1467
108	2-chloro-3-isobutylpyrazine	1187	2-chloro-3-isobutylpyrazine	1575
109	2-chloro-5-isopropyl-3-methylpyrazine	1173	2-chloro-5-isopropyl-3-methylpyrazine	1505
110	5-sec-butyl-2-chloro-3-methylpyrazine	1256	5-sec-butyl-2-chloro-3-methylpyrazine	1577
111	2-chloro-5-isobutyl-3-methylpyrazine	1264	2-chloro-5-isobutyl-3-methylpyrazine	1600
112	2-chloro-3-methyl-5-(2-methylbutyl) pyrazine	1371	2-chloro-3-methyl-5-(2-methylbutyl) pyrazine	1710
113	2-chloro-3-methyl-5-(2-methylpentyl) pyrazine	1456	2-chloro-3-methyl-5-(2-methylpentyl) pyrazine	1789
114	2-VinylPyrazine	907	2-VinylPyrazine	1392

Regression Analysis

The analysis of the multiple linear regressions was carried out with two methods by software Matlab (2009) for (Least Absolute Deviation) and Minitab (16) for (OLS).

We considers the multiple model of regression wich is given by (Berlin, 1982):

$$y_i = \beta_0 + \sum_{j=2}^{p-1} \beta_j x_{ij} + \varepsilon_i \quad (1)$$

Detection of meaningless statements and with action leverage according to the method of least squares is a problem which is largely studied. Diagnosis by the Least Absolute Deviation regression offers alternative

approaches whose principal characteristic is robustness. In our study a non-parametric method to detect the meaningless statements and point's lever is applied and compared with the traditional method of diagnosis (least squares).

Least Squares OLS Method

This is carried out with software Minitab 16, method OLS with is applied to multiple regression which consists in defining the β estimate which minimizes:

$$\sum ei^2 = \sum (y_i - \beta_0 - \sum x_{ij})^2 \quad (2)$$

Least Absolute Deviations (LAD) Method

The analysis of linear regression multiple is carried out with software Matlab (2009), by using the Least Absolute Deviations (LAD) method, which is one of the principal alternatives to the method of least squares when it is a question of estimating parameters of regression on, which minimizes the absolute values but not the values with square of the term of error. Least Absolute Deviation Method applied to the multiple regression consists in defining the β estimates which minimize (Dodge and Jureckova, 2000, Dodge, 2004):

$$\sum |ei| = \sum |y_i - \beta_0 - \sum \beta x_{ij}| \quad (3)$$

Results and Discussion

An ideal model is one that has a high R value, a smallest value of standard error, starting from independent variables. The best models found has 3 descriptors for each stationary phase by using the software Moby Digs are given below.

The criterion for identifying a compound as an outlier is that compound is diminished by three or more of six standard statistical tests used to detect outliers in regression analysis. These tests were (1) residual, (2) standardized residual, (3) Studentized residual, (4) leverage, (5) DFFITS, (6) Cook's distance. The residual is the difference between real value and the value predicted by the regression equation. The standardized residual is the residual divided by difference models of regression equation. The Studentized residual is the residual of forecast divided by proper model difference.

Leverage allows for the determination of a point the influence.

DFFITS describes difference in the fits of the equation caused by displacement of a given observation and Cook's distance describes the change of a model coefficient by the displacement of indicated point.

The definition of each descriptor is given Table 2.

The coefficient of multiple determinations (R^2) indicates the amount of variance in data is explained by the model. The standard error of regression coefficient is given in each case and n indicates of molecules involved in regression analysis procedure.

The Best Models

IR (OV-101) : (XMOD, FDI, Mor 06 v); S = 18.379, $R^2 = 99.4$, n = 89 compounds

IR (CRW20M) : (RDCHI, GATS1p, Mor 02 m); S = 34.933, $R^2 = 98.08$, n = 89 compounds

The best tree parametric model was constructed using:

[OV-101: Modified Randi connectivity index (XMOD) (is a molecular descriptor proposed as the sum of atomic properties, accounting for valence electrons and extended connectivities in the H-depleted molecular graph using a Randic connectivity index-type formula), Folding Degree Index (FDI) (is the largest eigenvalue 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) and (Mor06v) (3D-MORSE-signal 06/weighted by atomic Vander Waals volumes (Mor06v) (3D-MORSE) (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. 3D-MORSE the descriptors are calculated for five different atomic properties w: the unweighted case (u), atomic mass (m), the van der Waals volume (v), the Sanderson atomic electro negativity (e) and, the atomic polarizability (p). (CRW-20M: Reciprocal Distance Randi-type Index (RDCHI) (is defined on the analogy of the Randic connectivity index X1, where the vertex degrees are substituted by the row sums of the reciprocal distance matrix. Moreover, the reciprocal distance squared Randictype-index RDSQ is obtained from the RDCHI index substituting the exponent-1/2 with 1/2.), Geary Autocorrelation -log 1/weighted by atomic polariz abilities (GATS1p) (2D autocorrelations calculated by DRAGON are spatial autocorrelations calculated on a H-depleted molecular graph weighted by atom physico-chemical properties (i.e., the atom weightings w) and include: Autocorrelations GATS calculated by the Geary coefficient) and 3D-MORSE-signal 02/weighted by atomic masses (Mor02m)].

Table 2. Definitions of descriptors used in the retention index prediction models

Descriptors	The definition
XMOD	Modified Randi connectivity index
FDI	folding degree index
Mor06v	(3D-MORSE-signal 06/weighted by atomic Vander Waals volumes
RDCHI	reciprocal distance Randi-type index
GATS1p	Geary autocorrelation -log 1/weighted by atomic polarizabilities
Mor02m	3D-MORSE-signal 02/weighted by atomic masses

Using a significance level of 0.05, the Anderson-Darling normality test (Fig. 1) (A-Squared = 0,134; OV-101, A-Squared = 0,270; Crbowax- 20 $M < v_{cri} = 0.752$) indicates that the resting pulse data follow a normal distribution But it disturbance that if outliers may be present in the measurements.

