وزارة التعليم العالي و البحث العلمي

UNIVERSITE BADJI MOKHTAR-ANNABA
BADJI MOKHTAR UNIVERSITY -ANNABA

جامعة باجي مختار- عنابة-
Année 2016

Faculte des Sciences
Departement de Chimie

THÈSE
Presentée pour obtenir le diplôme de Doctorat en Sciences
Option : Chimie Analytique et Environnement

THEME

MODELISATION DE QUELQUES PROPRIETES (cteH, S, Pv, Koc(w)) CONTROLANT L'EVOLUTION DANS L'ENVIRONNEMENT D'UNE SERIE D'HERBICIDES

Par : M^lle BOUAKKADIA Amel

Devant le jury

Membres de Jury:

Présidente :	**M^me Salima ALI- MOKHNACHE**	**Professeur, Université d'Annaba**
Directeur de thèse :	**Mr Djelloul MESSADI**	**Professeur, Université d'Annaba**
Examinatrice :	**M^me Chahra HAIOUR - BIDJOU**	**Professeur, Université d'Annaba**
Examinatrice :	**M^me Amel MESSAI**	**MCA, Université de Khenchla**
Examinateur :	**MrMekki KADRI**	**Professeur, Université de Guelma**
Examinateur :	**Mr Noureddine ZENATI**	**MCA, Université de Souk- Ahras**

Dédicaces

A ma très chère mère

Amon père

A ma sœur Hayett

A mon frère Khaled et sa femme Imene

A ma grand-mère

A mes tantes et oncles

A mes cousines et cousins

A tous mes amis surtout Aicha, Imen, Mahira & Hamza.

Remerciements

Mes premiers remerciements vont à Monsieur le professeurMESSADIDjelloul, qui m'a fait l'honneur d'accepter de diriger ce travail.Pour son encadrement scientifique. Merci de m'avoir guidé avec patience et d'avoir consacré autant d'heures pour les corrections de ce manuscrit.

Mes vifs remerciments s'adressent à madame ***Salima ALI-MOKHNACHE*** *pour m'avoir fait l'honneur de présider mon jury de thèse*

Je tiens à remercier aussi les membres du jury, les professeurs : madame ***Chahra HAIOUR – BIDJOU*** *et monsieur* ***Mekki KADRI****, les maîtres de conférences madame* ***Amel MESSAI*** *et monsieur* ***Noureddine ZENATI****;pour avoir accepté de faire partie de mon jury de thèse en donnant de leur temps pour la lecture et la discussion de mon travail. Ils le valorisent à travers leurs remarques et conseils judicieux.*

Je tiens à exprimer mes sincères gratitudes à tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Résumé

Des modèles QSPR ont été développés pour la prédiction de cinq caractéristiques environnementales importantes d'un ensemble hétérogène de pesticides. Des approches basées sur la régression linéaire multiple (RLM), les moindres carrés partiels (PLS), les machines à vecteurs supports (SVM) et les réseaux de neurones artificiels (RNA), chaque fois en association avec une sélection des variables les plus importantes par algorithme génétique (GA), conduisent à des modèles de qualités différentes.

L'approche hybride algorithme génétique/ régression multilinéaire a été utilisée pour modéliser la constante de Henry (logH) de 27 pesticides appartenant à deux classes chimiques: triazines, et carbamates. Le modèle à 4 variables explicatives sélectionné est robuste, et présente une bonne qualité de l'ajustement ainsi que de bonnes capacités prédictives.

Une relation quantitative structure-propriété (QSPR) a été réalisée pour la prédiction de la solubilité aqueuse des pesticides appartenant aux quatre classes chimiques: acide, urée, triazine, et carbamate. L'ensemble des 77 pesticides a été divisé en un ensemble de calibrage de 58 pesticides et un ensemble de test de 19 pesticides selon la technique Snee. Un modèle à six descripteurs, avec un coefficient de détermination (R^2) de 0,8895 et une erreur standard de l'estimation (s) de 0,52, a été développé en appliquant une analyse de régression linéaire multiple baséeles moindres carrés ordinaires et les algorithmes-génétiques pour la sélection des sous-ensembles de variable. La fiabilité du modèle proposé a été en outre illustrée en utilisant diverses techniques d'évaluation: validation croiséepar leave- one- out, bootstrap, tests de randomisation, et validation extene sur l'ensemble de test.

Les études QSPR sont une alternative théorique puissante à la mécanique quantique pour la description et la prédiction des propriétés des systèmes moléculaires complexes dans divers environnements.

Le coefficient de partage octanol / eau (K_{ow}) de certains pesticides a été étudié parapproche QSPR hybride : algorithme génétique / régression linéaire multiple.

La robustesse et la performance prédictive des modèles proposés ont été validéesà l'aide de statistiques internes et externes. Un point influent qui renforce le modèle et une valeur aberrante ont été mis en évidence.

Dans notre étude, nous avons également utilisé la méthode des « supportvectormachines »enexploitant la fonction de RBF, pour les valeurs optimales des paramètres SVM, C = 10; γ = 0,5; et ε = 0,2 ; le modèle résultant conduit à de bonnes capacités prédictives internes et externes.

La modélisation de la pression de vapeur par la méthode des moindres carrées partiels nous a permis d'éliminer les autocorrélations des descripteurs La comparaison de la qualité des modèles RLM et PLS pour la pression de vapeur, les deux modèles RLM et PLS sont acceptable tant par la qualité de l'ajustement, la robustesse ou la capacité prédictive.

La modélisation de la pression de vapeur et du coefficient de partage octanol/carbone organique d'un mélange hétérogène de pesticides montre que les différentes statistiques établies pour les ensembles de calibrage et de validation (coefficients de détermination multiple et de prédiction; racines des erreurs quadratiques moyennes) attestent de la supériorité des modèles non linéaires (RNA), ainsi que de leur pertinence.

