وزارة التعليم العالى و البحث العلمى

Université BADJI Mokhtar – Annaba
BADJI Mokhtar – Annaba University

جـامـعـة بـاجـي مـخـتـار – عـنـابـة

Faculté des Sciences
Département de Chimie

MÉMOIRE

Présenté pour l'obtention du diplôme de Magistère en chimie Analytique

Par M. Hicham LAHMAR

Option : Chimie de l'environnement

THÈME

MODÈLE LINÉAIRE ET NON LINÉAIRE POUR L'ÉTUDE DES INDICES DE RÉTENTION DES PHÉNOLS EN CHROMATOGRAPHIE GAZEUSE A TEMPERATURE PROGRAMMEE

Devant le jury :

Président :	**M^{ME}. S. ALI-MOKHNACHE**	**Professeur**	**U. B. M. Annaba**
Rapporteur :	**M. A. TOUBAL**	**M. C.**	**U. B. M. Annaba**
Examinateurs :	**M.A. DJELLAL**	**M. C**	**U. B. M. Annaba**
	M. D. MESSADI	**Professeur**	**U. B. M. Annaba**
INVITEE	**M^{ELLE}.I.TOUHAMI**	**C.C**	**U. B. M. Annaba**

Année 2009

Dédicace

Je dédie ce modeste travail à :

- *La mémoire de mon père*
- *Ma mère*
- *Ma famille*
- *Mes amis*
- *Enfin, toute l'équipe du labo 34.*

H-Lahmar

REMERCIEMENTS

Ce mémoire n'aurait pas vu le jour
sans la confiance, la patience et la générosité du responsable de la P.G. Monsieur
le Professeur D. MESSADI que je remercie vivement. Je voudrais aussi le remercier pour le temps et la patience qu'il ma accordés tout au long de ces années et pour avoir accepté d'examiner ce travail.

Je tiens également à remercier :

Mr TOUBAl .A , pour la direction de ce travail ;

Mme ALI MOKHNACHE Salima, pour avoir accepté la présidence de ce jury ;

Mr DJELLAL .A , pour avoir accepté d'examiner ce travail.

M^{elle} TOUHAMI .I, pour avoir accepté l'invitation.

Enfin, je ne saurais oublier mes camarades de laboratoire et également tous ceux qui par leur présence ou par leur aide m'ont permis de mener à bien ce travail.

SOMMAIRE

PARTIE EXPERIMENTALE

Résumé:

Deux modèles QSRR ont été développés pour la prédiction de l'indice de rétention. Les données, concernant 56 composés dérivés du phénol , ont été séparées en deux sous-ensembles disjoints comprenant respectivement 44 éléments pour le calcul et le test (éventuel) du modèle, et 12 éléments pour sa validation externe. Deux modèles ont ainsi été créés sur le même ensemble de données: un modèle de régression multilinéaire et un modèle de réseaux de neurones artificiels.

Des descripteurs moléculaires théoriques ont été calculés en utilisant des logiciels de modélisation moléculaire du commerce. La taille du modèle à été déterminée en optimisant le FIT de KUBINYI, et la sélection des descripteurs réalisée par algorithme génétique.

Les valeurs des paramètres statistiques (R^2, Q^2, EQMC, EQMCP, EQMCPext) obtenues attestent de la pertinence des modèles développés, avec une supériorité établie pour les modèles de neurones artificiels.

Mots-clés:
Chromatohraphie en phase gazeuse – Dérivés du phénol – Indice de rétentien – Descripteurs moléculaires théoriques – Modèles hybrides.

ABSTRACT

Two QSRR models were developed for the prediction of retention index. A dataset of 56 compounds derivatives of phenols was subdivised into two disjoined subsets containing respectively 44 compounds for calculating and (possible) testing of the model, and 12 compounds used in external validation. Two models based on the same subset of data were thus created : a multiple linear regression model and an artificial neural network model.

Theoretical molecular descriptors were calculated using commercially available molecular modelling softwares. The model size was determined by optimizing the FIT of KUBINYI, and the selection of the descriptors realized by genetic algorithm.

Values obtained for the statistical parameters: R^2, Q^2, SDEC, SDEP and $SDEP_{ext}$, attest relevance of the models developed, with a clear superiority for the artificial neural network models.

Key words :

Gaz chromatogrphiy –Derivatives of phenol –Retention index – Theoretical molecular Descriptors – Hybrid models.

مـــلـــخـــص:

تم تطوير نموذجين بطريقة الـــ QSRR للتنبؤ بمؤشر اللأ ستبقاء.

المعطيات الخاصة بـــ ٥٦ مركب من مشتقات الفينول تم تقسيمها الى مجموعتين الاولى تحتوي ٤٤ عنصر لحـــساب و تجريب النموذج اما الثانية تحتوي على ١٢ عنصر للتصديق الخارجي للنموذج.

النموذجين المتحصل عليهما لنفس المعطيات هما : نموذج التراجع المتعـــدد الخطـــي و نمـــوذج الـــشبكة العـــصبونية الاصطناعية.

المواصفات الجزيئية النظرية تم حسابها باستعمال برمجيات النمذجة الجزيئية المتوفرة في السوق . حجم النمـــوذج تـــم تحديده عن طريق دالة الـــ FIT لـــ Kubinyi، اما اختيار المواصفات عن طريق الخوارزمية المورثية .

قيم المعالم الاحصائية ($SDEP_{ext}$, SDEP , SDEC , Q^2, R^2) المتحصل عليها تؤكد تعلق النماذج المطورة مع تفوق معتبر لنموذج الشبكة العصبونية الاصطناعي.

الكلمات الدالة:

الكروماتوغرافيا الغازية – مشتقات الفينول– مؤشر اللأ ستبقاء – مواصفات جزيئية نظرية – نماذج هجينة.

LISTE DES TABLEAUX

LISTE DES FIGURES

SYMBOLES ET ABREVIATIONS

AM1 :	Austin Model 1.
DFITS :	Statistique permettant de mesurer l'influence d'une observation i sur la valeur ajustée.
Di :	Distance de COOK.
d :	Statistique de Durbin-Watson.
di :	Résidu standardisé.
EQM:	Erreur quadratique moyenne.
EQMP:	Erreur quadratique moyenne sur l'ensemble de prédiction.
$EQMP_{ext.}$:	Erreur quadratique moyenne sur l'ensemble de prédiction externe.
e_i :	Résidu différence entre les valeurs observée (y_i) et estimée ($\hat{y}_i$).
F :	Statistique de Fisher.
FIT:	Fonction de KUBINYI.
GA:	Algorithme génétique (Genetic Algorithm).
$\underset{\sim}{\mathbf{H}}$ **:**	Matrice de projection, ou matrice châpeau.
hii :	Eléments diagonaux de la matrice châpeau.
IR:	indice de rétention.
IW :	Poids entrée-couche cachée .
LOO:	Cross-validation by leave-one-out: Validation croisée par omission d'une observation
LW :	Poids couche cachée-sortie.
MLR:	Régression linéaire multiple.
MCP:	Moindres carrés partiels.
n:	Dimension de la population (échantillon).
n-p :	Nombre de degrés de liberté.
PMC:	Réseaux multicouches.

PRESS : Somme des carrés des erreurs de prédiction.

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

p_c : Probabilité de croisement.

p_M: Probabilité de mutation.

QSRR : Quantitative Structure/ Retentions Relationships.
(Relations Structure/ Rétention Quantitatives).

Q^2_{LOO} : Coefficient de prédiction.

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

RNA: Réseaux de neurones artificiels.

R^2 : Coefficient de détermination.

r_i : Résidu studentisé interne.

S : Erreur standard.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

t : t de Student.

t_i : Résidu studentisé externe.

w_k : Poids à l'instant k.

w_{k+1} : Poids à l'instant k-1.

$\underset{\sim}{X}$ **:** Matrice des valeurs observées des variables explicatives.

$\underset{\sim}{X}'$ **:** Matrice transposée de $\underset{\sim}{X}$.

$\underset{-}{x_j}$: Variable explicative.

x_j : $j^{ième}$ valeur de x .

$x_{\max}$: Valeur maximale.

$x_{\min}$: Valeur minimale.

x_{norm} **:** Valeur normalisée.

$\underset{-}{y}$: Vecteur de dimension n.

y_i : Valeur observée.

$\hat{y}_i$: Valeur estimée.

α : Niveau de confiance; Facteur d'apprentissage.

σ^2 **:** Variance.

δ_k : Différence entre la sortie attendue et la sortie effective à l'instant k.

INTRODUCTION GENERALE

INTRODUCTION GENERALE

Les phénols sont des composés importants biologiquement et du point de vue de l'environnement. Non seulement ils possèdent d'importantes fonctions physiologiques et certaines activités pharmaceutiques, mais ils peuvent encore influencer la saveur de certaines boissons. D'-autres composés phénoliques sont utilisés pour le tannage, en cosmétique, dans l'industrie organique (fabrication de matières plastiques, produits pharmaceutiques, explosifs ……), ainsi que pour le développement photo, ce qui en fait d'importants polluants potentiels de l'environnement .

Origine

Ce sont des alcools aromatiques qui proviennent des végétaux. Les phénols simples, déchets du métabolisme végétal, sont assemblés en polyphénols comme la lignine. Les composés phénoliques définissent un ensemble de substances que l'on a appelées pendant longtemps " matières tannoïques " d'une façon générale et imprécise parce qu'on ne connaissait pas, avec suffisamment de précision, la nature de ces substances. Il y a quatre principales familles de composés phénoliques : les acides-phénols, les flavones, les anthocyanes, les tanins.

La lignine est un des principaux composants du bois, avec la cellulose. C'est un groupe de composés chimiques appartenant aux composés phénoliques (il existe donc plusieurs types de lignine). On la trouve principalement dans les parois pectocellulosiques de certaines cellules végétales.

Les anthocyanes ou anthocyanines (du grec anthos = fleur, kuanos = bleu sombre) sont des pigments naturels solubles dans l'eau allant du rouge au bleu dans le spectre visible. Ils appartiennent à la classe des composés nommés flavonoïdes.

Les anthocyanines sont présents dans un certain nombre de végétaux tels : myrtille, mûre, raisin noir .

Propriétés acido-basiques

Les phénols sont plus acides que les alcools.Un ion phénolate est stabilisé par résonance et est plus stable qu'un ion alcoolate. En effet, lors de la prise du proton du groupement hydroxyle, le doublet électronique est partagé sur quatre carbones; ainsi, la charge est

délocalisée sur autant de carbones et l'ion est beaucoup plus stable que sur un alcool où la charge négative serait trop importante et s'approprierait le proton laissé immédiatement après.

Cet acide est toutefois un acide relativement faible; en conséquence, sa base conjuguée, l'ion phénolate, est une base très forte.

Les techniques les plus courantes pour établir des modèles QSRR utilisent l'analyse de régression (régression multilinéaire : MLR; projection des structures latentes par les moindres carrés partiels : PLS), les réseaux neuronaux RNA, et les méthodes de classification.

Des logiciels informatiques spécialisés permettent le calcul de plus de 3000 descripteurs moléculaires appartenant à différentes classes. Plutôt que de rechercher à expliquer la variable dépendante (propriété physique considérée) par tous les régresseurs (descripteurs moléculaires), on peut chercher seulement un ensemble réduit de régresseurs qui donne une reconstitution aussi satisfaisante de la variable à expliquer. Parmi les stratégies mises en œuvre pour la sélection d'un ensemble limité de variables explicatives, on peut citer : les méthodes de pas à pas, ainsi que les algorithmes évolutifs et génétiques.

Nous avons appliqué des méthodes hybrides: algorithme génétique/régression multilinéaire (GA/MLR), et algorithme génétique/réseaux de neurones artificiels (GA/RNA) pour modéliser, l'indice de rétention de 56 composés de phénols.

Notre mémoire comporte en plus de la bibliographie, d'une introduction et d'une conclusion générales, deux grandes parties :

Dans la Partie Généralités, nous avons développé tout ce qui a trait au pré-traitement des molécules (introduction des molécules, optimisation de leur géométrie) en vue du calcul des descripteurs moléculaires à l'aide de différents logiciels de modélisation moléculaire. Nous y avons également développé les connaissances théoriques de base utilisées tout au long de ce travail: algorithmes génétiques, régression multilinéaire, réseaux de neurones artificiels, traitement statistique pour l'évaluation de la qualité de l'ajustement (robustesse des modèles;détection des observations aberrantes;test de randomisation;validation externe).

Dans la Partie Exprimentale, nous nous intéresserons à quelques phénols diversement substitués, analysés par CPG avec une phase stationnaire superox-20M, Dans ce cas, un modèle de rétention du soluté est important pour l'optimisation chromatographique. nous présentons et discutons les modèles calculés :

GA/MLR, GA/RNA, pour l'indice de rétention ;

Et la comparaison des résultats trouvé par les deuxs méthodes .

PARTIE THEORIQUE

Les relations structures / rétention quantitatives, désignées par l'abréviation QSRR (Quantitative Structure/ Retentions Relationships), constituent des modèles mathématiques pour l'approximation des relations, souvent complexes, entre la structure caractérisé par des descripteurs moléculaires et les indices de rétention des composés.

La méthodologie QSRR traite des rapports quantitatifs entre la structure moléculaire et l'indice de rétention dans différents types de chromatographie [1], QSRR est un sous-groupe de modelage QSPR qui traite des rapports quantitatifs entre la structure moléculaire et les propriétés physico-chimiques des composés.

Une révision de ces modèles a été donnée dans les publications [2,3].

Le but fondamental des processus QSRR est d'étudier le rapport entre une variable dépendante et une ou plusieurs variables indépendantes, y compris l'analyse et l'interprétation des fonctions de la corrélation obtenues. Temps de rétention et les volumes, dans tout leur formes [4], l'indice de Kovats [5], l'élément QSRR est considéré comme variable dépendant… Sa mise en œuvre comprend plusieurs étapes:

- La collecte des données ;
- L'optimisation de la géométrie moléculaire ;
- La génération des descripteurs moléculaires ;
- La sélection d'un sous-ensemble de descripteurs significatifs ;
- Le développement du modèle ;
- Et, finalement, l'évaluation des performances de ce modèle.

I – COLLECTE DES DONNEES

On a pris dans notre travail l'indice de rétention trouvé par programmation de température en chromatographie en phase gazeuse dans la perspective de l'analyse chimique. Les données utilisées dans ce travail, prélevées dans [6], concernent 56 composés de phénols substitués ; elles sont réunies dans le tableau I.

Les 56 observations de ce tableau ont été scindées aléatoirement en deux ensembles disjoints de 44 éléments [ensemble d'estimation qui sert à la construction du modèle (32 éléments) et pour le test (12 éléments =44-32)], et de 12 éléments [pour la validation externe]. Les éléments de l'ensemble d'estimation sont numérotés de 1 à 44, et ceux de l'ensemble de validation externe portent les numéros 45 à 56 dans le tableau I.