Auto Correlation of the Residus

Values of the statistics of Durbin-Watson (Durbin and Watson, 1951), [d = 1,47910; OV-101/D = 1,29968; Carbowax-20M] are the greater than higher values given by the tables, respectively for 3 regresses and for reasonable risk $\alpha = 0.05$, which expresses positive auto correlation of residues which establishes each time the independence of the residues include the absence of autocorrelation that if outliers may be present in the measurements.

Column RCW -20 M

Column OV -101

The diagnostic statistics joined together in Table 3 make it possible to make comparisons and to draw several conclusions.

All relevant statistical parameters are reported in Table 3.

Values of R^2 and R^2_{adj} attest the good fitting performances of the model which, moreover, is very highly significant (great value of the Fisher parameter F).

The model is robust, the difference between R^2 and Q^2 is small (0.05% of Colum OV-101 and 0.22% of Colum CRW-20M). The model demonstrates a very good stability in internal validation while bootstrapping confirms the internal (Q^2_{boot}) predictivity and stability of the model. SDE Pext is a little bit different from SDEP. The model works slightly worse in external prediction than in internal prediction.

Correlation Matrix between Retention Indices and the Selected Descriptors

Column OV-101

ov-101	XMOD	FDI
XMOD	0,986	
	0,000	
FDI	-0,039	-0,152

	0,715	0,154		
Mor06v	0,181	0,059	0,274	
	0,089	0,582	0,009	

Column CRW-20M:

	IR (cw)	RDCHI	GATS1p
RDCHI	0,893		
	0,000		
GATS1p	-0,375	0,044	
	0,000	0,681	
Mor02m	0,896	0,930	-0,024
	0,000	0,000	0,821

The matrix of correlation Table 4, obtained using the order Correlation of software MINITAB, shows that the descriptors are more or less correlated between them ($r \geq 0,39$ for a $p = 0,045 < \alpha = 0.05$).

All the descriptors respectively are correlated with the retention index of the CRW-20M phase except the GATS1p descriptor is correlated less and with the retention index of phase OV -101 descriptor (XMOD) is correlated and the Descriptors (FDI, Mor06v) less correlated.

The Least Squares method of estimation of parameters of linear (regression) models performs well provided that the residuals are well not behaved. However, models with the disturbances that are prominently non-normally distributed or follow a normal distribution But it disturbance and contain sizeable outliers fail estimation by the Least Squares method. An intensive research has established that in such cases estimation by the Least Absolute Deviation (LAD) method performs well.

Multiple linear Regression Comparison Robust Regression of OLS and Least Absolute Deviation

We will try More particularly 2 estimate methods for the vector $((\beta_0^*, \beta_1^*, \dots, \beta_k^*))$ of Parameters:

- Method of ordinary least squares, the most known and the most used.
- The method Least Absolute Deviation (LAD) (Sum of the absolute values of the errors) (Machabert, 2014).

Table 3. Statistics diagnostic for the selected models

Colum	Models	R ²	Q ²	Q ² boot	Q ² ext	R ² adj	Kx
OV-101	X1sol Mor06v AMR	99,44	99,39	99,35	97,5	99,42	51,36
		65,5	18,736	17,961	4987,6	18,38	s
		R2	Q2	Q2boot	Q2ext	R2adj	Kx
CRW-20M	RDCHI GATS1p Mor02m	98,08	97,86	97,72	77,02	98,01	46,61
		Kxy	SDEP	SDEC	F	s	
		63,91	36,044	34,139	1444,5	34,93	

Table 4. Least absolute deviation estimates for model

Predictor	Coef	SE Coef	T	P
Constant	-946	100,237	-9,44	0,000
XMOD	29,1	5,216	5,58	0,000
FDI	1174,4	65,36	17,97	0,000
Mor06v	70,4	10,909	6,453	0,000

Table 5. Least squaresestimates for model

Predictor	Coef	SE Coef	T	P
Constant	-809,4	107,2	-7,55	0,000
XMOD	292,454	0,2535	115,35	0,000
FDI	1028,3	108,5	9,48	0,000
Mor06v	70,453	6,266	11,24	0,000

Table 6. Least absolute deviation estimates for model

Predictor	Coef	SE Coef	T	P
Constant	859,72	94,47	9,10	0,000
RDCHI	527,46	44,679	11,805	0,000
GATS1p	-630,74	20,68	-30,5	0,000
Mor02m	28,36	19,582	1,45	0,000

Table 7. Least squares estimates for mode

predictor	Coef	SE Coef	T	P
Constant	852,37	44,50	19,15	0,000
RDCHI	512,52	33,40	15,34	0,000
GATS1p	-636,05	24,61	-25,85	0,000
Mor02m	32,671	4,612	7,08	0,000