Mots clés: *pesticides-constante de Henry- solubilité aqueuse- coefficient de partage octanol/eau et octanol/ carbone organique- pression de vapeur- QSPR- descripteurs moléculaires- régression linéaire multiple- les supports vecteurs machines- les réseaux de neurones artificiels*

Abstract

QSPR models were developed for the prediction of five important environmental characteristics of an heterogeneous set of pesticides. The approaches based on multilinear regression (MLR), partial least squares (PLS), support vectors machines (SVM) and artificial neural networks (ANN), every time associated with genetic algorithm (GA) selection of the most important variables, lead to models of very different qualities.

Genetic algorithm/ multi- linear hybrid approach was used to model the Henry constant (logH) of 27 pesticides belonging to two chemical classes: triazines, and carbamates. The 4 explanatory variables model selected is robust and has good fitness and good predictive ability.

A quantitative structure- property relationship (QSPR) was performed for the prediction of the aqueous solubility of pesticides belonging to four chemical classes: acid, urea, triazine, and carbamate. The entire set of 77 pesticides was divided into a training set of 58 pesticides and a test set of 19 pesticides according to the Snee technique. A six descriptor model, with squared correlation coefficient (R^2) of 0.8895 and standard error of estimation (s) of 0.52 log unit, was developed by applying multiple linear regression analysis using the ordinary least square regression method and genetic algorithm- variable subset selection. The reliability of the proposed model was further illustrated using various evaluation techniques: leave- one- out cross- validation, bootstrap, randomization tests, and validation through the test set.

QSPR studies become a powerful theoretical alternative to quantum mechanics for the description and prediction of the properties of complex molecular systems in various environments.

The octanol/water partition coefficient (K_{ow}) of some pesticides has been studied searching for quantitative structure- property relationships (QSPRs). Genetic algorithm/ multi- linear hybrid approach was used to model the octanol/water partition coefficient (K_{ow}).

The robustness and the predictive performance of the proposed models were verified using both internal and external statistical validation. One influential point which reinforces the model and an outlier were highlighted.

In our study we also used the method of support vector machines using the RBF function, for the optimal values of SVM parameters, C = 10; $\gamma = 0.5$; and $\varepsilon = 0.2$, the resulting model leads to good internal and external predictive abilities.

The modeling of vapor pressure by the partial least square method has allowed us to eliminate the autocorrelation descriptors. Comparing the quality of MLR and PLS models for vapor pressure, both MLR and PLS models are both acceptable: quality adjustment, robustness and predictive capacity.

Themodeling of the vapor pressure and the octanol / organic carbon of a heterogeneous mixture of pesticides show that the various statistics for the sets of calibrage and validation (multiple coefficients of determination and prediction; roots of squared errors averages) attest to the superiority of non-linear models (RNA) and their relevance.

Key Words: *pesticides- Henry constante - aqueous solubility - octanol/eau and octanol/ carbone organique partage coefficient- vapor pressur - QSPR- molecular descriptors - multiple. linéaire régression support vectors machines -artificial neural networks*

الملخص

تم تطوير نماذج بطريقة QSPR الـ لتنبؤ5 خصائص بيئية هامة لمجموعة غير متجانسة من المبيدات.عدة نهج تقوم على نموذج التراجع المتعدد الخطي، المربعات الصغرى الجزئية-شعاعالاصطناعيو نموذج الشبكة العصبونية، في كل مرة نشركها باختيار مجموعة من المتغيرات المهمة باستعمال الخوارزمية الجينية، أدت إلى نماذج مختلفة الدرجات.

تم استخدام خوارزمية وراثية لنهج التراجع الهجين / التراجع المتعدد الخطي من اجل التنبؤ لثابت هنريlogH الخاص ب27.مبيد يضم فئتينكيميائية : الثريازبن-الكربامات- النموذج المعرف ب4المتغيرات المختارة قوي ولديه نوعية جيدة فيما يخص القدرة على التنبؤ.

اجريت علاقة بين كمية الهيكل و الخاصية للتنبؤبانحلالية المبيدات منتمية إلى أربعة أقسام كيميائية: أحماض، اليوريا، تريازين، وكاربامات. المجموعة المكونة من 77 مبيد قسمت إلى مجموعة بناء من 58 مبيد ومجموعة اختبار من 19 مبيد بتقنية Snee. النموذج بستة متغيرات بمعامل ارتباط (R^2)يساوي 0.8895 و خطا معيار التقدير (s) يساوي 0.52 وحدة تم تطويره بتطبيق التراجع المتعدد الخطي باستخدام المربعات الصغرى واختيار مجموعة المتغيرات تم باستعمال الخوارزمية الجينية. قوة النموذج المقترح تأكدت باستخدام عدة تقنيات للتقييم leave- one- out ، bootstrap ،الاختبارات العشوائية، والتحقق من خلال مجموعة الاختبار.

دراسات QSPR اصبحت البديل النظري القوي لوصف و التنبؤ خصائص الجزيئية المعقدة في مختلف بيئات.

معامل تقسيم الأوكتانول /المياه(K_{ow}) لبعض المبيدات درس لبحث عن علاقات الكمية.بين هيكل و الخاصية (QSPRs) الخوارزمية الجينية / متعددة النهج الهجين الخطية استخدمت في تصميم نموذج لمعامل تقسيم الأوكتانول/المياه.(K_{ow})

التحقق من متانة و أداء التنبؤية للنماذج المقترحة استخدم على الصعيدين الداخلي و الخارجي.

في دراستنا كدلك قمنا باستعمال طريقة آلات ناقلات الدعم باستخدام وظيفة RBF والقيم المثلى للمعلمات منSVM، C=10؛ .$\gamma = 0.5$ و$\varepsilon = 0.2$،،ونموذج الناتج عنها يؤدي إلى قدر اتتنبؤية الداخلية و الخارجية جيدة.