Tableau I : **Nomenclature et valeurs des indices de rétention étudiés [6]**

N°	Composés	IR	N° CAS
1	2,6-dimethylphenol	78.17	[576-26-1]
2	2,6-di-tert-butylphenol	86.83	[128-39-2]
3*	phenol	100	[108-95-2]
4	o-cresol	101.39	[95-48-7]
5	2,3,6-trimethylphenol	110.55	[2416-94-6]
6*	2,4,6-tri-tert-pentylphenol	117.93	/
7	(R)-2,3-dihydro-1H-inden-1-ol	117.93	[36643-74-0]
8	2-ethylphenol	120.1	[90-00-6]
9*	2,4-dimethylphenol	121.72	[105-67-9]
10	m-cresol	122.32	[108-39-4]
11	2-isopropylphenol	133.27	[88-69-7]
12*	2,3-dihydro-1H-inden-2-ol	130.05	[4254-29-9]
13	4-ethylphenol	143.7	[123-07-9]
14	3,5-dimethylphenol	143.7	[95-87-4]
15*	3-ethylphenol	145.44	[620-17-7]
16	2,3,5,6-tetramethylphenol	145.44	[527-35-5]
17	(R)-2-sec-butylphenol	149.44	[89-72-5]
18	2-tert-butylphenol	151.61	[98-54-4]
19	2-isopropyl-5-méthylphenol	151.61	[89-83-8]
20*	(R)-1,2,3,4-tetrahydronaphthalen-1-ol	147.03	[529-33-9]
21	4-isopropylphenol	156.11	[99-89-8]
22	2,3,5-trimethylphenol	156.63	[697-82-5]
23	2-tert-butyl-4-methylphenol	161.59	[2409-55-4]
24*	4-propylphenol	166.14	[645-56-7]
25	4-tert-butylphenol	172.85	[98-54-4]
26	(R)-chroman-4-ol	180.6	[1481-93-2]
27	3,4,5-trimethylphenol	187.54	[527-54-8]
28*	2,3-dihydro-1H-inden-4-ol	186.72	[1641-41-4]
29	4-tert-pentylphenol	200	[80-46-6]
30	2,3,4,5-tetramethylphenol	202.38	[488-70-0]
31	2,3-dihydro-1H-inden-5-ol	203.74	[1470-94-6]
32*	6-methyl-2,3-dihydro-1H-inden-4-ol	206.06	[20294-32-0]
33	5,6,7,8-tetrahydronaphthalen-1-ol	221.82	[529-35-1]
34	7-methyl-2,3-dihydro-1H-inden-5-ol	229.23	/
35	5,6,7,8-tetrahydronaphthalen-2-ol	241.62	[1125-78-6]
36*	2-cyclohexylphenol	250.74	[119-42-6]
37	biphenyl-3-ol	354.07	[580-51-8]
38	biphenyl-4-ol	359.42	[92-69-3]
39*	2,5-dimethylphenol	121.01	[95-87-4]
40	3,4-dimethylphenol	153.61	[95-65-8]
41	3-isopropylphenol	156.85	[618-45-1]
42*	2-methylnaphthalen-1-ol	271.15	[7469-77-4]
43	naphthalen-2-ol	309.55	[875-83-2]
44	5-isopropyl-2-methylphenol	156.11	[499-75-2]

* composés de test éventuels ** composés de validation externe

Tableau I Suite et fin

N°	Composés	IR	N° CAS
45**	2,4,6-trimethylphenol	102.12	[527-60-6]
46**	p-cresol	121.01	[106-44-5]
47**	2,3-dimethylphenol	137.68	[526-75-0]
48**	(R)-4-sec-butylphenol	176.68	[99-71-8]
49**	3,5-diisopropylphenol	197.07	[26886-05-5]
50**	7-methyl-2,3-dihydro-1H-inden-4-ol	211.01	/
51**	benzo[d][1,3]dioxol-5-ol	251.61	[533-31-3]
52**	2-tert-butyl-6-methylphenol	106.61	[96-65-1]
53**	3-tert-butylphenol	173.57	[585-34-2]
54**	naphthalen-1-ol	300	[90-15-3]
55**	2-propylphenol	139.16	[644-35-9]
56**	biphenyl-2-ol	240.05	[90-43-7]

* composés de test éventuels ** composés de validation externe

II-OPTIMISATION DE LA GEOMETRIE MOLECULAIRE ET GENERATION DES DESCRIPTEURS MOLECULAIRES

Les structures des molécules ont été obtenues à l'aide du logiciel de modélisation moléculaire Hyperchem 7.5 [7], et les géométries finales à l'aide de la méthode semi empirique AM1 du même logiciel. 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 0,001 kcal.mol^{-1}. Les géométries obtenues ont été transférées dans les logiciels informatiques [7-8] utilisés pour le calcul de plus de 1700 descripteurs appartenant à 20 classes différentes.

Le logiciel Ecalc [9] est muni d'une interface graphique qui permet à l'utilisateur d'introduire les molécules, puis de calculer les indices électrotopologiques.

III- SELECTION D'UN SOUS-ENSEMBLE DE DESCRIPTEURS SIGNIFICATIFS

Des logiciels informatiques spécialisés permettent le calcul de plus de 3000 descripteurs moléculaires appartenant à différentes classes. Plutôt que de rechercher à expliquer la variable dépendante (propriété physique considérée) par tous les régresseurs (descripteurs moléculaires), on peut chercher seulement un ensemble réduit de régresseurs 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 évolutifs et génétiques.

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

III-1 Principe

Dans la terminologie des algorithmes génétiques, le vecteur binaire $\underset{\sim}{I}$, 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.

III – 2 Initialisation aléatoire du modèle

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

III – 3 Etape de croisement

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

III – 4 Etape de mutation

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

Après obtention du modèle transformé, on en calcule le paramètre statistique : si cette valeur est meilleure que la plus mauvaise de la population, le modèle est intégré à la population, à la place correspondant à son rang ; dans le cas contraire il n'est pas pris en compte.

Cette procédure est répétée pour tous les chromosomes, c'est-à-dire P fois.

III – 5 Conditions d'arrêt

Les étapes 2 et 3 sont répétées jusqu'à la rencontre d'une condition d'arrêt (par exemple un nombre maximum d'itérations défini par l'utilisateur), ou qu'il est mis fin arbitrairement au processus.

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

Notons que le taille du modèle est fixée par la valeur optimale de la fonction FIT de KUBINYI [10], calculée selon :

$$FIT = \frac{R^2}{1 - R^2} \frac{n - p - 1}{(n + p)^2} \qquad (1)$$

p désignant le nombre de variables du modèle et R^2 le coefficient de détermination.
Ce critère permet de comparer entre modèles construits sur un même nombre n de données, mais avec un nombre de variable p différent.

IV – DEVELOPPEMENT DES MODELES

Les techniques les plus courantes pour établir des modèles QSRR utilisent l'analyse de régression (régression linéaire multiple : MLR ; projection des structures latentes par les moindres carrés partiels : PLS), les réseaux neuronaux, et les méthodes de classification.

Nous avons utilisé la MLR et les réseaux de neurones artificiels (RNA). En imposant des transformations linéaires entre descripteurs moléculaires et propriétés étudiées, la MLR peut influencer négativement les capacités prédictives du modèle. Par contre, avec les réseaux de neurones il n'est nul besoin de postuler un modèle. Les réseaux de neurones ont la capacité de représenter n'importe quelle dépendance fonctionnelle qu'ils découvrent par eux-mêmes. Ainsi, la découverte et l'exploitation des dépendances non-linéaires de haut niveau peuvent améliorer la capacité de prédiction de la variable d'intérêt.

IV – 1 La régression linéaire multiple (MLR)

Supposons qu'on ait mesuré sur n individus (k+1) variables représentées par des vecteurs de $\Re^n$: $\underset{\sim}{y}, \underset{\sim}{x}_1, \underset{\sim}{x}_2, ..., \underset{\sim}{x}_k$; $\underset{\sim}{y}$ est la variable dépendante ou à expliquer (propriété physique d'intérêt) et les $\underset{\sim}{x}_j$ les variables explicatives ou encore prédicteurs (descripteurs moléculaires). On cherche alors à reconstruire $\underset{\sim}{y}$ au moyen des $\underset{\sim}{x}_j$ par une formule linéaire.

On pose :

$$\underset{\sim}{y} = \beta_0 \underset{\sim}{1} + \underset{\sim}{\mathbf{X}}(j)\,\beta(j) + \varepsilon(j) \qquad (2)$$

$\underset{\sim}{y}$ est un vecteur de dimension n contenant la propriété physique d'intérêt des hydrocarbures considérés, $\underset{\sim}{1}$ est un vecteur unité, c'est-à-dire une matrice colonne formée d'éléments égaux à 1, $\underset{\sim}{X}(j)$ indique la matrice (n×j), et $\varepsilon(j)$ correspond aux résidus qui doivent suivre une distribution Normale, posséder une espérance mathématique nulle et une matrice de dispersion I σ^2 [11]. Les estimateurs {β} sont calculés en utilisant la technique des moindres carrés ordinaires.

IV – 2 Les réseaux de neurones

Les réseaux de neurones ont été étudiés depuis les années 40 [12]. Les idées de base de cette technique viennent de la recherche cognitive, d'où vient le nom 'réseaux de neurones'.

La technique inspirait beaucoup de chercheurs à cette époque, mais beaucoup de l'intérêt disparaît après un article de Minsky et Papert [13], finalement relancée au début des années 80 après un quasi-oubli d'une vingtaine d'années. La cause de l'intérêt soudain était l'apparition de nouvelles architectures de réseaux de neurones.

IV – 2 -1 Le neurone artificiel :

L'élément de base d'un réseau de neurones est, bien entendu, le neurone artificiel. Un neurone (figure 1) contient deux éléments principaux :

- Un ensemble de poids associés aux connexions du neurone, et
- Une fonction d'activation (Figure 2).

Les valeurs d'entrée sont multipliées par leur poids correspondant et additionnées pour obtenir la somme S.

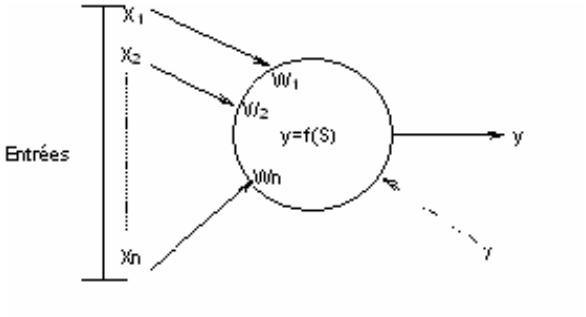
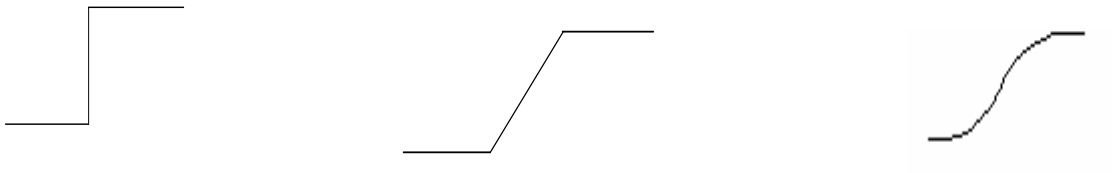


Figure – 1 *le neurone artificiel générique.*

Cette somme devient l'argument de la fonction d'activation, qui est le plus souvent d'une des formes présentées ci- dessous. Une fonction d'activation importante est la simple multiplication avec un, c'est-à-dire que la sortie est simplement une somme pondérée.



Fonction à seuil **Fonction à saturation** **Fonction sigmoïde**

Figure – 2 *Fonctions d'activation.*

Le choix de la fonction d'activation dépend de l'application. S'il faut avoir des sorties binaires c'est la première fonction que l'on choisit habituellement.

Une entrée spéciale est pratiquement toujours introduite pour chaque neurone. Cette entrée, normalement appelée biais (bias en anglais), sert pour déplacer le pas de la fonction d'activation sur l'axe S. La valeur de cette entrée est toujours 1 et le déplacement dépend alors seulement du poids de cette entrée spéciale.

IV – 2 - 2. Propriétés des réseaux de neurones :

Un réseau de neurones se compose de neurones qui sont interconnectés de façon à ce que la sortie d'un neurone puisse être l'entrée d'un ou plusieurs autres neurones. Ensuite il y a des entrées de l'extérieur et des sorties vers l'extérieur [14].

Rumelbart et al. donnent huit composants principaux d'un réseau de neurones [14] :

- Un ensemble de neurones.
- Un état d'activation pour chaque neurone (actif, inactif,…).
- Une fonction de sortie pour chaque neurone (f(S)).
- Un modèle de connectivité entre les neurones (chaque neurone est connecté à tous les autres, par exemple).
- Une règle de propagation pour propager les valeurs d'entrée à travers le réseau vers les sorties.
- Une règle d'activation pour combiner les entrées d'un neurone (très souvent une somme pondérée).
- Une règle d'apprentissage.
- Un environnement d'opération (le système d'exploitation, par exemple).

Le comportement d'un réseau et les possibilités d'application dépendent complètement de ces huit facteurs et le changement d'un seul d'entre eux peut changer le comportement du réseau complètement.

Les réseaux de neurones sont souvent appelés des "boites noires" car la fonction mathématique qui est représentée devient vite trop complexe pour l'analyser et la comprendre directement. Cela est notamment le cas si le réseau développe des représentations distribuées [14], c'est-à-dire que plusieurs neurones sont plus ou moins actifs et contribuent à une décision. Une autre possibilité est d'avoir des représentations localisées, ce qui permet d'identifier le rôle de chaque neurone plus facilement. Les réseaux de neurones ont quand même une tendance à produire des représentations distribuées.

IV – 2 -3.Les différents types de réseaux de neurones

Plusieurs types de réseaux de neurones ont été développés qui ont des domaines d'application souvent très variés. Notamment quatre types de réseaux sont bien connus :

- Le réseau de Hopfield (et sa version incluant l'apprentissage, la machine de Boltzmann).
- Les cartes auto-organisatrices de Kohonen .
- Les réseaux à fonction radiale que l'on nomme aussi RBF (pour " Radial Basic Functions ").
- Les réseaux multicouches ou perceptron multicouches PMC

Le réseau de Hopfield [15] est un réseau avec des sorties binaires où tous les neurones sont interconnectés avec des poids symétriques. C'est-à-dire que le poids du neurone N_i au neurone N_j est égal au poids du neurone N_j au neurone N_i. Les poids sont donnés par l'utilisateur. Les poids et les états des neurones permettent de définir l'«énergie» du réseau.