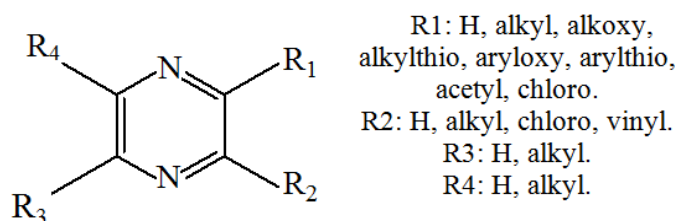
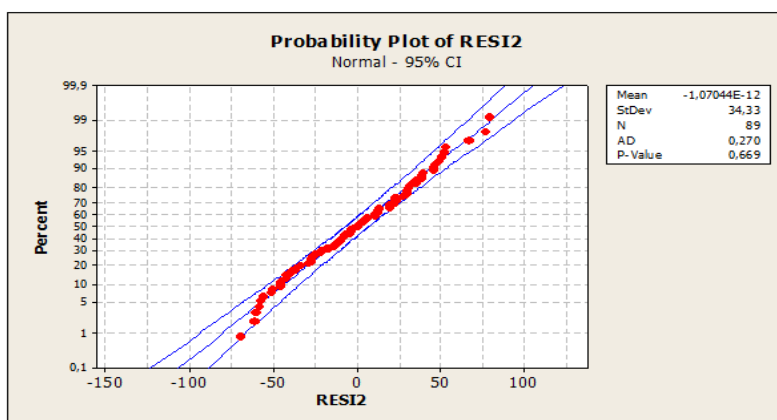
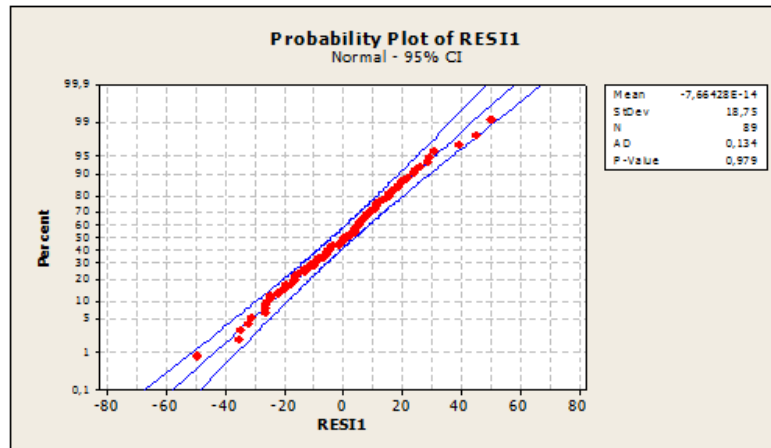


Fig. 1. Structure of pyrazine



(a)



(b)

Fig. 2. Diagram of percentage of normality's of the residues

The advantage large of the Least Absolute Deviation (LAD) method is robustness, i.e., that the estimators are not impact by the extreme values, (they are known as "robust"). It is thus particularly interesting to use the method Least Absolute Deviation LAD if one is in the presence of aberrant values in comparison with Least Squares (OLS) method.

Comparison of Hyperplanes of Regression

The model has been estimated by first by Least Squares (OLS,) and then by Least Absolute Deviation, Running the least squares and Least Absolute Deviation regression yields the estimates given in Table.

Column OV-101

Column CRW -20M

All the variables for the two models is strongly statistically significant in the two columns with method least squares and the method Least Absolute Deviation (Table 4-7).

We noticed that calculated of β least squares are not very different for the regression with β the Least Absolute Deviation on the two columns, except, calculated.

β_1 and β_3 least squares is almost the same ones as for the regression with β_1 and β_3 Least Absolute Deviation on column OV-101 (Table 4-7).

Thus it is relevant to remake a verification in presences of aberrant values using the following phases (Fig. 3):

Hyper plane of regression can radically vary with the change of hyper plane coefficients.

Graphical Comparisons of Alternative Regression Models

The application field has been discussed with the help of Williams diagram.

Column CRW-20M

Column OV-101

The analysis of the residues shows that the observations (82 68 14 1) raised residues in the two estimates and the observations (72, 2) raised residue with the Least Absolute Deviation estimate and lever by least square also observation (2, 4) raised residue and influential observations in the two estimates in the whole of validation on column OV -101 and column CRW -20 M the observations (1, 7, 85) raised residues in the two estimates, the observation (86) raised residues with the Least Absolute Deviation estimate and lever by least square also observation (2,3) raised residues and influential observations with Least Absolute Deviation but it with the least squares estimate the observation (2) influential observation butthe observation (3) lever whole of validation.

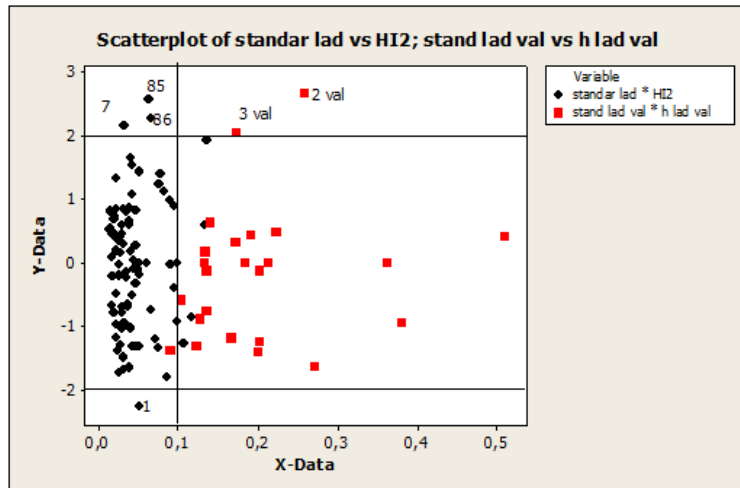
After elimination of the collective aberrant points between the two methods and after the secondary treatment one has the observation (83) raised residues in the two estimates also the observation 2influential observation in the whole of validation in the two estimates on column CRW -20 M and on column OV -101 the observations (1,69) raised residues in the two estimates and the observation 81 the observations raised residues in the least squares estimate also observation (2) influential observation in the least squares estimate.