ويبين نماذج من ضغط البخار و معامل تقسيم الأوكتانول / الكربون العضوي من خليط غير متجانس من المبيدات أن الإحصاءات المختلفة لمجموعات من المعايرة و التحقق (معاملات متعددة من التصميم و التنبؤ؛ جذور الأخطاء مربع المتوسطات) تشهد على تفوق النماذج غير الخطية (RNA) وأهميتها.

الكلمات الدالة : *المبيدات* ، لثابت هنري*-الانحلالية*-معامل تقسيمالأوكتانول/ كربون عضوي - الأوكتانول/المياه- *QSPR* - *الموصفات الجزيئية- التراجع المتعدد الخطي*.-ماكيناتناقلاتالدعم-الشبكاتالعصبونيةالاصطناعية.

Sommaire

Partie II: Application

SYMBOLES ET ABREVIATIONS

ACP: Analyse en composantesprincipales.

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: Ecart quadratique moyen.

EQMC: Ecart quadratique moyen calculé sur l'ensemble de calibrage.

EQMP Ecart quadratique moyen de prédiction.

$EQMP_{ext.}$: Ecart quadratique moyen calculé sur l'ensemble de validation externe.

e_i : Résidu : différence entre les valeurs observée (y_i) et estimée ($\hat{\mathbf{y}}_\mathbf{i}$).

F : Statistique de Fisher.

FIT: Fonction de KUBINYI.

FIV: Facteur d'inflation de la variance.

GA: Algorithme génétique (Genetic Algorithm).

hii : Eléments diagonaux de la matrice chapeau.

log cte H: logarithme de la constante de HENRY.

log S: logarithme de la solubilité.

log Pv: logarithme de la pression de vapeur.

logKoc: logarithme du coefficient de partage octanol/carbone organique

logKow: logarithme du coefficient de partage octanol/eau

LMO: Cross-validation by leave-many-out: Validation croisée par omission d'un ensemble d'observations.

LOO: Cross-validation by leave-one-out: Validation croisée par omission d'une observation

n: Dimension de la population (échantillon).

n-p : Nombre de degrés de liberté.

PLS(ou MCP): Moindres carrés partiels.

PRESS : Somme des carrés des erreurs de prédiction.

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

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

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

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

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

RLM (MLR): Régression linéaire multiple.

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

RNA: Réseaux de neurones artificiels.

R^2 : Coefficient de détermination.

r_i : Résidu studentisé interne.

S : Erreur standard.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

SVM : Support vector machine (machine à vecteur support)

t : t de Student.

t_i : Résidu studentisé externe.

yi: Valeur observée.

$\hat{y}_i$: Valeur estimée.

$\hat{y}_{(i)}$: Valeur prédite.

% OC pourcentage en carbone organique

Liste des tableaux

Liste des figures

Introduction générale

Depuis les années 1950, le développement de l'agriculture et la volonté d'augmenter les rendements ont conduit à une utilisation croissante des pesticides. Ces pratiques ont causé, suite à une utilisation massive, une contamination de l'environnement et en particulier une pollution diffuse d'un grand nombre d'aquifères sur l'ensemble du territoire. L'Algérie utilise environ 6 000 à 10 000 tonnes de pesticides par an (Moussaoui K M, 2001) (environ 400 produits phytosanitaires sont homologués en Algérie (Bouziani M, 2007)). Si l'intérêt s'est porté dans un premier temps sur la contamination du réseau de surface en raison de fortes concentrations détectées précocement, le problème des eaux souterraines est aujourd'hui pertinent. En effet, l'eau souterraine est d'une importance capitale dans la plupart des régions du monde. Toutefois, cette ressource qui était jadis de bonne qualité, se trouve actuellement menacée par diverses sources de contamination ponctuelles et diffuses.

Les sources ponctuelles ont fait l'objet de nombreux travaux et recherches sur le terrain et en laboratoire. Par contre, les sources diffuses et particulièrement la contamination par les pesticides (herbicides, fongicides, …) en zones agricoles, n'ont attiré l'attention des scientifiques et gouvernements que depuis la fin des années 1970 où des analyses d'eau souterraine, de surface et de drainage révélaient la présence des traiazines (herbicide) et d'autres produits phytosanitaires agricoles.

En raison de sa position d'interface dans l'environnement entre l'atmosphère et les eaux souterraines, le sol joue un rôle déterminant dans le devenir des herbicides. C'est un mélange hétérogène composé de nombreux constituants (matières organiques et inorganiques) dont la composition et l'activité de surface sont variables. Il apparaît donc nécessaire d'étudier le devenir des herbicides dans les sols afin de mieux en mesurer l'impact environnemental.

Si un pesticide est relativement soluble, son transfert de la zone d'utilisation au système aquatique, puis sa répartition dans celui- ci en est d'autant plus rapide. Au contraire, les pesticides insolubles mettent plus de temps pour atteindre le milieu aquatique et leur diffusion dans l'eau est très vite limitée par fixation sur les matières en suspension ou sur les sédiments. Ils affecteront alors la biologie du milieu uniquement, si le produit est très toxique, par

accumulation dans les tissus ou par fixation sur la matière organique servant d'aliment aux différents organismes.

Abstraction faite de la stabilité chimique et de la biodégradation, la diminution de la concentration en pesticides reste essentiellement liée à leurs caractéristiques physico-chimiques comme la pression de vapeur ou la constante de Henry.

Différents modèles permettent de prévoir la solubilité dans l'eau, la pression de vapeur ou la constante de Henry. Les méthodes incrémentielles sont basées sur des caractéristiques structurelles comme le type d'atome, le type de liaison et l'environnement structurel local ; d'autres modèles de régression mettent en jeu des propriétés physico- chimiques, des descripteurs structuraux comme les indices de connectivité, et des descripteurs reflétant la structure électronique. Signalons en plus, pour la solubilité dans l'eau et la constante de Henry, la possibilité d'utiliser des modèles basés sur la structure moléculaire et les modèles quantiques de solvatation (via l'enthalpie libre de solvatation ΔG_s). Les résultats de ces modèles révèlent des différences substantielles dans les domaines d'application et dans les capacités de prédiction.