C'est cette énergie que le réseau tente de minimiser pour trouver une solution. La machine de Boltzmann est en principe un réseau de Hopfield, mais qui permet l'apprentissage grâce à la minimisation de cette énergie.

Les cartes auto-organisatrices de Kohonen [16] sont utilisées pour faire des classifications automatiques des vecteurs d'entrées.

Les réseaux à fonction radiale sont des réseaux multicouches, à une couche cachée. Cependant, contrairement aux perceptrons multicouches, les fonctions de transfert de la couche cachée dépendent de la distance entre le vecteur d'entrée et le vecteur centre.

Les réseaux multicouches (PMC) sont les réseaux les plus puissants des réseaux de neurones qui utilisent l'apprentissage supervisé.

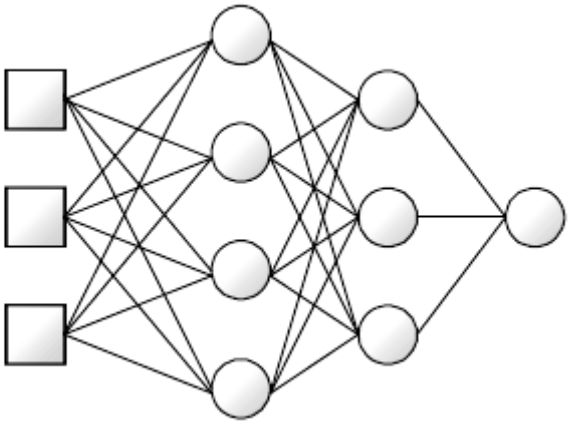
IV – 2 -.3-1 Les réseaux multicouches ou perceptron multicouches (PMC)

Les réseaux multicouches (PMC) (figure 3) se composent des entrées, une couche de sortie et zéro ou plusieurs couche cachées [14]. Les connexions sont permises seulement d'une couche inférieure (plus proche des entrées) vers une couche supérieure (plus proche de la couche de sortie). Il est aussi interdit d'avoir des connexions entre des neurones de la même couche.

Les entrées servent à distribuer les valeurs d'entrée aux neurones des couches supérieures, éventuellement multipliées ou modifiées d'une façon ou d'une autre.

La couche de sortie se compose normalement des neurones linéaires qui calculent seulement une somme pondérée de toutes ses entrées.

Les couches cachées contiennent des neurones avec des fonctions d'activation non linéaires, normalement la fonction sigmoïde.



Les entrées Couches cachées Couche de sortie

Figure – 3 *Structure générale du perceptron multicouches*

Il a été prouvé [17] qu'il existe toujours un réseau de neurones de ce type avec trois couches seulement (les entrées, couche de sortie et une couche cachée) qui peut approximer une fonction f : $[0.1]^n \Rightarrow R^n$ avec n'importe quelle précision $\varepsilon > 0$ désirée. Un problème consiste à trouver combien de neurones cachés sont nécessaires pour obtenir cette précision. Un autre problème est de s'assurer a priori qu'il est possible d'apprendre cette fonction.

Initialement tous les poids peuvent avoir des valeurs aléatoires, qui sont normalement très petites avant de commencer l'apprentissage.

IV – 2 -.4. Apprentissage :

L'apprentissage d'un réseau de neurones signifie qu'il change son comportement de façon à lui permettre de se rapprocher d'un but défini. Ce but est normalement l'approximation d'un ensemble d'exemples ou l'optimisation de l'état du réseau en fonction de ses poids pour atteindre l'optimum d'une fonction économique fixée a priori.

Il existe trois types d'apprentissages principaux. Ce sont l'apprentissage supervisé, l'apprentissage non-supervisé, et l'apprentissage par tentative (graded training en anglais) [17].

On parle d'apprentissage supervisé quand le réseau est alimenté avec la bonne réponse pour les exemples d'entrées donnés. Le réseau a alors comme but d'approximer ces exemples aussi bien que possible et de développer à la fois la bonne représentation mathématique qui lui permet de généraliser ces exemples pour ensuite traiter de nouvelles situations (qui n'étaient pas pressenties dans les exemples).

Dans le cas de l'apprentissage non-supervisé le réseau décide lui-même quelles sont les bonnes sorties. Cette décision guidée par un but interne au réseau qui exprime une

configuration idéale à atteindre par rapport aux exemples introduits. Les cartes auto-organisatrices de Kohonen sont un exemple de ce type de réseau [16].

'Graded learning' est un apprentissage de type essai-erreur où le réseau donne une solution en étant seulement alimenté avec une information indiquant si la réponse était correcte, ou si elle était au moins meilleure que la dernière fois.

Il existe plusieurs règles pour chaque type d'apprentissage. L'apprentissage supervisé est le type le plus utilisé. Pour ce type d'apprentissage la règle la plus utilisée est celle de Widrow-Hoff. D'autres règles d'apprentissage sont par exemple la règle de Hebb, la règle de perceptron, la règle de Grossberg etc …. [14, 17, 18].

IV – 2 -.4.- 1 L'apprentissage de Widrow-Hoff :

La règle d'apprentissage de Widrow-Hoff est une règle qui permet d'ajuster les poids d'un réseau de neurones pour diminuer à chaque étape l'erreur commise par ce réseau de neurones (à condition que le facteur d'apprentissage soit bien choisi).
Un poids est modifié en utilisant la formule suivante :

$$w_{k+1} = w_k - \alpha \delta_k x_k \qquad (3)$$

Où :

w_k est le poids à l'instant k ;

w_{k+1} le poids à l'instant k-1 ;

α est le facteur d'apprentissage ;

δ_k caractérise la différence entre la sortie attendue et la sortie effective d'un neurone à l'instant k ;

x_k la valeur de l'entrée avec laquelle le poids w est associé à l'instant k.

Ainsi, si δ_k et x_k sont positifs tous les deux, alors le poids doit être augmenté. L'ampleur du changement dépend avant tout de la grandeur de δ_k mais aussi de celle de x_k. Le coefficient α sert à diminuer les changements pour éviter qu'ils deviennent trop grands, ce qui peut entraîner des oscillations du poids.

Deux versions améliorées de cet apprentissage existent, la version 'par lois' et la version 'par inertie' (momentum en anglais) [17], dont l'une utilise plusieurs exemples pour calculer la moyenne des changements requis avant de modifier le poids et l'autre empêche que le changement du poids au moment k ne devienne beaucoup plus grand qu'au moment k-1.

IV – 2 -.4.- 2 L'apprentissage par rétro-propagation du gradient (Levenberg-Marquardt backpropagation)

L'algorithme d'apprentissage par rétro-propagation du gradient (figure 4) est un algorithme itératif qui a pour objectif de trouver le poids des connexions minimisant l'écart commis par le réseau sur l'ensemble d'apprentissage. Cette minimisation par une méthode du gradient conduit à l'algorithme d'apprentissage par rétro-propagation.

La procédure d'apprentissage se décompose en deux étapes. Pour commencer, les valeurs d'entrées sont présentées au réseau, qui propage ensuite ces valeurs jusqu'à la couche de sortie et donne ainsi la réponse au réseau. A la deuxième étape les bonnes sorties correspondantes sont présentées aux neurones de la couche de sortie qui calculent l'écart, modifient leurs poids et rétro-propagent l'erreur jusqu'aux entrées pour permettre aux neurones cachés de modifier leurs poids de la même façon. Le principe de modification des poids est normalement l'apprentissage de Widrow-Hoff.

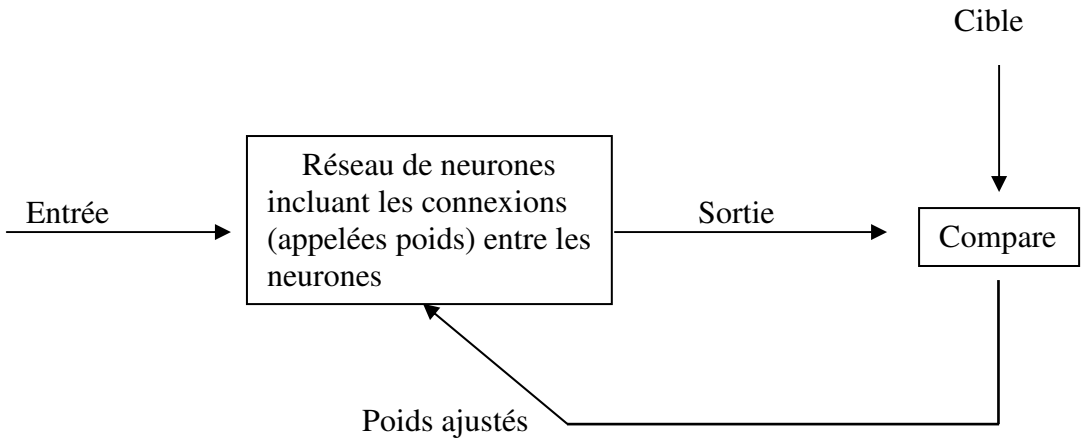


Figure – 4 *Apprentissage par un algorithme de rétro-propagation*

Généralement pour le calcul de l'écart on utilise l'erreur quadratique moyenne *MSE* (*Mean Square Error*) définie par la relation :

$$EQM = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} \qquad (4)$$

y_i est la valeur observée , $\hat{y}_i$ est la valeur estimée , et n le nombre d'observations.

VI – 2 –5 Critères d'arrêt

Plusieurs critères d'arrêt peuvent être utilisés avec l'algorithme d'apprentissage. Le premier critère consiste à fixer un nombre préalable de cycles ou d'itérations, mais il est difficile de savoir a priori combien d'itérations seraient appropriées pour arriver au but fixé.

Un deuxième critère consiste à fixer une borne inférieure sur l'erreur quadratique moyenne (MSE), il est parfois possible de fixer a priori un objectif à atteindre. Lorsque l'indice de performance choisi diminue en dessous de cet objectif, on considère simplement que le réseau a suffisamment bien appris ses données et on arrête l'apprentissage. L'inconvénient de ce critère est qu'il peut engendrer un phénomène de sur-apprentissage indésirable dans la pratique.

Le troisième critère est "l'arrêt précoce", qui consiste à suivre l'évolution des performances du réseau de généralisation durant le déroulement de l'apprentissage et à stopper celui-ci juste avant que ces performances ne se mettent à se dégrader, c'est-à-dire dès que l'indice de performance calculé sur les données de validation cesse de s'améliorer. Cette méthode, la plus utilisée pour éviter le sur-apprentissage, est celle pour laquelle nous avons optée dans ce travail. Le graphe suivant illustre ce critère :

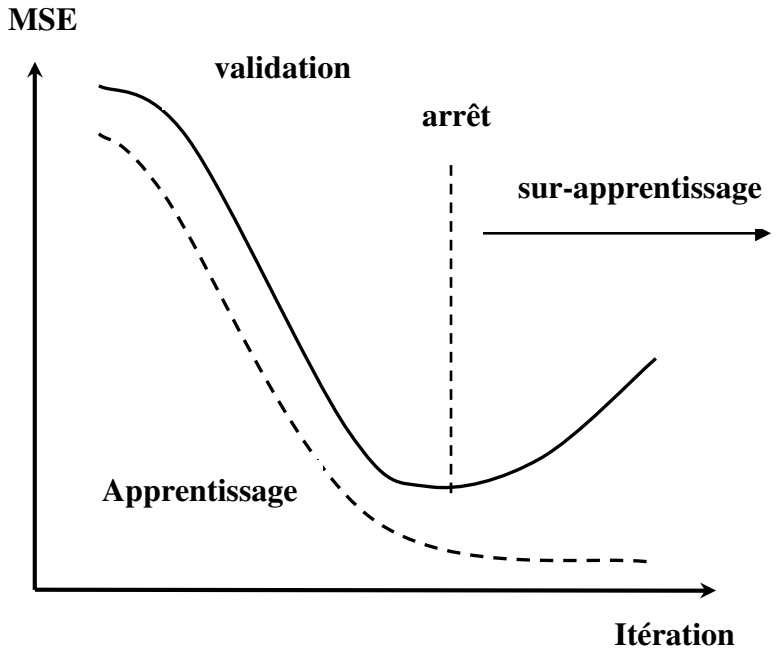


Figure – 5 *Illustration de l'arrêt précoce*

VI – 2 –6 Construction d'un modèle

La construction d'un modèle implique dans un premier temps le choix des échantillons des données d'apprentissage, de test et de validation. Le choix du type de réseau intervient dans une seconde étape.

Les quatre grandes étapes de la création d'un réseau de neurones sont détaillées comme suit :

VI – 2 –6- 1 Construction de la base de données

Le processus d'élaboration d'un réseau de neurones commence par la construction d'une base de données.

Afin de développer une application à base de réseaux de neurones, il est nécessaire de disposer de deux bases de données, une pour effectuer l'apprentissage et l'autre pour tester le réseau obtenu et déterminer ses performances. Pour cette raison nous avons partagé notre base des données (tableau I) aléatoirement en trois sous-ensembles comme suit :

- Un ensemble de 44 composés pour l'apprentissage du réseau de neurones.
- Un deuxième de 12composés pour la validation.
- Et un troisième de 12 composés choisis aléatoirement de l'ensemble d'apprentissage pour le test.

Généralement, les bases de données subissent un prétraitement qui consiste à effectuer une normalisation appropriée tenant compte de l'amplitude des valeurs acceptées par le réseau.

Les valeurs d'entrées et de sortie sont normalisées dans un intervalle spécifique afin de donner à chaque paramètre la même influence statistique. Les valeurs d'apprentissage et de test ont été normalisées dans la marge [- 1, 1], au moyen de l'équation

$$x_{norm} = 2 \times \frac{\left(x_j - x_{\min}\right)}{\left(x_{\max} - x_{\min}\right)} - 1 \qquad (5)$$

où x_{norm} est la valeur normalisée ; x_j est la j^{ième} valeur ; $x_{\max}$ est la valeur maximale ;

$x_{\min}$ est la valeur minimale

VI – 2 –6- 2 Définition de la structure du réseau

Nous avons retenu le Perceptron Multicouches comme base du modèle. Nous structurons ce réseau en précisant le nombre de couches et de neurones cachés pour que le réseau soit en mesure de reproduire ce qui est déterministe dans les données.

VI – 2 – 6 - 3 Nombre de couches et de neurones cachés

Mis à part les entrées et la couche de sortie, il faut décider du nombre de couches cachées. Sans couche cachée, le réseau n'offre que de faibles possibilités d'adaptation. Néanmoins, il a été démontré qu'un Perceptron Multicouches avec une seule couche cachée

pourvue d'un nombre suffisant de neurones, peut approximer n'importe quelle fonction avec la précision souhaitée [19].

Chaque neurone peut prendre en compte des profils spécifiques de neurones d'entrée. Un nombre plus important permet donc de mieux "coller" aux données présentées mais diminue la capacité de généralisation du réseau. Il faut alors trouver le nombre adéquat de neurones cachés nécessaire pour obtenir une approximation satisfaisante.