Thus finally the models in which the meaningless statements were removed become:

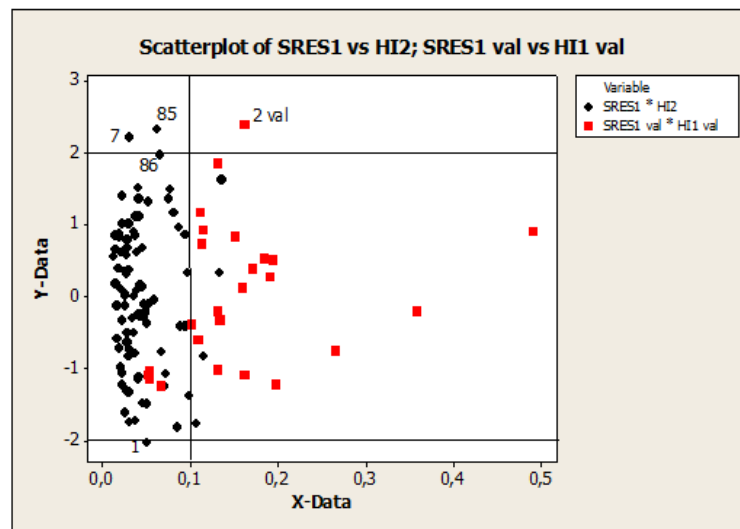
Column OV-101

Least Absolute Deviation:

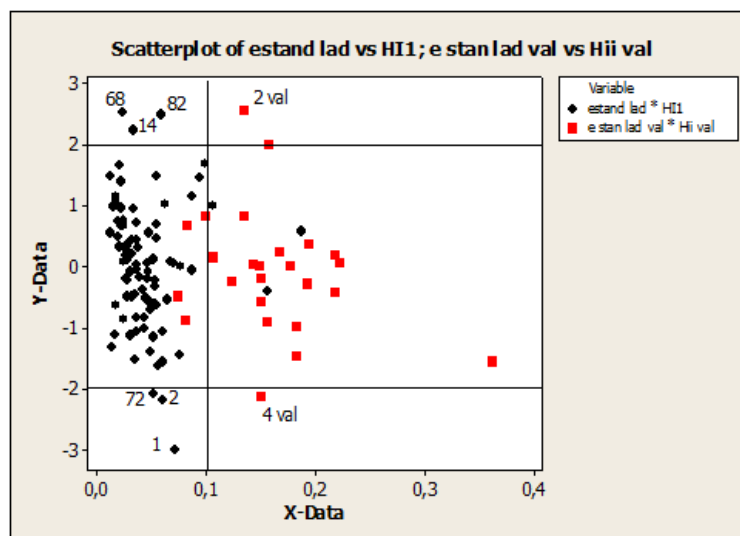
$$y = -946 + 29.1 XMOD + 1174.4 FDI + 70.4 Mor06v \quad (4)$$



(a)



(b)



(c)

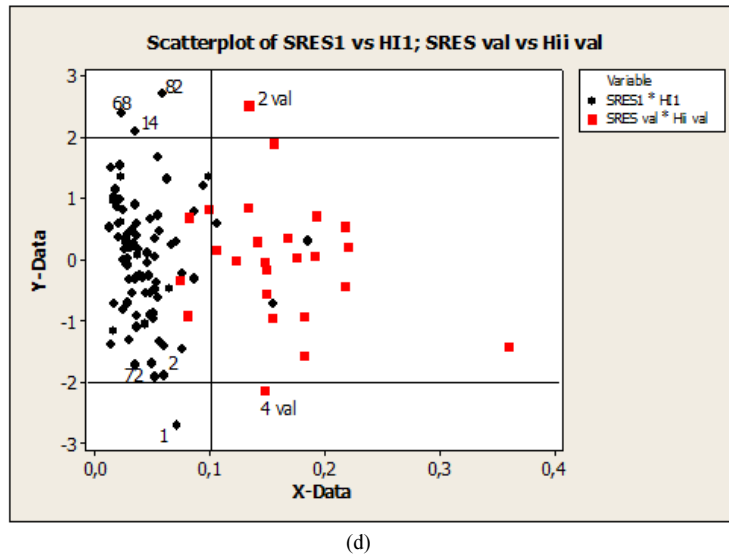


Fig. 3. Diagram of Williams of the residues of prediction standardized according to the lever (a, c) Least absolute deviation method (Training, Test); (b, d) Least squares method (Training, Test)

Least Squares:

$$y = -886 + 29,1 XMOD + 1115 FDI + 70.9 Mor06v \quad (5)$$

Column CW -20M

Least Absolute Deviation:

$$y = -859,72 + 527.46 RDCHI - 630.74 GATS1p + 28.37 Mor02m \quad (6)$$

Least Square:

$$y = 842,527 RDCHI - 625 GATS1p + 29,2 Mor02m \quad (7)$$

We noticed besides that calculated β can approach that regression with β Least Absolute Deviation on the two columns into precise calculated (β_1 and β_3) least squares are almost the same ones as for regression with (β_1 and β_3) Least Absolute Deviation and on the order same with (β_0 and β_2) on OV 101 and calculated β_1 least squares are almost the same ones as for regression with β_1 Least Absolute Deviation on CRW -20 M and on the order same with (β_1 , β_3 and β_4).