Notre travail a porté sur la modélisation de quelques propriétés physico- chimiques: solubilité dans l'eau, pression de vapeur, coefficient de partage octanol/eau et octanol/ carbone organique, et constante de Henry, d'une série hétérogène de pesticides appartenant à des familles différentes. Nous avons utilisé une approche QSPR hybride associant algorithme génétique pour la sélection de sous-ensembles de variables significatives parmi quelques 2000 calculées théoriquement, et soit une régression linéaire (RLM, PLS), soit une régression non linéaire (RNA, SVM).

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

Dans la partie Généralités, après une disgression sur l'état de l'art des méthodes QSAR/QSPR, 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, analyse en composantes principales, régression en composantes principales, moindres carrés partiels, réseaux de neurones

artificiels, traitement statistique pour l'évaluation de la qualité de l'ajustement (robustesse des modèles ; détection des observations aberrantes et/ou influentes ; test de randomisation dans le cas des régression linéaires ; validation externe ; définition des domaines d'application).

Dans la Partie Application, nous présentons et discutons les modèles calculés.

I. Généralités sur les pesticides :

Le terme pesticide est un mot anglais, daté du milieu du XX^{ème} siècle qui fit son apparition en France en 1959 comme substantif. Ce mot se compose de pest, signifiant insecte ou plante nuisible ou encore parasite, qui fut lui-même emprunté au mot français peste au XVI^{ième} siècle et de –icide, du latin caedere signifiant frapper, abattre, tuer. Ce mot en français est mal formé puisque peste ne signifie pas parasite (Boivin A, 2003).

L'utilisation des pesticides a commencé avec l'existence de l'Homme. Il apprend au fur et à mesure à promouvoir les espèces exploitées et à contrer leurs compétiteurs. Qu'elle soit avant et après la récolte, la lutte contre les nuisibles tant des cultures que des stocks se fitd'abord par des procédés physiques et manuels, puis par des méthodes chimiques. On note, au moyen âge, la première utilisation de sels d'arsenic comme insecticides.

Jusqu'à la première moitié du XX^{ème} siècle, les produits phytosanitaires sont essentiellement des dérivés minéraux (arsenic, sulfate de cuivre) ou végétaux (roténone, pyrèthre), donc le développement des pesticides a ensuite suivi celui de la chimie minérale.

Dès la seconde guerre mondiale, les pesticides ont profité du développement rapide de la chimie organique. Les composés synthétiques, qui sont majoritaires, ont d'ailleurs été à l'origine de l'expansion rapide des pesticides à partir des années 1940. La forte utilisation des pesticides était considérée comme unpréalable à la réussite d'une stratégie de développement agricole rapide. Cetteconception a favorisé l'utilisation importante des produits phytosanitaires afind'augmenter la production agricole, surtout celles des cultures destinées à l'exportation (Fleischer G *et al.,* 1998).

Dans les années 1950, des insecticides comme le dichlorodiphényldichloroéthane (DDD) et le dichlorodiphényltrichloroéthane (DDT) sont utilisés en grandes quantités en médecine préventive (pour détruire le moustique responsable de la malaria) et en agriculture (élimination du doryphore). D'autres biocides sont mis au point pour l'industrie textile et du bois, pour les usages domestiques (aérosols tue-mouches), pour l'entretien des routes et pour une utilisation en médecine.

L'usage de ces produits a connu un très fort développement au cours des décennies passées, les rendant quasiment indispensables à la plupart des pratiques agricoles, quel que soit le niveau de développement économique des pays.

Si les pesticides ont constitué un énorme progrès dans la maîtrise des ressources alimentaires et l'amélioration de la santé publique (en particulier dans la lutte contre les insectes, vecteurs des maladies), le revers de la médaille est apparu rapidement: des phénomènes de résistance chez les insectes, puis des troubles de la reproduction chez les oiseaux, ont montré de façon spectaculaire les limites et les dangers de ces substances pour l'environnement , pour les écosystèmes mais également pour les êtres humains.

Il ne faut pas toutefois perdre de vue, que les pesticides ont constitué un énorme progrès pour l'agriculture et ont permis d'assurer une production alimentaire de qualité. L'augmentation des rendements des terres agricoles a permis de limiter la déforestation, ainsi les experts estiment que leur utilisation, en 50 ans, a permis de préserver 50% de la surface de la forêt actuelle (Kleter G A *et al.,* 2007).

Leur utilisation a également permis d'éradiquer un grand nombre de maladies parasitaires très meurtrières, ou d'en limiter la propagation.

I. 1- Définitions des pesticides :

Dans les textes relatifs à la réglementation nationale et européenne, les pesticides sont aussi appelés « produits phytosanitaires» (Colas A, 1971).

Les définitions des pesticides sont nombreuses, nous retiendrons celle de Colas (Colas A, 1971) qui est satisfaisante par sa généralité : « on appelle pesticide, toute substance naturelle ou synthétique, qu'elle soit ou non mélangée à d'autres produits (support, adjuvant, tensio-actif) utilisée :

- Dans la lutte contre les vecteurs des maladies humaines et animales, à l'exclusion des médicaments,
- Pour la lutte contre les ennemis des plantes et des récoltes,
- Pour la protection des matériaux et produits stockés ou mis en œuvre ».

La directive européenne 91/414/CE du 15 juillet 1991 (Directive européenne, 1991) concernant la mise sur le marché de **produits phytosanitaires**, les définit comme étant :

« Les substances actives et les préparations contenant une ou plusieurs substances actives qui sont présentes sous la forme dans laquelle elles sont livrées à l'utilisateur et qui sont destinées à :

- Protéger les végétaux ou les produits végétaux contre tous les organismes nuisibles ou à prévenir leur action,
- Exercer une action sur les processus vitaux des végétaux, pour autant qu'il ne s'agisse pas de substances nutritives,
- Assurer la conservation des produits végétaux, pour autant que les substances ou produits ne fassent pas l'objet de dispositions particulières du Conseil ou de la Commission concernant les agents conservateurs,
- Détruire les végétaux indésirables,
- Détruire les parties des végétaux, freiner ou prévenir une croissance indésirable des végétaux ».