VI – 2 – 6 - 4 Présentation de l'environnement utilisé

Dans cette optique, le logiciel MATLAB [20], qui contient un module consacré au développement de réseaux de neurones, a été retenu ; un PC Dell P4 avec une Ram de 512 et une vitesse de 3.4 GHZ a été utilisé.

Le réseau de neurones stocke l'information dans une chaîne d'interconnexions neuronales, en faisant appel à la notion de poids (poids entrée - couche cachée = *IW* -i*nitial weights*, poids couche cachée - sortie = *LW-last weights*).

Une capacité d'apprentissage est nécessaire pour ajuster les poids des réseaux de neurones pendant la phase d'apprentissage au cours de laquelle toutes les données sont présentées au RNA à plusieurs reprises.

Les fonctions sigmoïde de transfert, tangente hyperbolique et linéaire, ont été adoptées comme fonctions d'activation pour les couches cachée et de sortie.

Nous présentons l'algorithme du réseau de neurones utilisé dans la page suivante :

Algorithme du réseau de neurones utilisé :

```
P= [les descripteurs];
T= [la propriété physique étudiée];
N = 56 ;       % tous les composés
N1 = 44 ;      % Composés d'apprentissage
N2 = 12 ;      % Composés de validation
P0= (P)';      % Transposition de la matrice P
T0= (T)';      % Transposition de la matrice T
[pn,minP,maxP,tn,minT,maxT] = premnmx(P0,T0);     % Normalisation entre [-1,+1]
P1n= (pn)';
T1n= (tn)';
% Apprentissage
P1=Pn(1:N1,:)';     % Descripteurs normalisés d'apprentissage
T1=Tn(1:N1,:)';     % Propriété physique normalisée d'apprentissage
T10=T(1:N1,:)';
%  Test
[R, Q] = size (P1);
iitst = [3:3:Q 2:42:Q];   % Choix aléatoire de 11 composés du test
test.P = P (:,iitst);
T20=T10 (:,iitst);
%  Validation
val.P = Pn(N1+1:N,:)';   % Descripteurs normalisés de validation
val.T = Tn(N1+1:N,:)';   % Propriété physique normalisée de validation
T30=T (N1+1:N, :)';
net = newff(minmax(P),[ S1 S2],{ TF1 TF2}, BTF);     % Création d'un réseau
 % S1 : Neurones de la couche cachée – S2 : la sortie (=1)
 % TF1, TF2 : Fonctions de transferts – BTF : Fonction de transfert de rétro-propagation
net.trainParam.epochs =250;            % Nombre d'itération
net.trainParam.goal= 0.0000001;      % Erreur désirée
net = init(net);        % Initialisation du réseau
[net,tr]=train (net, P1, T1, [], [], val);      % Entrainement du réseau
plotperf(tr)
a1n=sim(net,P1);       % Simulation du réseau pour les données d'apprentissage
[a1]=postmnmx(a1n,minT0,maxT0);%Remettre les résultats d'apprentissage à leurs valeurs réels
E1= T10-a1; % Calcul de l'erreur
a2n=sim(net,test.P);  % Simulation du réseau pour les données du test
[a2]=postmnmx(a2n,minT0,maxT0);%Remettre les résultats du test à leurs valeurs réels
E2= T20-a2; % Calcul de l'erreur
a3n=sim(net,val.P);   % Simulation du réseau pour les données de validation
[a3]=postmnmx(a3n,minT0,maxT0);%Remettre les résultats de validation à leurs valeurs réels
E3= T30-a3;  % Calcul de l'erreur
```

V – PARAMETRES D'EVALUATION DE LA QUALITE DE L'AJUSTEMENT

Deux paramètres sont couramment utilisés :

- Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \qquad (6)$$

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 SDEP ; Cf infra) :

$$\sigma_N = \sqrt{\frac{1}{n} \sum_1^n (y_i - \hat{y}_{(i)})^2} \qquad (7)$$

V – 1 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) [21]. Elle consiste à recalculer le modèle sur (n – 1) composés de phénol, le modèle obtenu servant alors à estimer l'indice de rétention du composé éliminé noté $\hat{y}_{(i)}$. On répète le procédé pour chacun des phénols.

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

$$PRESS = \sum_1^n (y_i - \hat{y}_{(i)})^2 \qquad (8)$$

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

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \qquad (9)$$

Contrairement à R^2, qui augmente avec le nombre de paramètres de la régression, le facteur Q^2_{LOO} 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^2_{LOO}. Une valeur de $Q^2_{LOO} > 0{,}5$ est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [22].

V – 2 Détection des observations aberrantes

Elle a été basée sur la non – satisfaction à trois au moins (pour n'en privilégier aucun) des six tests statistiques couramment utilisés pour la détection de telles observations en analyse de régression :

1°/ Les résidus ordinaires e_i, différences entre les valeurs observées (y_i) et estimées par le modèle $(\hat{y}_i)$.

2°/ Les résidus normalisés d_i, obtenus en divisant les e_i par l'écart type s de l'équation de régression.

3°/ Le résidu studentisé interne r_i, est le résidu d'une prédiction divisé par son écart type propre $\left(r_i = e_i / s\sqrt{1 - h_{ii}}\right)$.

4°/ Les leviers, h_{ii}, permettent de juger de l'influence d'une observation i dans la détermination de l'équation de régression.

5°/ La statistique représentée par le symbole DFITS :

$$\text{DFITS} = \frac{1}{p}\sqrt{\left(\frac{h_{ii}}{1 - h_{ii}}\right)}\, t_i \qquad (10)$$

permet de mesurer l'influence d'une observation i sur la valeur ajustée ou prédite. Belsley, Kuh et Welsch [23] considèrent qu'une observation pour laquelle DFITS > $2\sqrt{p/n}$ (p étant le nombre de paramètres de la régression) est inhabituelle.

6°/ La distance de Cook D_i :

$$D_i = \frac{1}{p}\frac{h_{ii}}{1 - h_{ii}} d_i^2 \qquad (11)$$

permet d'étudier l'influence d'une observation i sur les coefficients de régression estimés par les moindres carrés. Cook [24] et Weisberg [25] suggèrent de comparer D_i au paramètre de Fisher $F_{(0,5,p,n-p)}$ et de contrôler les observations avec distances de Cook > $F_{(0,5,p,n-p)}$. Comme $F_{(0,5,p,n-p)} \approx 1$, on considère que les observations pour lesquelles $D_i > 1$ sont influentes.

V – 3 Test de randomisation

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

V – 4 Validation externe

En plus du test de randomisation, il est intéressant [26], pour juger de la qualité du modèle, de considérer la racine de l'écart quadratique moyen (RMSE, pour Root Mean Squared Error), calculée sur différents ensembles:

- Ensemble d'estimation (appelée EQMC)
- Ensemble de validation croisée (appelée également EQMP)
- 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.

$$EQMC = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \tag{12}$$

$$\sigma_N = EQMP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}{n}} = \sqrt{\frac{PRESS}{n}} \tag{7}$$

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}\left(y_i - \hat{y}_{(i)}\right)^2}{n_{ext}}} \tag{13}$$

Nous traiterons l'indice de rétention comme propriété étudier avec un ensemble d'estimation et de validation. Nous favoriserons l'approche hybride algorithme génétique / régression multilinéaire et/ou réseaux de neurones :

GA / MLR et GA / RNA

INDICE DE RETENTION

I -1 Calcul du modèle

Le graphe de la figure 6 reproduit les variations du FIT (éq. 1) en fonction du nombre de variables du modèle. Il est évident que le modèle optimum (FIT maximum) comporte 7 variables.

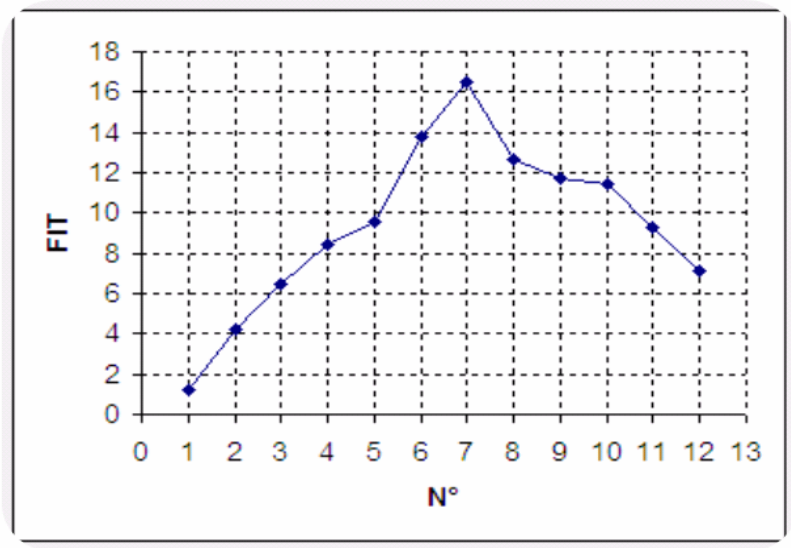


Figure 6 – Variation *du FIT en fonction de nombre de descripteurs.*

Dans les études QSRR, cinq observations, à la limite, doivent être associées à chaque variable explicative. Le nombre de degrés de liberté final doit être au moins égal à 10, soit, en désignant par k le nombre de descripteurs :

$$n-k-1\geq 10 \qquad (14)$$

Cette condition est bien vérifiée pour le modèle à 7 variables.

Pour les modèles à plus de 2 descripteurs, de faibles coefficients de corrélation croisés n'assurent pas forcément l'orthogonalité des descripteurs. Une indépendance globale acceptable des descripteurs sera vérifiée lorsque les facteurs d'inflation de la variance (FIV) calculés pour chacun d'eux sont inférieurs à 5.

Parmi les modèles optimaux générés, celui qui fournit la valeur maximale pour les paramètres statistiques Q^2, R^2 et Q^2_{ext} tout en vérifiant la condition : FIV comporte les 7

descripteurs calculés par le logiciel DRAGON, dont les symboles, la classe et la signification sont réunis dans le tableau II

Tableau II - Descripteurs moléculaires intervenant dans la modélisation de l'indice de rétention [8].

N°	Descripteur	Classe	Signification
1	VRp2	Indice topologique	Moyenne de type Randic (eigenvector) de la matrice distance pondérée par l'indice de polarisabilité
2	MATS3e	Indice d'autocorrélation -2D	Indice d'autocorrélation de Moran
3	Mor15u	Descripteurs de MoRSE -٣D	Fournissant des informations Signal MoRSE-3D-15 à partir des coordonnées 3D en utilisant la même transformation que dans la diffraction des électrons non pondéré
4	Mor21m		Signal MoRSE-3D-21 Pondéré par la masse atomique
5	Mor25m		Signal MoRSE-3D-25 Pondéré par la masse atomique
6	Mor19e		Signal MoRSE-3D-19 Pondéré par les électronégativités atomiques de Sanderson
7	SPP	Descripteur de charge	Paramètre de polarité sous-moléculaire

L'équation de régression ainsi établie est reproduite ci-après :

$$IR = 444(\pm 48.52) - 310(\pm 28.95)\ MATS3e + 154(\pm 6.800)\ VRp2 - 73.5(\pm 3.669)\ Mor15u - 190(\pm 27.85)\ Mor21m + 127(\pm 27.85)\ Mor25m - 33.5(\pm 8.351)\ Mor19e - 1855(\pm 97.28)\ SPP \quad (15)$$

I – 2 Analyse de régression

Les valeurs des paramètres statistiques montrent que les sept descripteurs (VRp2 ; MATS3e; Mor15u ; Mor21m ; Mor25m ; Mor19e ; SPP) (tableau-II) permettent de corréler l'indice de rétention des 44 composés de phénol.

Régresseur	Coef	Er-T coef	T	P	FIV
Constante	455.64	48.52	9.39	0.000	
MATS3e	−321.25	28.95	−11.10	0.000	1.5
VRp2	153.065	6.800	22.51	0.000	4.5
Mor15u	−72.817	3.669	−19.85	0.000	1.3
Mor21m	−197.24	27.85	−7.08	0.000	3.6
Mor25m	125.15	14.53	8.61	0.000	3.5
Mor19e	−32.127	8.351	−3.85	0.000	7.4
SPP	−1875.62	97.28	−19.28	0.000	1.4

S = 8.557 n = 44 ; σ_N=9,576 ; R^2 = 98,50 % ; Q^2=0,9768 ; F = 334,38

Matrice de Correlation:

	IR	MATS3e	VRp2	Mor15u	Mor21m	Mor25m	Mor19e
MATS3e	−0.199						
p	0.191						
VRp2	0.404	0.271					
p	0.006	0.072					
Mor15u	−0.190	−0.056	0.296				
p	0.211	0.715	0.048				
Mor21m	−0.385	−0.110	−0.024	−0.039			
p	0.009	0.470	0.878	0.798			
Mor25m	0.435	−0.399	−0.410	−0.096	−0.469		
p	0.003	0.007	0.005	0.529	0.001		
Mor19e	−0.156	0.453	0.700	0.059	0.384	−0.793	
p	0.307	0.002	0.000	0.699	0.009	0.000	
SPP	−0.212	−0.071	0.216	−0.012	−0.254	−0.041	0.157
p	0.162	0.641	0.154	0.935	0.092	0.791	0.302

En effet, la valeur du cœfficient de détermination (R^2) signifie que 98,5% de la variabilité de IR peut être expliquée par ces sept descripteurs, alors que la racine de l'erreur quadratique moyenne de prédiction est de l'ordre de (σ_N=9,576) ; en outre ce modèle est significatif (avec une valeur du paramètre de Fisher : F=334,38).

La commande « régression » de MINITAB fournit les valeurs des résidus caractéristiques réunis dans le tableau (III), ainsi que les valeurs hii, Di et DFITS qui permettent d'établir des diagnostics d'influence.