The analysis of the residues shows that in this case All the observation of Least Absolute Deviation method between (-2, 2), but it the analysis of the residues of least squares method shows that the observations [OV-101: Training - test (2), CRW-20 M: Training- (46)] the Least Absolute Deviation estimate given good result On the other hand estimate least squares Fig. 4:

Graphical Comparisons of Alternative Regression Models

Column CRW-20M

Column OV-101

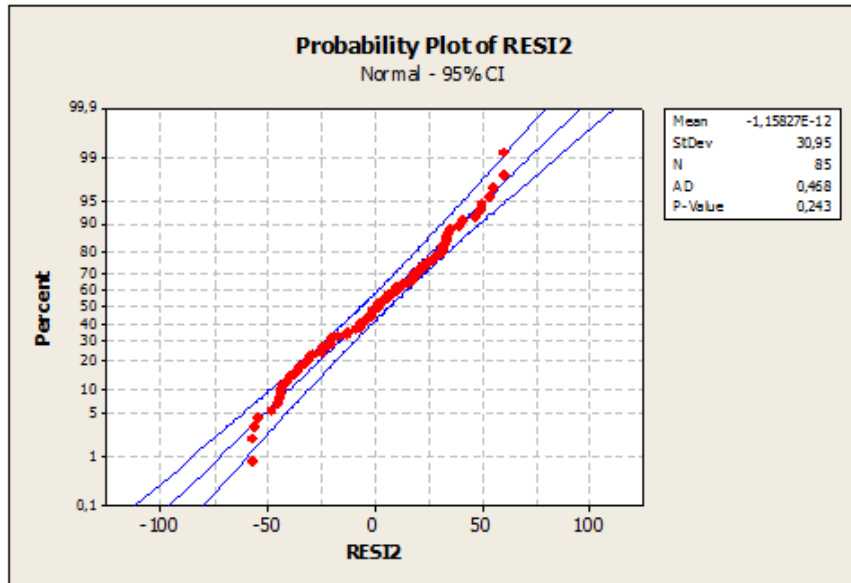
We notice no change of the coefficients of the right-hand side after feeding of the aberrant point what translates the line is stable which expresses that the Least Absolute Deviation method born not sensitive to the presences of the aberrant values thus we report that the Least Absolute Deviation method is a stable method and more robust.

To conform the approach between the two methods and to deduce the robust method between them, There is a set of tests of normality (of standard errors or residues...) indeed, thanks to robustness concept, we can used simple techniques (descriptive e.g. Statistics, technical graphs) to check if the distribution of data is really approximate.

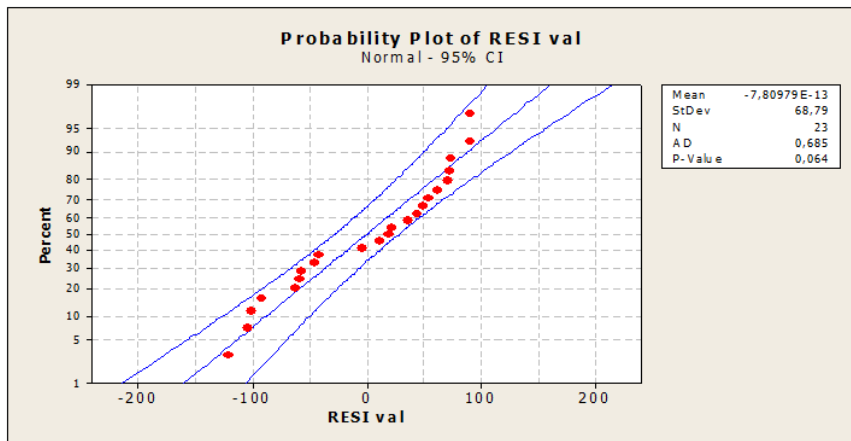
Any test is associated a risk known as of first species years works us, we will adopt it risk $\alpha = 5\%$.

Comparisons of the Tests of Normality of the Errors between Method Least Absolute Deviation and Least Squares in Approached State

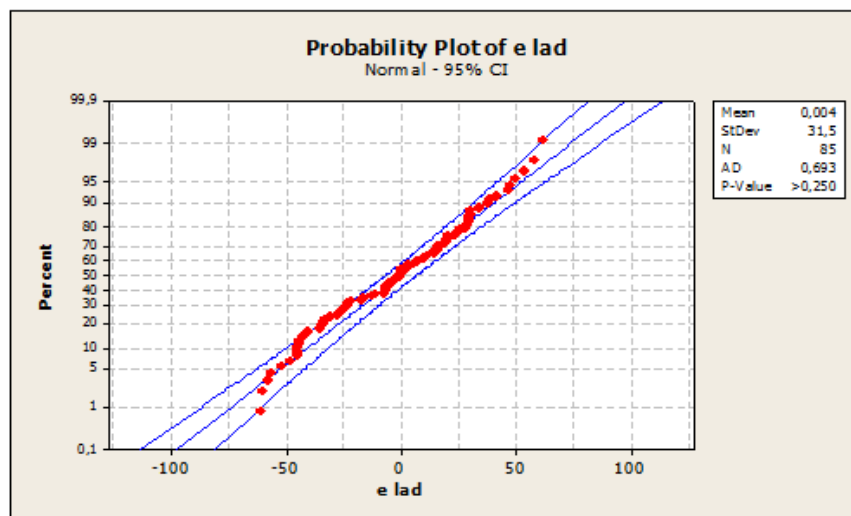
Software Minitab 16 proceeds automatically in estimating two principal parameters of the normal law (μ the Mean (OV-101:0, CRW-20M: 0), σ the variation-type (OV-101:10.35, CRW-20M:14.84) for least squares one applying the same principle with the Least Absolute Deviation method but one used (the median (OV-101: -1.57, CRW-20M:0.01) σ variation-type (OV-101:10.26, CRW-20M:15.08) and with the principal number in the state approached to the two columns (OV-101: n = 83, CRW-20 M: n = 85).