Les **pesticides** sont définis plus simplement comme des substances dont les propriétés chimiques contribuent à la protection des plantes cultivées et des produits récoltés.Ils sont caractérisés par leur stabilité et leur résistance aux processus de dégradation dans l'environnement, ainsi que par leur tendance à s'accumuler dans les chaînes alimentaires (Marliere F, 2000).

Les pesticides regroupent plus de 1000 substances chimiques appartenant à près de 150 familles chimiques différentes.

Une famille chimique regroupe (Pesticides Manual, 2006) l'ensemble des molécules dérivées d'un groupe d'atomes qui constituent une structure de base. Cependant les propriétés des molécules ne dépendent pas uniquement d'un groupe donné. Elles résultent également de l'existence de motifs moléculaires particuliers (ex : noyaux aromatiques…) et de la présence d'atomes et/ou d'autres groupes fonctionnels (ex : groupe alcoolique). Ainsi il n'existe pas toujours de relation simple entre une famille chimique et les propriétés des substances qui la composent. Les pesticides constituent donc un ensemble de substances et de produits hétérogènes tant du point de vue de leurs propriétés physico-chimiques, que de celui de leur devenir dans l'environnement ou de leurs propriétés toxicologiques et écotoxicologiques.

I. 2- Historique des pesticides :

Depuis la révolution industrielle, l'exploitation des terres agricoles s'intensifie au rythme de la croissance exponentielle de la population mondiale. La mécanisation et la modernisation des techniques de travail ont favorisé l'augmentation de la production agricole répondant ainsi à une demande de plus en plus forte. En plus de ces progrès technologiques,

l'agriculture se dote aujourd'hui de produits chimiques plus performants afin de lutter contre l'infestation de mauvaises herbes. Dans le but d'augmenter la qualité et la production des récoltes, les agriculteurs épandent différents herbicides afin d'éliminer entièrement ou partiellement les parasites végétaux. Cette intervention est devenue inévitable, car les mauvaises herbes provoquent une compétition active avec les plantes cultivées en matière d'éléments nutritifs, d'eau et d''air (Edelahid M C , 2004).

L'utilisation de substances chimiques pour contrôler la végétation remonte à plus d'un siècle. Au XV^{ème} siècle, certains métaux toxiques non biodégradables tels que l'arsenic, le plomb et le mercure étaient utilisés sur les cultures pour éloigner les insectes. C'est en Allemagne, vers les années 1850, que la première substance herbicide vit le jour, un mélange de sel et de chaux fut alors utilisé (Fortier J C & Messier C, 2005).

De nombreux pesticides naturels tirés de plantes, aux propriétés insecticides particulièrement puissantes, sont couramment utilisés dès le milieu du XIX^{ème} siècle.

Certaines de ces substances végétales, notamment la nicotine et le pyrèthre (tiré des chrysanthèmes), sont encore utilisées de nos jours. Cependant, le désherbage chimique des cultures a véritablement débuté avec l'emploi de l'acide sulfurique dans la culture, préconisé en France dès 1911 en remplacement du sulfate de cuivre précédemment utilisé.

Les premiers désherbants organiques ont fait leur apparition en 1932 avec les dinitrophénols. Par ailleurs, la découverte et l'identification en 1934 de la première hormone végétale, l'acide indole-acétique, ont conduit pendant la deuxième guerre mondiale à la fabrication de produits de synthèse analogues (Flogeac K, 2004).

Le premier herbicide synthétique, l'acide 2,4-dichlorophénoxyacétique (2,4-D), fut conçu en 1945 (Edelahid M C, 2004, Perrin R & Scharff J P, 1997). A partir de ce moment, plusieurs substances chimiques ont été développées telles que les triazines (1955) et les chloroacétamides (1956). Dans les années du siècle passé, de nouvelles familles d'herbicides à faible dose d'application se sont développées, comme les sulfonylurées et les phosphonates (ayant des propriétés fongicides et herbicides) (Fdil F, 2004).

Aujourd'hui on compte plus de 30 000 types de mauvaises herbes dans le monde et plus de 200 groupes d'herbicides permettant de les contrôler. Les herbicides sont les

pesticides les plus utilisés dans le monde toutes cultures confondues (50 % du tonnage mondial en 2002, fongicides, 22% ; insecticides, 25% ; et divers, 3%) . Les herbicides représentent 60% des ventes totales mondiales de pesticides (Edelahid M C, 2004).

I. 3- Classification des pesticides :

Les pesticides disponibles aujourd'hui sur le marché sont caractérisés par une telle variété de structures chimiques, de groupes fonctionnels et d'activités que leur classification est complexe.

D'une manière générale, ils peuvent être classés en fonction de la nature de l'espèce à combattre mais aussi en fonction de la nature chimique de la principale substance active qui les compose(Kleter G A, 2007). Les systèmes de classification sont universels :

Le premier système de classification repose sur le type de parasites à contrôler. Il existe principalement trois grandes familles chimiques qui sont : les herbicides, les fongicides, et les insecticides.

I. 3- 1- Les insecticides :

Ce sont des matières actives organiques de synthèse, ils sont utilisés pour la protection des plantes contre les insectes. Ils interviennent en les éliminant ou en empêchant leur reproduction.

Certains insecticides ciblent le système nerveux des insectes. De ce fait, l'impact comportemental peut suffire à induire une mortalité (prédation, orientation,…) mêmeà des doses sublétales (Willy J & Peumans J M, 1995).