Tableau III - ***Diagnostics d'influence IR par MLR***

Observation i	Colonne ١	Colonne ٢	Colonne ٣	Colonne ٤	Colonne ٥	Colonne ٦	Colonne ٧	Colonne ٨
	e_i	d_i	r_i	h_{ii}	t_i	IR estimé	D_i	DFITS
١	-٠,١٨٨	٠,٠٢٤١٦	-٠,٠٢٤٢٦	٠,١٧٦٩٤٥	٠,٠٢٣٨٣	٧٨,٣٥٨	٠,٠٠٠٠١٦	٠,٠١١٠٥
٢	٣,٢٣١	- ٠,٥١٦٤٣	٠,٥١٢٠٧	٠,٤٦٥٤١٦	-٠,٥١١١٠	٨٣,٥٩٩	٠,٠٢٩٠٢٤	-٠,٤٧٦٨٩
٣	١,٨٨٥	-٠,٢٤٠٤٨	٠,٢٤١٠٤	٠,١٦١٢٥٩	-٠,٢٣٧٣٠	٩٨,١١٥	٠,٠٠١٣٩٠	-٠,١٠٤٠٥
٤	-١٨,٣٤٥	٢,٢٩٣٩٢	-٢,٢٩٢٢٩	٠,١٢٦٥٢٩	٢,٤٤٧٧٩	١١٩,٧٣٥	٠,٠٩٥٢٨١	٠,٩٣١٦٣
٥	١٥,٨٤٤	- ٢,٠١٢٩١	٢,٠٢١٧٨	٠,١٥٣٨٧٠	-٢,١٠٦٨٦	٩٤,٧٠٦	٠,٠٩٢١٠٤	-٠,٨٩٨٤٥
٦	-٠,٩٥٦	٠,١٩٦٠٣	-٠,١٩٣٧٣	٠,٦٧٥٠٠٧	٠,١٩٣٣٩	١١٨,٨٨٦	٠,٠٠٩٩٧٧	٠,٢٧٨٧١
٧	-٦,٧٠٥	٠,٩٤٦٠٤	-٠,٩٤٧٥٨	٠,٣١٣٩٢٢	٠,٩٤٤٦٢	١٢٤,٦٣٥	٠,٠٥١١٨٩	٠,٦٣٨٩٧
٨	-٣,٥٠٣	٠,٤٣٢٨٤	-٠,٤٣٣٠٦	٠,١٠٥٥٥٢	٠,٤٢٧٩٠	١٢٣,٦٠٣	٠,٠٠٢٧٦٤	٠,١٤٦٩٩
٩	-٦,٨١٦	٠,٨٢٥٠٨	-٠,٨٢٥٣٨	٠,٠٦٨٠٩١	٠,٨٢١٣٤	١٢٨,٥٣٦	٠,٠٠٦٢١٨	٠,٢٢٢٠٢
١٠	٣,٤٠١	-٠,٤٢٣٦٤	٠,٤٢٣٦٦	٠,١١٩٧٦٤	-٠,٤١٨٧٦	١١٨,٩١٩	٠,٠٠٣٠٥٢	-٠,١٥٤٤٧
١١	-٤,٤٩٨	٠,٥٤٥٦٢	-٠,٥٤٥٦٤	٠,٠٧١٦١٠	٠,٥٤٠٢٢	١٣٧,٧٦٨	٠,٠٠٢٨٧٠	٠,١٥٠٠٤
١٢	١٥,٥٥٣	- ٢,١٧٣٧	٢,١٧٢٥٢	٠,٣٠٠٨٨٠	-٢,٢٩٩٦١	١١٤,٤٩٧	٠,٢٥٤٢٠٤	-١,٥٠٨٦٠
١٣	٨,٣٧٦	-١,٠١٢٥٠	١,٠٤٩٥١	٠,٠٦٥٣١٥	-١,٠١٢٨٧	١٣٥,٣٢٤	٠,٠٠٨٩٥٥	-٠,٢٦٧٧٥
١٤	٨,٩٣٩	-١,١٢٤٢٦	١,٠٧٣٩١	٠,١٣٦٥٩٩	-١,١٢٨٥٣	١٣٤,٧٦١	٠,٠٢٤٩٩٧	-٠,٤٤٨٨٨
١٥	-١,٣٠٥	٠,١٥٦٦٥	-٠,١٨٠٣٤	٠,٠٥٢٢٩٧	٠,١٥٤٥١	١٤٦,٧٤٥	٠,٠٠٠١٦٩	٠,٠٣٦٣٠
١٦	٩,٨٥٥	-١,٣٥٨٣٨	١,٣٢٩٣٠	٠,٢٨١٢٠٣	-١,٣٧٥٠٩	١٣٥,٥٨٥	٠,٠٩٠٢٣٣	-٠,٨٦٠٠٧
١٧	٩,٤٢٦	-١,٢٧٤٤٠	١,١٧٥٧١	٠,٢٥٢٨٢٣	-١,٢٨٥٩١	١٤٠,٠١٤	٠,٠٦٨٦٩٣	-٠,٧٤٨٠١
١٨	٠,٠٦١	-٠,٠٠٧٥٥	٠,٠٠٧٣٠	٠,١٢٠٩٤٠	-٠,٠٠٧٤٥	١٥١,٥٤٩	٠,٠٠٠٠٠١	-٠,٠٠٢٧٦
١٩	٦,٧١٢	-٠,٨٠٣١٩	١,٠٢٨٧١	٠,٠٤٦٣٥٣	-٠,٧٩٩١٥	١٤٤,٨٩٨	٠,٠٠٣٩٢٠	-٠,١٧٦١٩
٢٠	-٩,٣٠٢	١,٤٢٨٨٧	-١,١٢٢٧٣	٠,٤٢١٢١١	١,٤٥٠٦١	١٥٦,٣٣٢	٠,١٨٥٧٢٦	١,٢٣٧٤٩
٢١	-٧,٢٦٢	٠,٨٧٥٢٦	-٠,٨٦٦٩٨	٠,٠٥٩٩٠١	٠,٨٧٢٣٥	١٦٣,٣٧٢	٠,٠٠٦١٠٢	٠,٢٢٠٢٠
٢٢	٤,٨٩	-٠,٦١٠٦١	٠,٦٠٨٦٥	٠,١٢٤٠٢٤	-٠,٦٠٥٢١	١٥١,٧٤٠	٠,٠٠٦٥٩٩	-٠,٢٢٧٧٣
٢٣	٤,٠٥٥	-٠,٥٧٨٧٤	٠,٥٧٩٢٣	٠,٣٢٩٦٢٧	-٠,٥٧٣٣٢	١٥٧,٥٣٥	٠,٠٢٠٥٨٦	-٠,٤٠٢٠٢
٢٤	٢,٥٩٢	-٠,٣١٤٥١	٠,٣١٤٤٨	٠,٠٧٢٠٤٨	-٠,٣١٠٥٤	١٦٣,٥٤٨	٠,٠٠٠٩٦٠	-٠,٠٨٦٥٣

Tableau III – suite et fin.

Observation i	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5	Colonne 6	Colonne 7	Colonne 8
	e_i	d_i	r_i	h_{ii}	t_i	IR estimé	D_i	DFITS
25	1.814	-٠,٢٢٢٢٩	٠,٢٢١٨٥	٠,٠٩٠٣٩٢	-٠,٢١٩٣٣	١٧١,٠٣٦	٠,٠٠٠٦١٤	-٠,٠٦٩١٤
26	0.514	-٠,٠٧١٠٨	٠,٠٧٠٩٠	٠,٢٨٦٨٢٥	-٠,٠٧٠٠٩	١٨٠,٠٨٦	٠,٠٠٠٢٥٤	-
27	-6.098	٠,٧٥٦٥٣	-٠,٧٥٦٠٦	٠,١١٢٥٨٦	٠,٧٥١٩٥	١٩٣,٦٣٨	٠,٠٠٩٠٧٧	٠,٢٦٧٨٤
28	7.854	-٠,٩٧٣٦٩	٠,٩٧٨٤٣	٠,١١١٣٥٠	-٠,٩٧٢٩٧	١٧٨,٨٦٦	٠,٠١٤٨٤٩	-٠,٣٤٤٤١
29	-7.31	٠,٩٠٣٢٤	-٠,٩٠٠١٤	٠,١٠٥٤٦٧	٠,٩٠٠٨٧	٢٠٧,٣١٠	٠,٠١٢٠٢٤	٠,٣٠٩٣٣
30	-18.271	٢,٢٤٩٧٧	-٢,٢٤٣١٨	٠,٠٩٩٢٤٥	٢,٣٩٢٨٩	٢٢٠,٦٥١	٠,٠٦٩٧٠٩	٠,٧٩٤٢٨
31	-8.464	١,٠٤٤٥٠	-١,٠٤٣٣٣	٠,١٠٣١١٩	١,٠٤٥٨٦	٢١٢,٢٠٤	٠,٠١٥٦٧٩	٠,٣٥٤٦٣
32	7.514	-٠,٩٤٦٠٣	٠,٩٤٩٢٠	٠,١٣٨٤٩٩	-٠,٩٤٤٦١	١٩٨,٥٤٦	٠,٠١٧٩٨٥	-٠,٣٧٨٧٥
33	-4.403	٠,٥٦٧١١	-٠,٥٦٧٥٢	٠,١٧٦٨٣٠	٠,٥٦١٦٩	٢٢٦,٢٢٣	٠,٠٠٨٦٣٦	٠,٢٦٠٣٣
34	-9.444	١,٢٠٦٩٤	-١,٢٠٠٠٠	٠,١٦٣٨٤١	١,٢١٤٨٩	٢٣٨,٦٧٤	٠,٠٣٥٦٧٩	٠,٥٣٧٧٨
35	8.557	-١,١٣١٨٥	١,١٢٨٣٧	٠,٢١٩٣٨١	-١,١٣٦٤٢	٢٣٣,٠٦٣	٠,٠٤٥٠٠٤	-٠,٦٠٢٤٥
36	-0.551	٠,٠٧٩١٢	-٠,٠٧٩٢٥	٠,٣٣٧٩٦٩	٠,٠٧٨٠٢	٢٥١,٢٩١	٠,٠٠٠٣٩٩	٠,٠٥٥٧٥
37	-13.179	١,٧٨٦٤٣	-١,٧٨٣٦٠	٠,٢٥٦٦٨٩	١,٨٤٥١٢	٣٦٧,٢٤٩	٠,١٣٧٧٥٨	١,٠٨٤٢٨
38	4.32	-٠,٦٠٠٨٩	٠,٥٩٨٧٩	٠,٢٩٤٠٣٧	-٠,٥٩٥٤٨	٣٥٥,١٠٠	٠,٠١٨٧٩٨	-٠,٣٨٤٣٠
39	-7.314	٠,٨٩٠٩٠	-٠,٨٩٠٥٠	٠,٠٧٩٦٠٣	٠,٨٨٨٢٩	١٢٨,٣٢٤	٠,٠٠٨٥٨١	٠,٢٦١٢٤
40	3.377	-٠,٤١٥٣٨	٠,٤١٥٦٧	٠,٠٩٧٥٥٤	-٠,٤١٠٥٦	١٥٠,٢٣٣	٠,٠٠٢٣٣١	-٠,١٣٤٩٩
41	4.678	-٠,٥٥٩٨٥	٠,٥٦٥٦٧	٠,٠٤٦٢٩٧	-٠,٥٥٤٤٤	١٥٢,١٧٢	٠,٠٠١٩٠٢	-٠,١٢٢١٦
42	-0.508	٠,٠٦٩٩٦	-٠,٠٦٣٢٠	٠,٢٧٩٢٥٥	٠,٠٦٨٩٩	٢٧١,٦٥٨	٠,٠٠٠٢٣٧	٠,٠٤٢٩٤
43	2.383	-٠,٣٢٣٥٥	٠,٣٢٤٠١	٠,٢٥٨٩٩٩	-٠,٣١٩٤٩	٣٠٧,١٦٧	٠,٠٠٤٥٧٤	-٠,١٨٨٨٩
44	-1.408	٠,١٦٨٠٦	-٠,١٨٨٧٩	٠,٠٤٠٨٧٩	٠,١٦٥٧٨	١٥٧,٥١٨	٠,٠٠٠١٥٠	٠,٠٣٤٢٢

L'analyse des résidus (colonne 1 tableau III) permet, en particulier, de voir que les résidus ordinaires $e_ε$, e_{30}, sont supérieurs à 2 fois l'erreur standard ($|e_i|>2S$), soit
2x8, 557=17,114.

Tous les résidus standardisés d_i de la colonne (2) sont compris entre les limites ± 2, à l'exception des points 4, 5, 12, et 30.

La colonne (3) rassemble les résidus studentisés internes r_i qui sont du même ordre de grandeur que les d_i correspondants. On a ici p = 8 et n = 44, et on constate que tous les r_i sont inférieurs en valeur absolue à $t_{(0,025;n-p)}$ [= 2,0.546] à l'exception de r_4, r_{12}, r_{30}, qui est le 0,975 quantile d'une loi de Student avec (n-p) degrés de liberté.

La colonne (4) donne les valeurs de h_{ii}, $i^{ème}$ terme diagonal de la matrice de projection : $\underset{\sim}{H} = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\ \underset{\sim}{X}'$ où $\underset{\sim}{X}$ est la matrice des valeurs observées des variables explicatives et $\underset{\sim}{X}'$ sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques. La valeur critique pour déterminer les points leviers correspond à

$h^* = \frac{3\times7}{44} = \frac{3p}{n} = 0,48$. On constate que tous les hii sont inférieures à cette valeur critique, à l'exception du composé 6.

Les résidus studentisés externes, t_i, rassemblés dans la colonne (5) sont du même ordre de grandeur que les r_i correspondants, on remarque que tous les t_i sont inférieurs en valeur absolue à $t_{(0,025;n-p-1)}$ [= 2,0525] à l'exception de t_4, t_5, t_{12} et t_{30}.

Le diagramme de Williams (di en fonction de hii) de la figure (7) fait ressortir en plus des observations aberrantes qui sont 4, 5, 12, et 30 , le point 6, qui possèdent un bras de levier important et de plus influent.

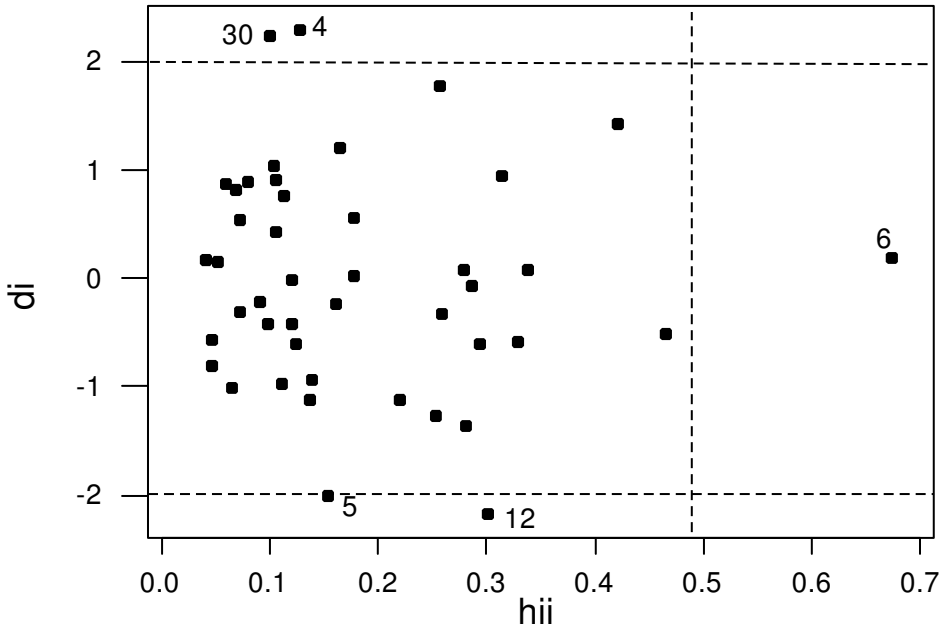


Figure – 7: Diagramme de Williams.