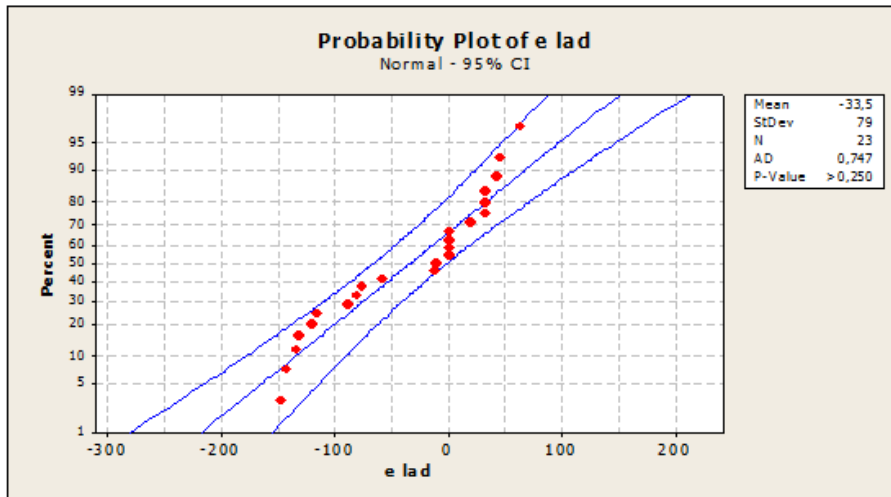
(a)



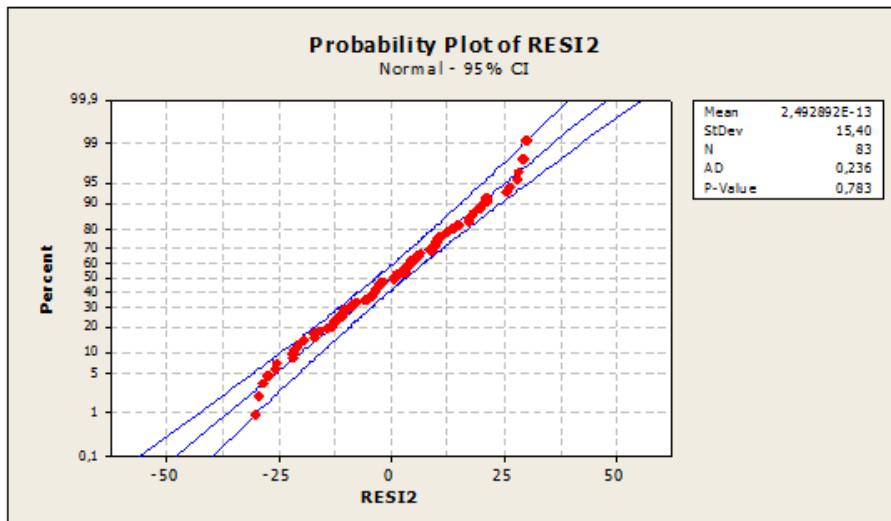
(b)



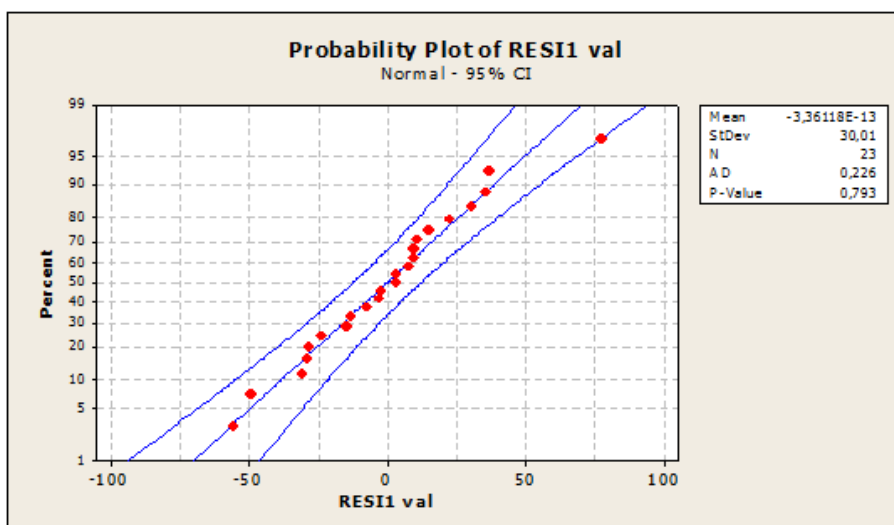
(c)



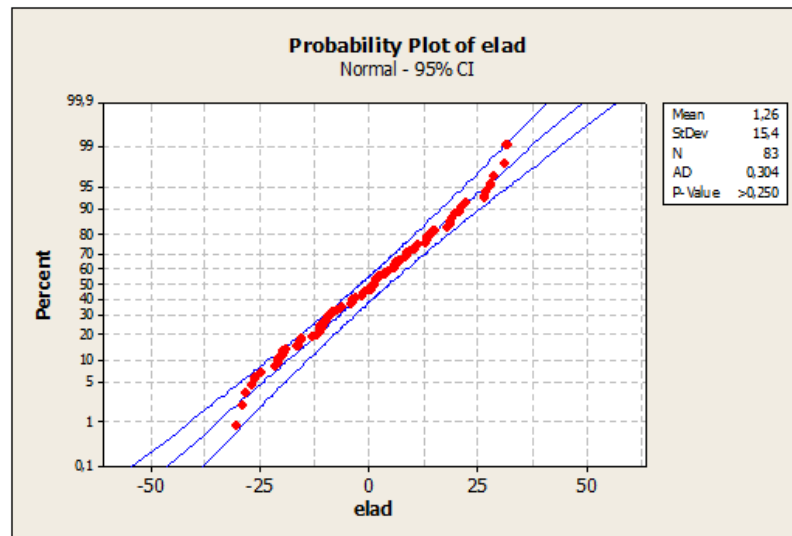
(d)