I. 3- 2- Les fongicides :

Ces molécules ciblent différents types de champignons pouvant directement infester les cultures à différents stades de développement. Les fongicides peuvent agir différemment sur les plantes ; on distingue :

- les inhibiteurs respiratoires,
- les inhibiteurs de la division cellulaire,
- les perturbateurs de la biosynthèse des acides aminés ou des protéines,
- les perturbateurs du métabolisme des glucides.

I. 3- 3- Les herbicides :

Sont des pesticides utilisés pour détruire des plantes ou interrompre leur développement normal. Au cours des quarante dernières années, les herbicides ont largement remplacé les méthodes mécaniques utilisées pour le contrôle des adventices. La notion d'adventice définit une espèce de plante se développant à un endroit où elle n'est pas désirée, ou en dehors de son lieu habituel de croissance, ou une plante offrant plus d'effets nuisibles que d'effets bénéfiques. Les herbicides sont un moyen, a priori, plus efficace que les anciennes techniques pour le contrôle des adventices. Leur utilisation a permis de réduire l'augmentation des coûts et de diminuer l'intensité deslabours. Suivant leur mode d'action, leur dose et leur période d'utilisation, ces composés peuvent être sélectifs ou non-sélectifs. Les herbicides sélectifs peuvent détruire certaines espèces de plantes et n'infliger que de faibles dégâts, voir aucun, à d'autres espèces. Ce principe repose sur de nombreux facteurs dont la dose d'application. Ainsi, l'atrazine employée à forte dose peut stériliser les sols alors que son utilisation courante se fait en tant qu'herbicide sélectif appliqué sur les cultures de maïs (Orlando S *et al.,* 1997).

Un herbicide, dans la plus large définition, est un composé qui est capable de massacrer les mauvaises herbes, il peut être employé pour limiter la croissance des plantes ou les éliminer (Jager G, 1983).Vers la fin des années 30 du siècle passé, beaucoup d'études avaient été lancées pour trouver les agents qui détruiraient sélectivement certaine espèce de plante. Plusieurs de ces produits chimiques étaient plus efficaces mais possédaient une considérable toxicité des mammifères.Cependant quelques composés de produits chimiques de prototype ont servi au développement ultérieur.

Dans les deux dernières décennies, les herbicides ont représenté une croissance très rapide de la plupart de la section des affaires de pesticides agrochimiques dueau mouvement dans les pratiques monoculturales, et à lamécanisation des pratiques agricoles (plantant, tendre, moissonnant) en raison du coût de la main-d'œuvre. Le taux de la croissance de production d'herbicide sur une base mondiale entre 1980 et 1985 étaient de 1,9 % par an, plus que le double du taux de croissance des insecticides pendant la même période (Marquis J K, 1982).

I. 3- 3- a- Composition :

Comme tous les autres pesticides, un produit herbicide correspond d'abord au nom commercial du produit commercialisé par un distributeur ou un fabricant. Ce produit commercial ou spécialité commerciale se compose de deux types de constituants : les matières actives qui lui confèrent son activité herbicide et les formulants qui complètent la formulation. Les formulants sont soit des charges ou des solvants qui n'ont qu'un rôle de dilution des matières actives, soit des produits qui améliorent la préparation (http://agroecologie-cirad.fr mars 2000).

- pour sa qualité :

la stabilité (émulsifiant, dispersif, etc…),

la présentation (colorant, parfum, répulsif, etc…),

la facilité d'emploi (vomitif, etc…),

- pour son comportement physique lors de la pulvérisation : mouillant, adhésif, etc…
- pour son activité biochimique : surfactant, phytoprotecteur (*safener*).

I. 3- 3- b- La formulation :

La formulation correspond à la forme physique sous laquelle le produit phytopharmaceutique est mis sur le marché ; obtenue par le mélange des matières actives et des formulants, elle se présente sous une multitude de formes, solides ou liquides. Les plus couramment répandues sont les suivantes :

- pour les formulations solides : les granulés solubles, les poudres mouillables
- pour les formulations liquides :
    - les concentrés solubles, composés de produits solubles dans l'eau,
    - les concentrés émulsionnables, composés de produits liquides en émulsion dans le produit,
    - les suspensions concentrées, appelées (parfois *flow* de l'anglais *flowable*), composées de particules solides dispersées dans le produit.

Le type de formulation a une grande importance dans la manipulation des produits : fabrication, transport, stockage, préparation des bouillies ; par exemple, les suspensions

concentrées auront tendance à sédimenter au cours du temps et il sera indispensable de les agiter avant l'emploi (Edelahid M C, 2004).

I. 3- 3- c- La caractérisation :

La caractérisation d'un produit herbicide signifie la désignation de la matière active, le nom du produit commercial, le fabricant et éventuellement du distributeur local, la teneur de la matière active dans le produit, le type de formulation, le mode d'emploi, la dose d'emploi et la culture cible.

La teneur en matière active s'exprime en g/l pour les formulations liquides et en pourcentage (%) pour les formulations solides. La dose d'emploi en produit commercial s'exprime en l/ha pour les formulations liquides et en kg/ha (ou parfois en g/ha) pour les formulations solides. La dose d'emploi en matière active s'exprime toujours en g/ha (EdelahidM C, 2004).

I. 3- 3- d- La classification :

Les herbicides peuvent être classés par leur structure chimique. La deuxième méthode de classification concerne comment et quand les agents sont appliqués : *Preplanting*, *Preemergent*, *Postemergent*.

Les herbicides exploités aujourd'hui sont d'origine minérale ou d'origine organique. Mais l'épandage moderne fait principalement appel aux composés organiques de synthèses. Chaque herbicide possède des caractéristiques propres selon sa composition, son mode d'absorption, son effet sur la mauvaise herbe et son élimination progressive (ScheyerA, 2000).