I – 3 Autres diagnostics d'influence

Les autres mesures d'influences utiles, dont l'étude complète la recherche des observations aberrantes sont présentées dans les colonnes 7 et 8.du tableau III

On remarque que les distances de Cook sont toutes inférieures à 1 ; et que les DFITS des observations 4, 5, 12, 16, 20, et 37 sont supérieurs à la valeur critique $2\sqrt{p/n} = 2\sqrt{8/44} = 0{,}852$. Ces observations sont donc inhabituelles.

L'analyse des résidus studentisés internes et externes permet de détecter les observations aberrantes qui sont 4, 5, 12, et 30.

I – 4 Vérification de la qualité de l'ajustement

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par « leave –one –out ». La figure (8), qui reproduit les valeurs prédites $I\hat{R}$ en fonction de celles observées, fait ressortir une faible dispersion caractéristique d'un bon ajustement, d'ailleurs confirmé par la grande valeur de Q^2 (=0,976).

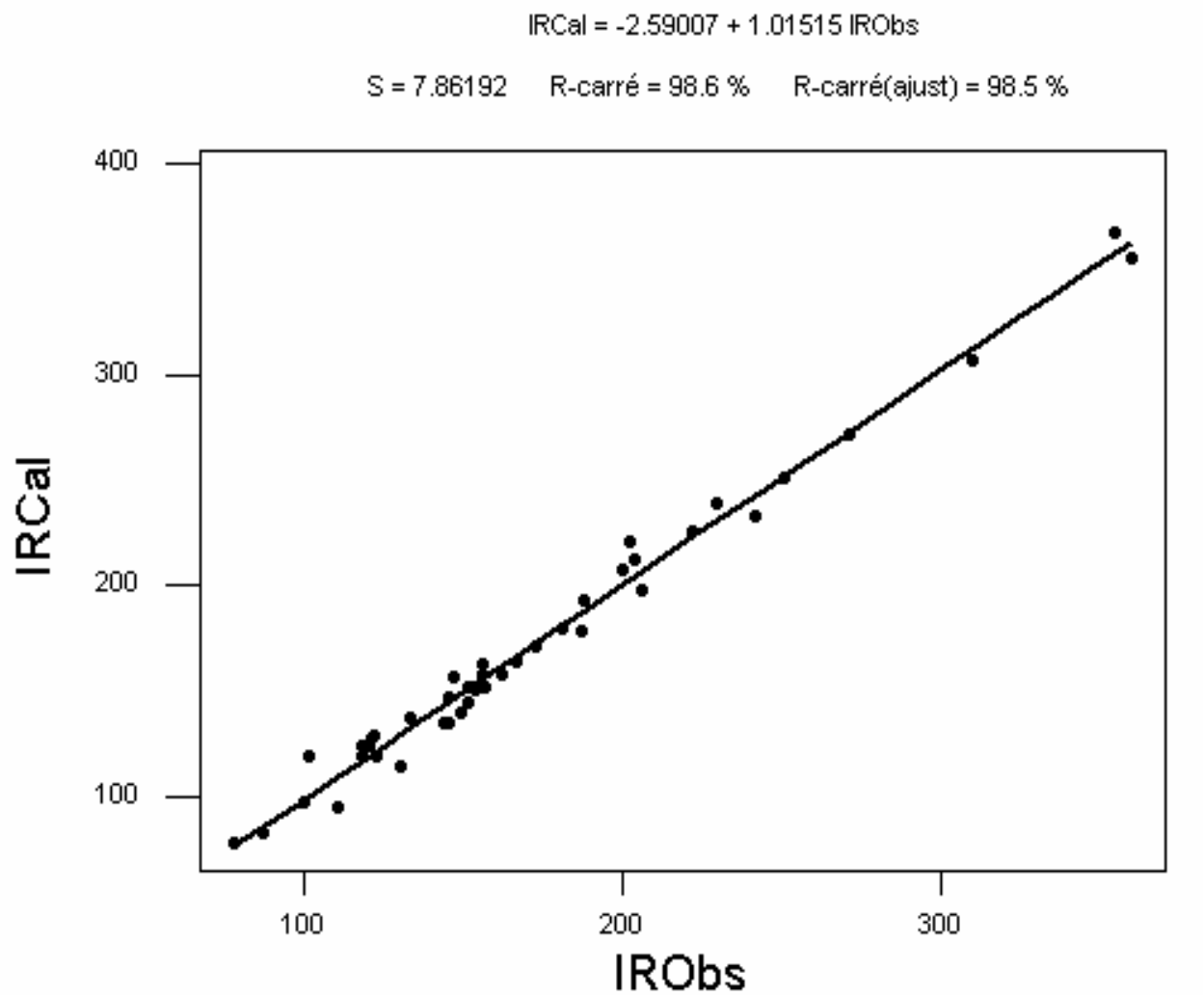


Figure 8– Graphe des valeurs calculées IR Cal en fonction des valeurs observées.

Dans le but d'établir que le modèle obtenu n'est pas dû au hasard ou à une sur-spécification, nous avons appliqué le test de randomisation.

Ainsi 100 nouveaux vecteurs de températures critiques ont été générés par permutation des positions des composantes du vecteur réel:

$$y = (y_1, y_2, ..., y_{27})' \xrightarrow{RND} y_{RND} = (y_8, y_5, ..., y_2)'$$

et utilisés comme sources d'observations pour des modèles QSRR dans les conditions optimales établies (7paramètres) .

La figure 9 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (cercles noircis) au modèle réel de départ (astérisque).

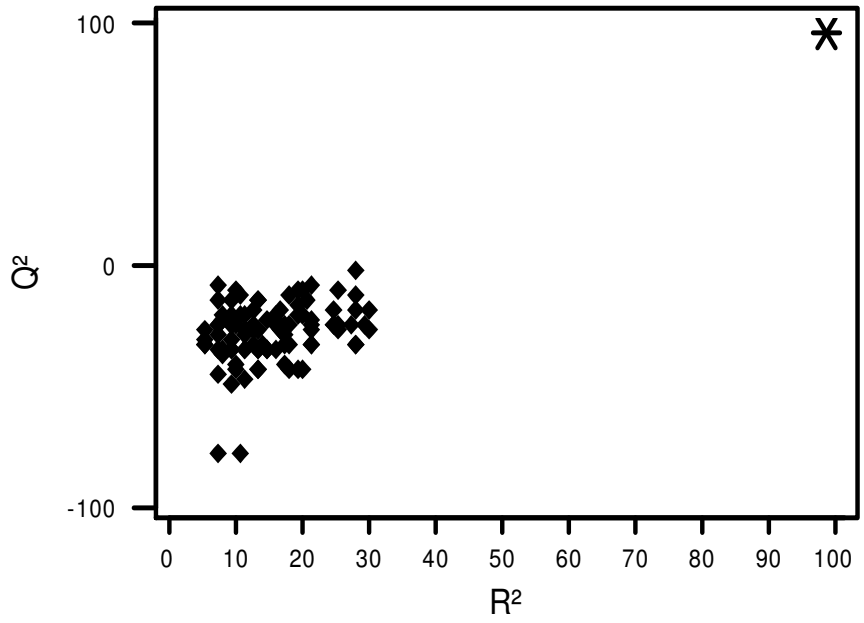


Figure 9 – *Test de randomisation associé au modèle QSRR. Les points noircis représentent les indices de rétentions ordonnés de façon aléatoire, et l'astérisque correspond au modèle réel.*

Il est clair que les statistiques obtenues pour les vecteurs modifiés des l'indices de rétentions sont plus petits que celles du modèle QSRR réel, et pour la majeure partie on obtient un $Q^2 < 0$. Ceci permet d'assurer qu'une relation structure/ indice de rétention réelle a été établie.

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

I – 5 Validation externe

Pour savoir la capacité prédictive du notre modèle, nous avons opéré par validation externe sur l'ensemble de 12 composés choisis aléatoirement et qui ne font pas partie de l'ensemble d'essai (composés numérotés de 45 à 56 dans le tableau I).

Une validation rigoureuse du modèle se traduit par une proportion importante de prédictions exactes données sur l'ensemble de la validation. La performance du modèle est alors mesurée par le coefficient de régression R^2.

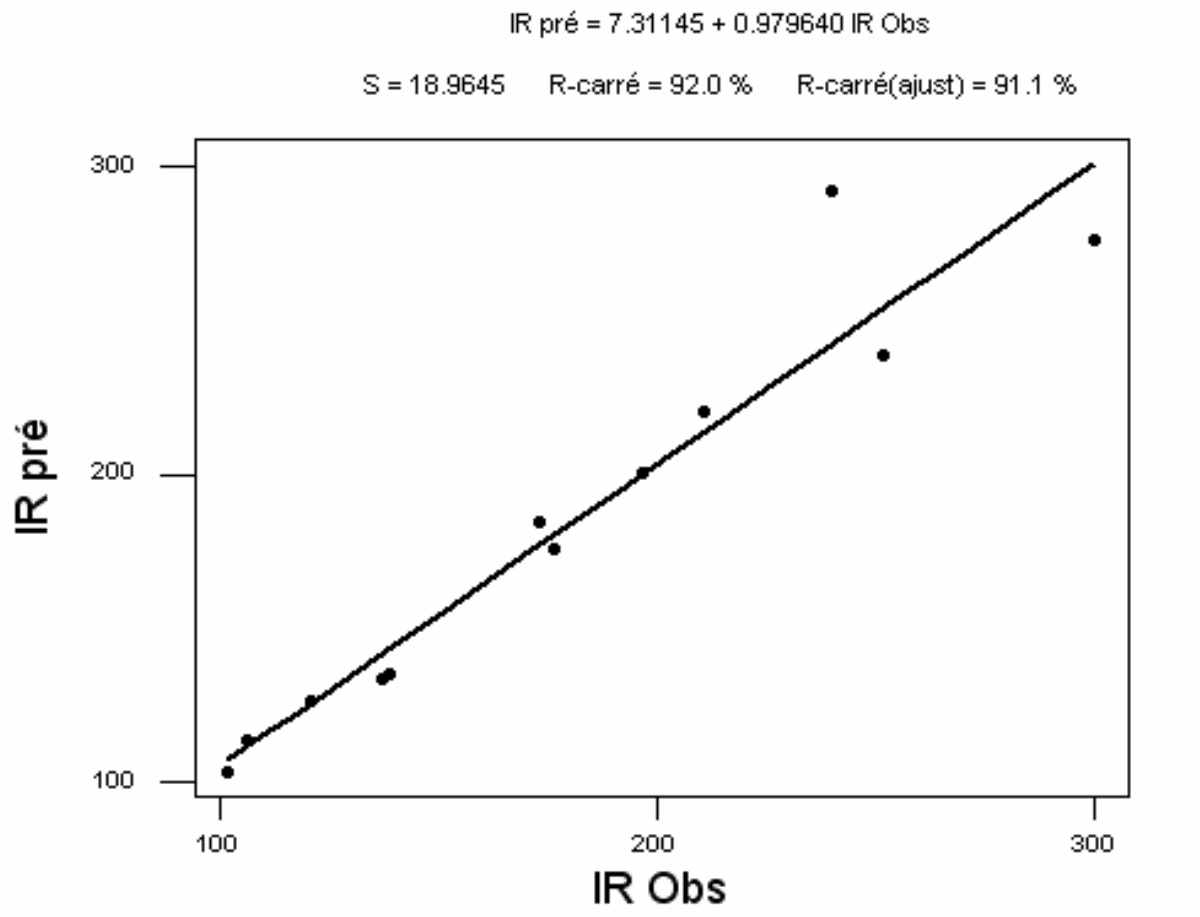


Figure 10 – *Graphe des $I\hat{R}$ prédites en fonction des* IR *observées pour validation externe.*

Tableau IV - Valeurs des indices de rétentions observées IR et prédites $\hat{IR}$ en exploitant la MLR, ainsi que leurs différences, pour l'ensemble de validation externe.

	IR	$\hat{IR}$	$e_{(i)}$
45	١٠٢,١٢	١٠٣,١٦	-1.04
46	١٢١,٠١	١٢٦,٣٤	-5.33
47	١٣٧,٦٨	١٣٣,٣٣	4.35
48	١٧٦,٦٨	١٧٥,٧١	0.97
49	١٩٧,٠٧	٢٠٠,٤٦	-3.39
50	٢١١,٠١	٢٢٠,٣٥	-9.34
51	٢٥١,٦١	٢٣٩,٢١	12.4
52	١٠٦,٦١	١١٣,٩٦	-7.35
53	١٧٣,٥٧	١٨٤,٥٢	-10.95
54	٣٠٠,٠٠	٢٧٦,٣٦	23.64
55	١٣٩,١٦	١٣٤,٧٩	4.37
56	240,05	292.21	-52.16

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

Les valeurs de tous ces paramètres statistiques, réunies ci-après,

EQMC	= 7,47	(44 objets)	R^2	= 98,50
EQMP	= 9,57	(44 objets)	Q^2	= 97,68
EQMP (ext)	= 17,89	(12 objets)	Q^2 (ext)	= 91,22

montrent tous à la fois, une mauvaise capacité prédictive (valeurs des RMSE élevées). On peut dire aussi que les valeurs des $\hat{IR}$ prédites sont proches des IR observés (sauf pour certains composés tels que 54 et 56 possèdent des erreurs élevées qui sont responsables de la mauvaise capacité prédictive de notre modèle développé).

II-1- modèle hybride algorithme génétique / réseaux de neurones artificiels

II - 1- 1 choix des paramètres statistiques

Les descripteurs choisi par algorithme génétique sont utilisés pour la configuration du réseau de neurones, qui est perfectionnée en phase d'apprentissage ; les paramètres de fonctionnement sont déterminés de façon à obtenir une bonne adéquation entre les valeurs simulées et les données d'apprentissage, combinée à une généralisation correcte de ces simulations.

II - 1- 2 Choix du nombre de couches cachées

Quelle que soit la problématique étudiée, l'utilisation d'une seule couche cachée permet d'obtenir de meilleures configurations des réseaux de neurones.

II - 1- 3 Choix du nombre d'itérations et de neurones dans la couche cachée

Le choix de ce nombre est très important. Au départ, on fixe un nombre de neurones (2 à 8) et on fait varier le nombre d'itérations pour calculer à chaque fois l'erreur quadratique moyenne EQM.

Le nombre de neurones de la couche cachée ainsi que le nombre d'itérations et fixé par la valeur minimale de l'erreur quadratique moyenne EQM.