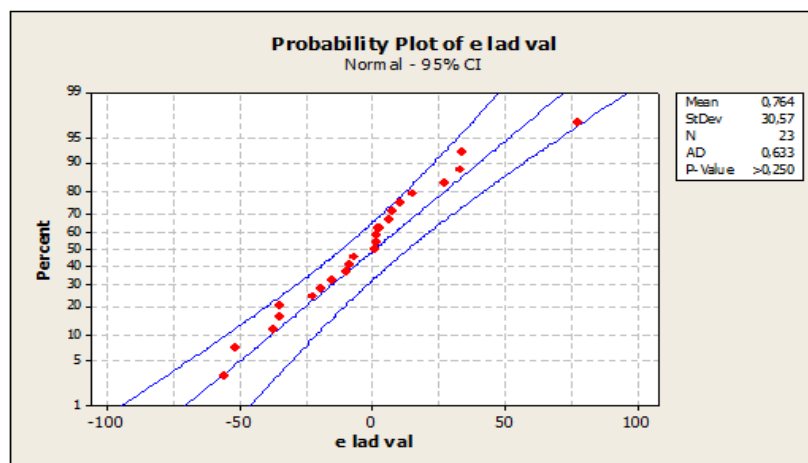
(e)



(f)



(g)



(h)

Fig. 4. Diagram of normality percentage of residues (Training, Test) (a, c, e, g) Training; (b, d, f, h) Test

Graphic Tests

Probability Plot of Error

To check normality of errors of a model of regression is to carry out Probability stud of residues.

Column CW -20M

Least Squares Method

Least Absolute Deviation Method

Column OV-101

Least Squares Method

Least Absolute Deviation Method

A normal distribution with the two columns appears to fit your data sample fairly well.

The plotted points form a reasonably straight line.

Test of Anderson-Darling

In our work, one finds us that Anderson-Darling (AD) [OV- 101: (Least Absolute Deviation) = 0.364 with value of $p > 0.250$, (least squares) = 0.236 with value of $p = 0.783$, $n = 83$], [CRW-20M: (Least Absolute Deviation) Anderson-Darling (AD) = 0,693 with value of $p > 0.250$, (least squares) = 0,468 with value of $p = 0.243$ $n = 85$] $<$ AD critique = 0.752 with $p > 0.1$ to 5%, the assumption of normality is compatible with our data with Least Absolute Deviation method and least squares.

Interval of Confidence

The interval confidence and the risqe a constitute a complementary approach thus (an estimate approach) the most used interval confidence is interval confidence has $100(1-a) = 95\%$.

The Column OV-101:

Training : Least Absolute Deviation: (-31.52, 29), least squares (-30.18, 30.18)

Test : Least Absolute Deviation (-59.15, 60.68), least squares (-58.82, 58.82)

The Column CRW-20M:

Training : Least Absolute Deviation: (-61.73, 61.74), least squares (-60.66, 60.66)

Test : Least Absolute Deviation (-135.9, 135.8), least squares (-136.6, 136.6)

The data may be compatible with the hypothesis also that the limited values of the interval are center which expresses the mean and the median which verifies position 95% that the 50th percentile for the population the center of the acceptance zone the null hypothesis.

Completely all the graphic and statistical tests is accepted data of the approached state between the two methods especially the test of Anderson-Darling the value of the Least Absolute Deviation method closer to least squares method and Interval of The value of confidence these result is formed L approximate of two method.

Conclusion

PYRAZINES are compounds naturally presents in food and taking part in their odour, contrary to their biodegradation, pyrazine formation has been intensively studied.

Modeling of retention indices of 114 pyrazines (89 Training and 25 Test) eluted out of two columns various OV -101, the best tree parametric model was constructed using.

[OV-101 with Modified Randi connectivity index (XMOD), Folding Degree Index (FDI) and (3D-MORSE-signal 06/weighted by atomic Vander Waals volumes (Mor06v); CRW-20M with Reciprocal distance Randi-type Index (RDCHI), Geary autocorrelation -log 1/weighted by atomic polarizabilities (GATS1p) and 3D-MORSE-signal 02/weighted by atomic masses (Mor 02 m)].

The Column of OV-101 and CRW-20M by two methods Least Absolute Deviation and least squares are based on the following comparisons.

The comparison of the equations of the hyper planes:

L'equations of least squares is closer to Least Absolute Deviation after elimination of the aberrant points for the β_2 (Least Absolute Deviation) $\cong \beta_2$ (least squares) and the other coefficient remaining with the same order for column OV-101 for the column CRW-20 M the β_1 (Least Absolute Deviation) $\cong \beta_1$ (least

squares) and the other coefficient remaining with the same order after the secondary treatments for the checking of presence of aberrant values (training: 1, 2, 14, 68, 72, 82 test: 2, 4) (training: 1, 7, 85, 86, test: 2, 3) on column (OV -101) and (training: 1, 7, 85, 86, test: 2, 3) for the CRW-20M- column) and to be able to compare them By using the following stage.

Graphic comparison: The applicability is discussed using the diagram of Williams in dependence.

Lastly, it is noted that Least Absolute Deviation is a robust estimator not sensitive to the presences of the aberrant values thus we report that the Least Absolute Deviation method is a stable and robust method.