Les herbicides peuvent être répertoriés suivant leurs caractéristiques physico-chimiques selon les familles suivantes: (Scheyer A, 2000)

- **Les triazines :** ce groupe présente une structure cyclique. Les triazines (atrazine, simazine, métribuzine, …) sont en général peu solubles dans l'eau. Leur persistance peut ainsi atteindre 6 à 12 mois pour certains. Elles possèdent une grande stabilité chimique et sont assez fortement adsorbées sur le complexe argilo-humique, c'est pourquoi l'atrazine a été interdite en 2003 en France en raison de l'importance de la contamination des eaux.

- **Les acétamides :** comme l'Alachlore et le Métolachlore. Ces deux substances sont très similaires chimiquement du fait d'un groupement commun N-$COCH_2Cl$. Les propriétés physico-chimiques sont également semblables : ils présentent une forte solubilité dans l'eau et une pression de vapeur plutôt élevée.
- **Les aryloxyacides:** ces molécules sont constituées d'un noyau benzénique, naphténique ou anthracénique dont un des atomes d'hydrogène est substitué par un atome d'oxygène lié à une chaîne hydrocarbonée comportant un groupe carboxyle (CO_2H). Les aryloxyacides sont très polaires et peu volatils. Ces herbicides acides sont très solubles dans l'eau et ils se retrouvent sous leur forme dissociée à pH neutre.
- **Les urées :** les urées sont thermosensibles et sont facilement dégradées en isocyanates, leur dégradation est par contre lente dans l'environnement. Les urées sont assez persistantes et se retrouvent assez souvent dans les eaux.
- **Les toluidines** : comme la Trifluraline. Celle-ci est fortement adsorbée dans le sol. Sa demi-vie par évaporation à partir des surfaces de sol humide ou des eaux peu profondes varie de quelques heures à 50 heures. La photo-décomposition, la volatilisation et la dégradation microbienne sont les principaux processus responsables de l'élimination de la trifluraline dans les eaux de surface. La concentration maximale de la trifluraline dans l'eau potable est fixée à 0,045 mg/l (Document d'aide technique, 2003).

I. 3- 3- e- Modes d'action des herbicides :

Les herbicides se distinguent par leur voie de pénétration et leur mode d'action dans les végétaux: (CIRAD-CA GEC AMATROP, 2000)

- herbicides à pénétration racinaire : appliqués sur le sol, ils pénètrent par les organes souterrains des végétaux (racines, graines, plantules). Ce sont les traitements herbicides de prélevée, effectués avant la levée de la plante considérée (culture ou mauvaise herbe).
  - actions sur la photosynthèse : triazines, diazines – uraciles, triazinones, urées substituées (Fdil F, 2004).
  - action sur la division cellulaire : toluidines.
  - action sur l'élongation cellulaire : alachlore, métazachlore, métolachlor, etc.
  - inhibition de la synthèse des caroténoïdes : isoxaflutole, clomazone.

- herbicides à pénétration foliaire : appliqués sur le feuillage, ils pénètrent par les organes aériens des végétaux (feuilles, pétioles, tiges). Ce sont les traitements herbicides de post-levée, effectués après la levée de la plante considérée (culture ou mauvaise herbe).
    - actions sur la photosynthèse : bipyridyles, diazines.
    - actions sur les membranes cellulaires : dinitrophénols, benzonitriles.
    - action sur la division cellulaire : carbamates.
    - action sur l'élongation cellulaire : aryloacides, dérivés picoliniques.
    - action sur la biosynthèse : acides aminés, lipides.
- herbicides de contact : herbicides qui agissent après pénétration plus ou moins profonde dans les tissus, sans aucune migration d'un organe à un autre de la plante traitée.
- herbicides systémiques : herbicides capables d'agir après pénétration et migration d'unorgane à un autre de la plante traitée.

I. 3- 3- f- Sélectivité des herbicides :

Parmi les différents herbicides, certaines substances procurent un désherbage total en éliminant toute végétation qui se voit exposé et affecté par le produit chimique tandis que d'autres assurent un désherbage sélectif impliquant un seul type de mauvaises herbes sans que la culture saine en soit grandement affectée . La sélectivité peut être due à la morphologie de la plante et à la physiologie particulière de l'espèce.

La sélectivité des herbicides correspond à une modification d'au moins une des phases de l'action des produits dans la plante : mise en contact du produit avec la cible, pénétration, transport éventuel, site d'activité et métabolisme de dégradation. On distingue divers types de sélectivité :

- sélectivité de position : l'herbicide de prélevée, appliqué en surface, ne se répartit que dans la couche superficielle du sol à quelques centimètres de profondeur.
- sélectivité d'application : il s'agit d'éviter le contact du produit avec la plante cultivée lors de la pulvérisation.
- sélectivité anatomique : ce type de sélectivité concerne principalement les produits de post-levée : la pénétration par les feuilles peut être gênée par la présence de poils ou par l'épaisseur de la cuticule de l'épiderme.
- sélectivité physiologique : la sélectivité peut être obtenue par des différences de

comportement physiologique entre les végétaux.

I. 3- 3- g-Principales familles d'herbicides

- Les herbicides minéraux

Ils furent surtout utilisés au début du vingtième siècle. Les plus utilisés actuellement sont :

- le cyanure de calcium ($Ca(CN)_2$), il rentre par les racines et pénètre la sève brute pour ensuite s'accumuler dans les feuilles.
- le sulfate de fer ($FeSO_4$), herbicide de contact utilisé pour lutter contre les mousses et qui accélère de plus l'humification des déchets végétaux,
- le chlorate de sodium ($NaClO_3$) qui détruit les plantes à fort enracinement. Oxydant puissant, le chlorate de soude pénètre principalement par les racines et est transporté par la sève brute vers les feuilles. Son action n'est pas sélective et peut perdurer jusqu'à six mois dans la terre. Il est détruit par le calcaire, les matières organiques et les corps réducteurs, il peut être aussi lessivé par les eaux d'infiltration.