Le graphe EQM = f (nombre d'itérations) à balayage de nombres de neurones de la figure suivante permet de visualiser EQM_{min} qui correspond à 250 comme nombre d'itérations et à 06 comme nombre de neurones de la couche cachée pour une meilleure configuration du réseau.

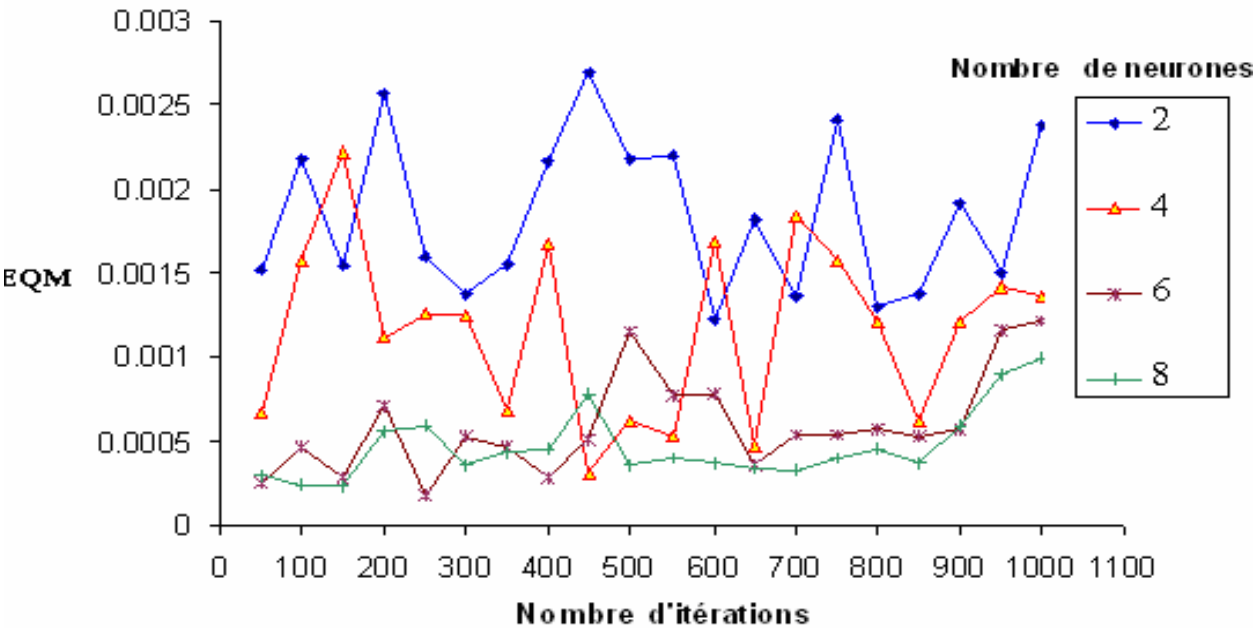


Figure 11 – *Choix du nombre d'itérations, et de neurones dans la couche cachée*

II - 1- 4 Choix de la fonction de transfert

Les réseaux de neurones les plus adaptés à notre étude ont l'architecture suivante :

- Fonction de transfert tangente hyperbolique (tansig) pour la couche cachée
- Fonction de transfert linéaire (purelin) pour la couche de sortie.

II - 1- 5 Choix des paramètres d'apprentissage

Ces paramètres sont également importants et ont permis d'affiner la configuration des réseaux de neurones pour obtenir les meilleures prédictions.

- ♣ Indice de performance choisi: EQM (pour l'erreur quadratique moyenne).

L'apprentissage du réseau de neurones représente un fragile équilibre entre tous ces paramètres, d'où la difficulté pour l'atteindre. Une fois cet apprentissage achevé, le réseau de neurones devient un outil viable et peut être utilisé pour la simulation de nouvelles données. Le tableau VI précise la structure optimale du réseau de neurones.

Tableau – V Structure optimale du réseau de neurones.

Nombre d'entrées	07 (les descripteurs)
Nombre de sorties	01 (l'indice de rétention)
Nombre de couches cachées	Une couche cachée
Nombre d'itérations	250
Nombre de neurones dans la couche cachée	06
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonctions d'apprentissage	Tangente hyperbolique (couche cachée) Linéaire (couche de sortie)

II - 1- 6 Résultats et discussion

II - 1- 6 – 1 Evaluation de la qualité de l'ajustement

Nous avons deux paramètres utilisés pour évaluer la qualité de l'ajustement; la valeur du cœfficient de détermination R^2 = 99, 88 % qui explique très bien la variabilité de IR en fonction des descripteurs choisis; la racine de l'erreur quadratique moyenne de prédiction σ_N= 3,62 dont la petite valeur indique un modèle très hautement significatif, que justifie la grande valeur du paramètre de Fisher : F=4390,630.

II - 1- 6 – 2 Vitrification de la qualité de l'ajustement :

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par « leave –one –out ». La figure (12), reproduit les valeurs prédites $\hat{IR}$ en fonction de celles observées, fait ressortir un bon ajustement, d'ailleurs confirmé par la grande valeur de Q^2 (=0,9984).

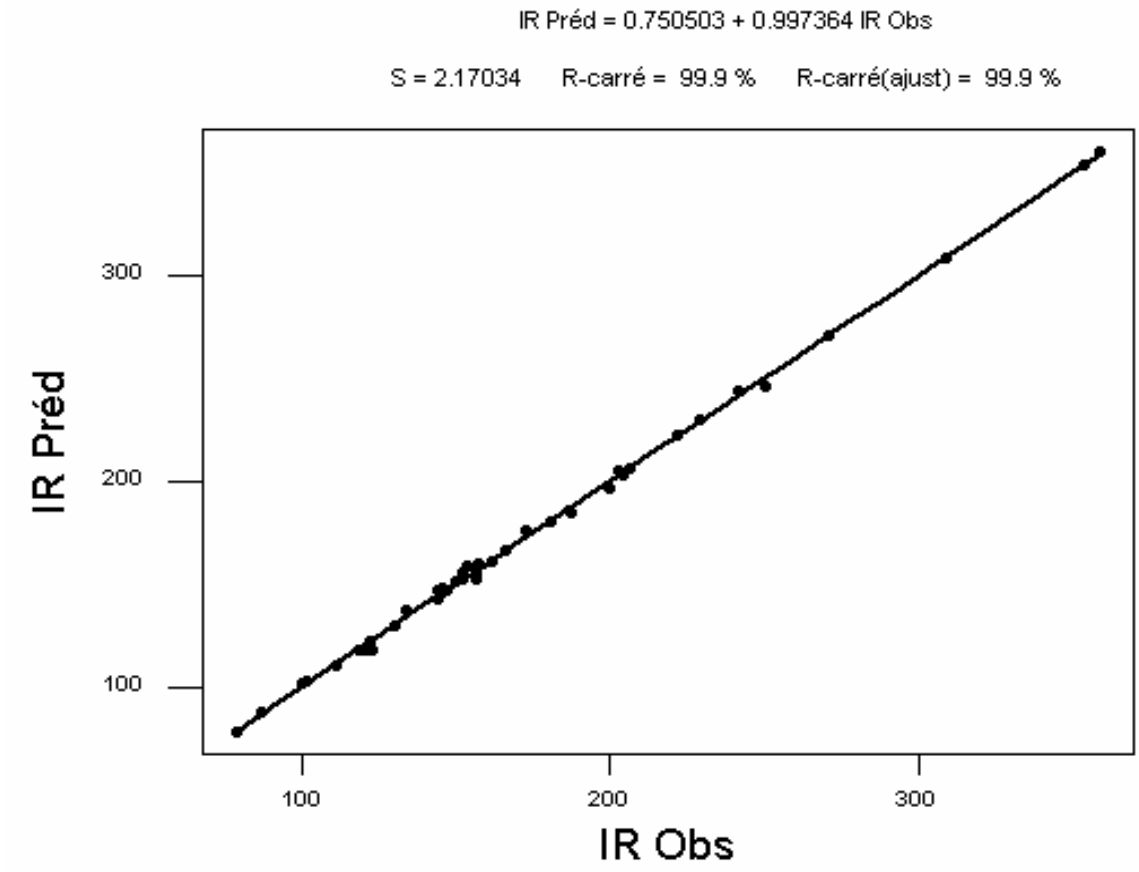


Figure 12 – *Graphe des valeurs prédites $\hat{IR}$ en fonction des valeurs observées IR.*

II-2 Validation externe

L'évaluation de la capacité de généralisation du réseau est réalisée sur la base de la validation externe, la performance du réseau est alors mesurée par le coefficient de régression R^2.

Les résultats obtenus montrent que les valeurs prédites (tableau - VI) sont très proches des valeurs observées (figure-13). La valeur de R^2 est égale à 99,4 %, qui confirme que le modèle neuronal décrit de façon adéquate la relation entre les indices de rétentions prédites et observées.

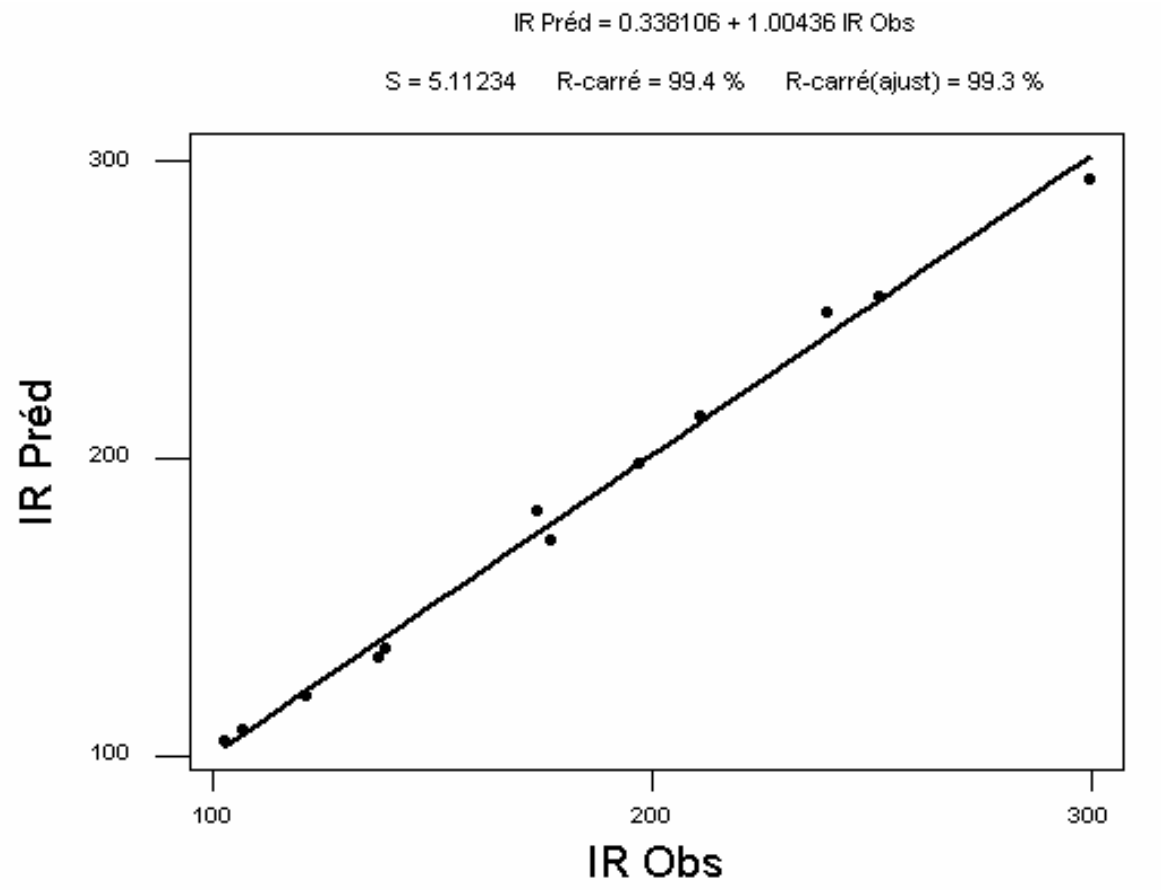


Figure 13 – *Graphe des $I\hat{R}$ prédites en fonction des IR observées pour validation*

Tableau - VI Les valeurs IR observés, prédits et les erreurs pour l'ensemble de validation externe trouvé par RNA.

	IR	$\hat{IR}$	$e_{(i)}$
45	١٠٢,١٢٠	١٠٥,١٥٠	-٣,٠٢٩٧
46	١٢١,٠١٠	١٢٠,١٩٩	٠,٨١١١
47	١٣٧,٦٨٠	١٣٣,٢٥٤	٤,٤٢٥٧
48	١٧٦,٦٨٠	١٧٢,٣٤٣	٤,٣٣٧٠
49	١٩٧,٠٧٠	١٩٨,٧٠٤	-١,٦٣٤١
50	٢١١,٠١٠	٢١٤,٢١٢	-٣,٢٠٢٣
51	٢٥١,٦١٠	٢٥٥,٠٦٦	-٣,٤٥٦١
52	١٠٦,٦١٠	١٠٨,٩٠٤	-٢,٢٩٣٦
53	١٧٣,٥٧٠	١٨٢,٢٩٥	-٨,٧٢٥٢
54	٣٠٠,٠٠٠	٢٩٤,٥٤٦	٥,٤٥٣٥
55	١٣٩,١٦٠	١٣٦,١٤١	٣,٠١٨٧
56	٢٤٠,٠٥٠	٢٤٩,٢١١	-٩,١٦٠٥

Les valeurs RMSE sont réunies ci-après :

EQMC	= ٢,١٥	(44 objets)	R^2	= 99,88
EQMP	= ٢,٥٠	(44 objets)	Q^2	= 99,84
EQMP (ext)	= ٤,٠٦	(12 objets)	Q^2 (ext)	= 99,08

Les faibles valeurs des RMSE montrent une bonne capacité prédictive du modèle et une possibilité d'extension suffisante (valeurs proches ou similaires).