Used test of normality's of the errors by graphic and statistical test. One applied compatibility with the normal law, but using the degree $\alpha = 0.05$. Too one confirmed approached graphically by Probability plot of the error One notes that the test to accept the assumption of normality is that of Anderson-Darling, in finished by the confidence interval with one p-been worth sup 0.1 on the columns.

It general this study is shown that results by the two estimates theoretical (equation) and graph give good results expressed by the models.

Acknowledgement

We would like to thank Salima Khanouch, PhD (informatique) for writing the algorithm of least absolute deviation Method and Mr hessene (Teacher of English Language) for reviewing the language in the manuscript.

Author's Contributions

Fatiha Mebarki: Good Developed methods of least absolute deviation and least squares, Developed deference's Softwares (Matlab, Minitab, Tanagra, genetic Algorithm)and participated in all experiments, coordinated the data-analysis.

Khadija Amirat: Developed deference's Softwares (Matlab, Minitab, Tropsha, SVM, genetic Algorithm) and participated in all experiments, coordinated the data-analysis and contributed to the writing of the manuscript.

Salima Ali Mokhnach: Correction of the work and chef laboratories.

Djellol Messadi: Is the main researcher of project and chef of tree laboratories, designed the research plan and organized the study.

Ethics

This article is original and to the best knowledge of the authors has not been published before. The authors confirm that there are no ethical issues involved.

References

- Berlin, G.B., 1982. The Pyrazine. 1st Edn., J. Wiley, New York, ISBN-10: 0471381195, pp: 687.
- Buchbauer, G., 2000. Threshold-based structure-activity relationships of pyrazines with bell-pepper Flavor. *J. Agric. Food Chem.*, 48: 4273-4278. PMID: 10995349
- Damodar, N.G. and C.D. Porter, 2009. Basic Econometrics. 5th Edn., McGraw-Hill Irwin., Boston, ISBN-10: 0071276254, pp: 922.
- Dewar, M.J.S., E.G. Zebisch, E.F. Ealy and J.J.P. Stewart, 1985. AM1: A new general purpose quantum mechanical model. *J. Am. Chem. Soc.*, 107: 3902-3909.
- Dodge, Y. and J. Jureckova, 2000. Adaptive Regression. 1st Edn., Springer Science and Business Media, New York, ISBN-10: 1441987665, pp: 177.
- Dodge, Y., 1987. Statistical Data Analysis Based on the Li-Norm and Related Methods. 1st Edn., North-Holland, Amsterdam, ISBN-10: 0444702733, pp: 464.
- Dodge, Y., 1997. L1-Statistical Procedures and Related Topics. 1st Edn., Institute of Mathematical Statistics, Hayward, ISBN-10: 0940600439, pp: 498
- Dodge, Y., 2004. Statistique: Dictionnaire Encyclopédique. 1st Edn., Springer Science and Business Media, Paris, ISBN-10: 2287720944, pp: 662.
- Dragon 5.4, <http://www.disat.unimib.it>
- Eriksson, L., J. Jaworska, A. Worth, M. Cronin and R.M. Mc Dowell *et al.*, 2003. Methods for reliability, uncertainty assessment and applicability evaluations of regression based and classification QSARs. *Environ. Health Perspect.*, 111: 1361-1375.
- Gonin, R. and A.H. Money, 1989. Linear L_p -norm Estimation. 1st Edn., Marcel Dekker, New York.
- Holder, A.J., 1998. AM1, Encyclopedia of Computational Chemistry. Scheleyer, P.V.R., N.L. Allinger, T. Clarck, J. Gasteiger and P.A. Kollman *et al.* (Eds.), Wiley, Chichester, pp: 1-8.
- Hyperchem 6.03, (Hypercube), <http://www.hyper.com>
- Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.C. Lee, 1985. The Theory and Practice of Econometrics. 2nd Edn, Wiley, New York, ISBN-10: 047189530X, pp: 1019
- Li, W., C.L. Heth and S.C. Rasmussen, 2014. Thieno[3,4-b] pyrazine-based oligothiophenes: Simple models of donor-acceptor polymeric materials. *Phys. Chem. Chimica Phys. J.*, 28 : 7231-40.
- Machabert, T., 2014. Modèles en très grande dimension avec des outliers. Théorie, Simulations, Applications Paris.
- Matlab, R., 2009. Minitab, release 16.1, statistical software, 2003.
- Mebarki, F., K. Amirat, S.A. Mokhnache and D. Messadi, 2016. Treatment by alternative methods of regression gas chromatographic retention indices of 35 pyrazines. *Int. J. Instrument. Control Syst.*, 6: 1-14.
- Mihara, S. and N. Enomoto, 1985. Calculation of retention indices of pyrazines on the basis of molecular structure. *J. Chromatogr.*, 324: 428-430.
- Moby Digs 1.1, <http://www.disat.unimib.it>
- Nornadiah, M.R. and Y.B. Yah, 2011. Power Comparisons of shapiro-wilk, Kolmogorov-smornov, lilliefors and Anderson-Darling tests. *J. Statistique Modell. Analyt.*, 2: 21-33.
- Small, G.W. and P.C. Jurs, 1983. Interactive computer system for the simulation of carbon-13 nuclear magnetic resonance spectra. *Anal. Chem.*, 55: 1121-1127. DOI: 10.1021/ac00258a033
- Stanton, D.T. and P.C. Jurs, 1989. Computer-assisted predict of gaschromatographicretention indexes of pyrazines. *Anal. Chem.*, 61: 1328-1332.
- Todeschini, R., D. Ballabio, V. Consonni, A. Mauri and V. Pavan, 2009. MobyDigs 1.1, Copyright TALETE srl.
- Tropsha, A., P. Gramatica and V.K. Grombar, 2003. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Combi. Sci.*, 22: 69-76.