Tableau – VII comparaisons des IR observés, prédits et les résidus trouvés par MLR et RNA pour l'ensemble de validation externe**.**

N°	$IR_{(ob)}$	$I\hat{R}_{(MLR)}$	$I\hat{R}_{(RNA)}$	$e_{(i)\,(MLR)}$	$e_{(i)\,(RNA)}$
45	١٠٢,١٢٠	١٠٣,١٦	١٠٥,١٥٠	-1.04	-٣,٠٢٩٧
46	١٢١,٠١٠	١٢٦,٣٤	١٢٠,١٩٩	-5.33	٠,٨١١١
47	١٣٧,٦٨٠	١٣٣,٣٣	١٣٣,٢٥٤	4.35	٤,٤٢٥٧
48	١٧٦,٦٨٠	١٧٥,٧١	١٧٢,٣٤٣	0.97	٤,٣٣٧٠
49	١٩٧,٠٧٠	٢٠٠,٤٦	١٩٨,٧٠٤	-3.39	-١,٦٣٤١
50	٢١١,٠١٠	٢٢٠,٣٥	٢١٤,٢١٢	-9.34	-٣,٢٠٢٣
51	٢٥١,٦١٠	٢٣٩,٢١	٢٥٥,٠٦٦	12.4	-٣,٤٥٦١
52	١٠٦,٦١٠	١١٣,٩٦	١٠٨,٩٠٤	-7.35	-٢,٢٩٣٦
53	١٧٣,٥٧٠	١٨٤,٥٢	١٨٢,٢٩٥	-10.95	-٨,٧٢٥٢
54	٣٠٠,٠٠٠	٢٧٦,٣٦	٢٩٤,٥٤٦	23.64	٥,٤٥٣٥
55	١٣٩,١٦٠	١٣٤,٧٩	١٣٦,١٤١	4.37	٣,٠١٨٧
56	٢٤٠,٠٥٠	٢٩٢,٢١	٢٤٩,٢١١	-52.16	-٩,١٦٠٥

Le jugement de la qualité des modèles (MLR et RNA) a été vérifié en calculant les RMSE; ces paramètres statistiques sont réunis ci-après:

Tableau - VIII Valeurs des paramètres statistiques trouvés par les deux méthodes.

	MLR	**RNA**
n	12	12
S	8,55	2,15
σ_N	9,57	2,50
R^2	98,50	99,88
Q^2	97,68	99,84
Q^2_{ext}	91,22	99,08
EQMC	7,47	2,15
EQMP	9,57	2,50
$EQMP_{ext}$	17, 89	4,06

CONCLUSION GENERALE

Nous avons appliqué la méthodologie QSRR pour relier les indices de rétentions de composés de phénols etudiés , à des descripteurs moléculaires théoriques reflétant certaines particularités des molécules considérées.

Le mélange pris en compte comprend 56 composés substitué par des groupement alkyles.

Les modèles QSRR ont été établis en utilisant l'analyse de régression multilinéaire et /ou les réseaux de neurones standards à 3 couches (les entrées, une couche cachée et une couche de sortie), avec algorithme d'apprentissage par rétro- propagation du gradient (Levenberg- Marquard)).

Les 56 composés de base ont été éclatées aléatoirement en deux ensembles disjoints.

-un ensemble principal de 44 composés utilisés pour le calcul et, éventuellement, les essais du modèle ;

- un ensemble de 12 composés pour la prédiction externe.

La taille du modèle est fixée par la valeur optimale de la fonction FIT de KUBINYI. La sélection des variables explicatives a été réalisée par algorithme génétique, dans la version MOBYDIGS de TODESCHINI[27], en maximisant Q^2_{L00}.

Les statistiques réunies ci-après permettent de faire des comparaisons, et de tirer plusieurs conclusions comparons les résultats trouvés par MLR et RNA .

		(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
		R^2 (%)	**σ_N**	**Q^2 (%)**	**F**	**Q^2_{ext} (%)**	**RMSE**	**Points * aberrants**	**Points ** aberrants**
IR	(MLR)	98,50	9,57	97,68	334,38	91,22	7,47-17,89	4-5-12-30	54-56
	(RNA)	99,88	2,50	99,84	4390,63	99,06	2,15- 4,06	30	-

Points de l'ensemble d'essai (*) et de l'ensemble de prédiction externe(**).

La numérotation des composés est celle du tableau I (pp. 6 et7).

Les statistiques (R^2 ; σ_N ; F) calculées permet de jugé la qualité de notre modèle développé.

L'analyse des résidus a permis de détecter les observations aberrantes précisées dans les deux dernières colonnes.qui devrait se traduire par une diminution de Q^2, ce qui n'est pas observé (colonnes (III) et (IV)).

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par "leave – one - out". Les valeurs de Q^2 obtenues (proches de 100 % - colonne (III)) reproduisent pratiquement celles du coefficient de détermination multiple correspondant (colonne (I)), ce qui fait ressortir la qualité de l'ajustement du notre modèle obtenu pour chaque méthode (MLR / RNA).

Les valeurs RMSE réunies dans la colonne (VI) sont elevées dans le cas du modèle (MLR) par contre ils sont faibles et proches dans le cas de (RNA), ce qui permet,de s'assurer de la bonne capacité prédictive de ce dernier.

Le test de randomisation montre, que le modèle obtenu n'est pas dus au hasard.

Ce travail sera étendu à d'autres composés à hauts poids moléculaires et, en diversifiant la structure des ensembles de données.

Enfin, d'autres méthodes (comme la logique floue …….) qui peuvent s'avérer plus avantageuses pour la précision des modèles, et du point de vue de la capacité de généralisation, seront testées.

Annexe 1: Liste des descripteurs

N°	COMPOSE	MATS۳e	VRp۲	Mor۱۵u	Mor۲۱m	Mor۲۵m	Mor۱۹e	SPP
1	2,6-dimethylphenol	0.041	3.421	0.441	-0.131	0.191	0.695	0.471
2	2,6-di-tert-butylphenol	-0.006	4.455	0.669	0.104	-0.028	2.027	0.486
3	phenol	-0.108	2.871	0.024	-0.108	0.333	0.236	0.47
4	o-cresol	-0.031	3.165	0.228	-0.104	0.267	0.534	0.473
5	2,3,6-trimethylphenol	0.024	3.648	0.459	-0.112	0.138	0.799	0.471
6	2,4,6-tri-tert-pentylphenol	0.037	3.651	0.397	-0.212	0.145	0.625	0.473
7	(R)-2,3-dihydro-1H-inden-1-ol	0.031	5.415	1.468	-0.039	-0.338	3.019	0.49
8	2-ethylphenol	-0.03	3.812	0.414	-0.304	0.302	0.753	0.523
9	2,4-dimethylphenol	0.025	3.39	-0.028	-0.135	0.213	0.737	0.47
10	m-cresol	-0.124	3.156	0.175	-0.11	0.264	0.394	0.47
11	2-isopropylphenol	-0.048	3.416	0.362	-0.107	0.198	0.646	0.468
12	2,3-dihydro-1H-inden-2-ol	-0.124	3.155	0.238	-0.133	0.262	0.399	0.47
13	3,5-dimethylphenol	0.012	3.589	-0.011	-0.101	0.17	0.963	0.471
14	3-ethylphenol	-0.149	3.797	0.277	-0.246	0.279	0.783	0.521
15	2,3,5,6-tetramethylphenol	-0.034	3.421	0.153	-0.178	0.197	0.395	0.476
16	(R)-2-sec-butylphenol	-0.136	3.408	0.366	-0.158	0.194	0.485	0.47
17	2-tert-butylphenol	-0.056	3.377	-0.086	-0.121	0.232	0.689	0.471
18	2-isopropyl-5-methylphenol	0.034	3.863	0.402	-0.271	0.061	0.614	0.476
19	(R)-1,2,3,4-tetrahydronaphthalen-1-ol	0.036	3.758	-0.417	-0.101	0.231	1.399	0.477
20	4-isopropylphenol	-0.04	3.782	0.122	-0.06	0.087	0.93	0.469
21	5-isopropyl-2-methylphenol	-0.006	3.801	0.035	-0.143	0.092	1.083	0.471
22	2,3,5-trimethylphenol	-0.011	3.946	-0.542	-0.258	0.078	1.038	0.529
23	2-tert-butyl-4-methylphenol	-0.062	3.578	-0.041	-0.075	0.145	0.895	0.47
24	4-propylphenol	-0.006	3.793	0.095	-0.119	0.111	1.065	0.469
25	4-tert-butylphenol	-0.048	3.647	0.376	-0.218	0.127	0.552	0.47
26	(R)-chroman-4-ol	-0.054	3.982	0.211	-0.03	0.02	0.941	0.469
27	3,4,5-trimethylphenol	-0.033	3.554	-0.189	-0.113	0.249	0.904	0.469
28	2,3-dihydro-1H-inden-4-ol	-0.109	3.77	0.315	-0.109	0.149	0.952	0.469
29	4-tert-pentylphenol	-0.033	3.751	-0.354	-0.058	0.129	1.136	0.47
30	2,3,4,5-tetramethylphenol	-0.109	3.965	0	-0.208	0.352	0.757	0.521

Annexe 1: Suite et fin

N°	COMPOSE	MATS۳e	VRp۲	Mor۱۵u	Mor۲۱m	Mor۲۵m	Mor۱۹e	SPP
31	2,3-dihydro-1H-inden-5-ol	-0.121	3.648	0.019	-0.108	0.16	0.929	0.47
32	6-methyl-2,3-dihydro-1H-inden-4-ol	0.016	3.831	0.456	-0.241	0.336	0.709	0.467
33	5,6,7,8-tetrahydronaphthalen-1-ol	-0.052	4.143	-0.21	-0.064	0.026	1.674	0.471
34	7-methyl-2,3-dihydro-1H-inden-5-ol	-0.07	3.942	0.172	-0.137	0.134	1.258	0.469
35	5,6,7,8-tetrahydronaphthalen-2-ol	-0.036	3.865	0.164	-0.181	0.072	0.865	0.469
36	2-cyclohexylphenol	-0.052	3.821	0.542	-0.247	0.287	0.737	0.471
37	biphenyl-3-ol	-0.006	4.04	0.607	-0.257	0.276	0.914	0.466
38	biphenyl-4-ol	0.005	4.054	0.445	-0.244	0.25	0.846	0.466
39	2,5-dimethylphenol	0.029	3.967	-0.614	-0.174	0.098	1.194	0.474
40	3,4-dimethylphenol	-0.063	4.042	0.595	-0.274	0.207	0.871	0.471
41	4-ethylphenol	-0.033	3.963	-0.633	-0.189	0.097	1.14	0.47
42	2-methylnaphthalen-1-ol	0.04	4.223	-0.809	-0.129	0.106	1.714	0.47
43	naphthalen-2-ol	-0.211	3.858	0.339	0.054	0.351	0.114	0.469
44	3-isopropylphenol	-0.076	4.214	0.293	-0.302	0.535	0.185	0.471
45	2,4,6-trimethylphenol	-0.076	4.219	0.307	-0.326	0.615	0.1	0.468
46	p-cresol	0.015	3.98	0.335	-0.04	0.021	1.401	0.473
47	2,3-dimethylphenol	-0.048	3.414	0.395	-0.139	0.186	0.749	0.468
48	(R)-4-sec-butylphenol	0.041	3.561	-0.12	-0.139	0.2	1.008	0.47
49	3,5-diisopropylphenol	-0.123	3.415	0.081	-0.11	0.216	0.69	0.47
50	7-methyl-2,3-dihydro-1H-inden-4-ol	-0.056	3.38	-0.125	-0.105	0.267	0.641	0.47
51	benzo[d][1,3]dioxol-5-ol	-0.062	3.576	-0.072	-0.11	0.173	0.885	0.47
52	2-tert-butyl-6-methylphenol	-0.109	3.769	0.204	-0.162	0.118	0.98	0.471
53	3-tert-butylphenol	0.046	4.171	0.591	-0.304	0.539	0.416	0.469
54	naphthalen-1-ol	-0.011	3.97	0.386	-0.256	0.639	0.293	0.471
55	2-propylphenol	-0.091	3.959	0.44	-0.279	0.667	0.178	0.469
56	biphenyl-2-ol	0.002	4.223	0.352	-0.255	0.468	0.474	0.475

REERENCES BIBLIOGRAPHIQUES

١- Kaliszan. Anal Chem 64, 619A (1992).

٢- E. Cornwell. Bol. Soc. Chil. Quim. 45, 649 (2000).

٣- E. Cornwell. Bol. Soc. Chil. Quim. 47, 53 (2002).

٤- C. F. Poole and S. K. Poole "Chromatography today" Edit Elsevier (1991).

٥- Report for Analytical Chemistry. Anal. Chem. 36, 31A (1964).

٦- Curt M.white*[1] and Norman C.Li Anal Chem 54,1564-1570 (1982).

7- Hyperchem™ Release 7.5 for windows, Molecular Modelling system, 2000.

8- R. Todeschini, V. Consonni, M. Pavan, DRAGON, Software for the calculation of Molecular Descriptors. Release 5.3 for Windows, Milano, 2005.

9- E-calc computes all the E-stat values and displays them in a convent form of the screen. The computational parts of this program have been taken from:

-Molconn-Z™ software. Lowell H., Hall Associates Consulting, 2 Devis street, Quincy, MA02170 for DOS version only.

-Sci. QSAR™ 2D. Sci. Vision, Inc, 200 Wheeler Rd, Burlington, MA, 01803 for PC versions.

10- H.Kubinyi ,Quant. Struct. – Act. Relat., 13, 1994, 285.

11- D.C. Montgomery, E.A. Peck, Introduction to linear Regression Analysis, Second Edition, Wiley-Interscience Publication, New York, 1992.

12- Mc Culloch-Pitts. a logical Calculus at the ideas imminent in Nervous Activity. Bulletin at math. Biophysics.1943,Vol. 5, p.115-133.

13- M. Minsky,S. Papert, Perceptrons. Massachusetts: MIT press, 1969.

14- D. E. Rumelbart, J. L. Mcclelland et al . Parallel Distributed processing Vol. 1. Massachusetts: MIT press, 1988. 547 p.

15- J. J. Hopfield. Neural Networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of sciences. USA. 1982. Vol.79. p. 2554-58.

16- T. Kohonen Self-organization and associative memory. Bulletin: Springer-Verlag. 984.

17- R. Hecht-Nielson Neurocomputing. Addison-Wesly Publishing Company. 1990. 433 p.

18- F. Fogelmen-Soulié. Méthodes connexionnistes pour l'apprentissage. Actes des journées Nationales sur l'intelligence Artificielle. Paris: Teknea. 1988. p. 275-293.

19- K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks, 4 (1991). 251-257.

20- Matlab Version 7.0.0.19920 (Release 14) The Language of Technical Computing The MathWorks, Inc. May 06, 2004.

21- N.R Draper, H. Smith, Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New york, 1998.

22- L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Perspective, 111(10), (2003), 1361-1375.

23- D.A. Belsley, E. Kuh, R.E. Welsch, Regression Diagnostics : Identifying Influential Data and Sources of Collinearity, Wiley, New York, 1980.

24- D. Cook, Detection of Influential Observations in linear Regression. Technometrics. 19, (1977), 15-18.

25- S. Weisberg, Applied linear Regression. J. Wiley, Inc., New York, 1980.

26- R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, R. Mannhold, H. Kubinyi, H. Timmerman eds., Wiley- VCH, Verlage Gmbh, Weinheim, 2000.

27– R. Todeschini, D. Ballabio, Consonni, A. Mauri and M. Pavan Milano Chemometrics and QSAR Research Group Moby Digs Professional – Version 1.0 – 2004.