Il est peu toxique pour l'homme. Il peut être employé pour la dévitalisation des souches. Du fait de son danger (risque d'explosion), il est de plus en plus remplacé par des substances organiques.

- Les herbicides organiques

Ils constituent la très large majorité des herbicides du marché actuel. Par commodité, on les regroupe suivant leur type de pénétration dans le végétal :

- Le **glyphosate** : est un désherbant total, c'est-à-dire un herbicide non-sélectif. Le mécanisme d'action de ce pesticide est systémique. Il agit en bloquant l'enzyme enoylpyruvylshikimate 3-phosphate synthase (EPSPS). C'est un produit irritant et toxique, surtout connu sous la marque Roundup.
- **Les Urées Substituées (NH_2-CO-NH_2) :** ce sont exclusivement des herbicides. Leur absorption est essentiellement racinaire.Véhiculées par la sève brute, elles s'accumulent dans les feuilles où inhibent la photosynthèse. Elles ont une très faible solubilité dans l'eau et présentent une assez longue persistance d'action dans le sol (2 à 3 mois) mais variable selon les conditions écologiques rencontrées (sol, pluie, température). Elles ont une bonne action sur les graminées et sur certaines

dicotylédones. Elles sont utilisées en pré ou post-levée. Leur toxicité est quasiment nulle. Leur nom se termine par le vocable "uron".

Exemples : chlortoluron, chloroxuron, cycluron, diuron, éthidimuron, fénuron, isoproturon, linuron, monolinuron, méthabenzthiazuron, métobromuron, métoxuron, monuron, thiazafluron, tebuthiuron, thiazafluron, siduron, néburon …

- **Les triazines :** ce groupe présente une structure cyclique. Elles agissent par inhibition compétitive au niveau du photosystème II. EIles sont appliquées directement sur le sol. Le maïs est une plante très tolérante à l'atrazine. Le sorgho est également tolérant mais le blé et le soja y sont sensibles. Leur toxicité est faible et leur sélectivité souvent bonne. Leur solubilité dans l'eau est réduite et sont donc peu entraînées dans le sol.

Exemples : atrazine,cyanazine, méthoprotryne, propazine, terbuthylazine,simazine, simétryne, secbumeton, secbumeton, terbuméton, amétryne, desmétryne, prométryne,terbutryne…

- **Les sulfonylurées :** elles agissent sur la même enzyme que les imidazolinones, l'acétolactatesynthase (ALS)

Exemples : amidosulfuron, azimsulfuron, chlorsulfuron…

- **Les diphényls-éthers :** synthétisées à partir de 1964, ces molécules possèdent 2 noyaux benzènes reliés par un oxygène. Ils sont absorbés par les feuilles et les racines. Leur transport dans la plante est très limité, ils ont une action de contact. Ils ont un effet inhibiteur sur la croissance des méristèmes et sont de ce fait généralement utilisés en prélevée ou en post-levéeprécoce contre les graminées. Ils inhibent également la respiration. Leur solubilité dans l'eau est faible et ils persistent dans les sols de 2 à 4 mois. Leur toxicité vis-à-vis des mammifères est faible. Leur nom se termine généralement par le vocable "fène"

Exemples : acifluorfène-sodium, aclonifen, bifénox, bromofénoxime, chlométhoxyfène, diclofop-méthyle, fluorodifène, fomesafen, lactofène,nitrofène, oxyfluorfèneé

- **Les carbamates :** conçus en 1945 pour la destruction des graminées, cesherbicides se subdivisent en 4 catégories :

1. les dérivés de l'acide carbamique (NH_2-COOH) qui agissent sur la division cellulaire.

2. les dérivés de l'acide thiocarbamique (NH_2-CO-SH) qui inhibent la synthèse des lipides à longue chaîne et des gibbérellines.

3. les dérivés de l'acide dithiocarbamique (NH_2-CS-SH) qui empêchent la germination.

4. les biscarbamates qui empêchent la photosynthèse.

Ces herbicides ont en commun leur faible toxicité et une volatilité plus ou moins grande. Ils perturbent la division cellulaire (antimitotique) et la physiologie générale de la plante, provoquant le phénomène de l'anse en panier.

Elles s'emploient le plus souvent en pré-levée (thiocarbamates) ou post-semis, parfois en post-levée (phenmediphame, barbame).

À l'exception des composés allates, qui persistent plusieurs mois dans le sol, leur persistance est quasiment nulle.

Exemples :

1. Asulame, barbame, chlorbufame, chlorprophame, prophame, carbétamide ;

2. Thiocarbamates: butilate, cycloate, diallate, triallate, EPTC, molinate, prosulfocarbe, vernolate, pédulate, thiobencarbe ;

3. Dithiocarbamates : métam sodium, nabame ;

4. Biscarbamates: desmédiphame, phenmédiphame, karbutylate.

Les pesticides peuvent être aussi classés en fonction de leurs familles chimiques. Les familles les plus importantes sont les organophosphorés, les organochlorés, les carbamates, les urées et les triazines.

* **Les carbamates** ou **uréthanes** sont une famille de composés organiques porteur d'une fonction R-HN-(C=O)O-R'. Il s'agit en fait des **esters** substitués de l'**acide carbamique** ou d'un **amide** substitué.

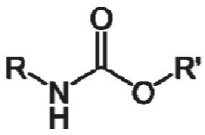


Figure 1: Formule générale des carbamates

Le groupe fonctionnel carbamate peut être formé lorsqu'une molécule de **dioxyde de carbone** ou un dérivé carbonate réagit avec la terminaison amino d'une chaîne de **peptides** ou le groupe aminé d'un **acide aminé**, y ajoutant un groupe COO^- et libérant un cation (H^+).

*** Les triazines** : Groupe d'herbicides (atrazine, simazine et terbuthylazine) très utilisés en grande culture (maïs principalement) et dont la solubilité relativement importante et la dégradation plus lente que celle des autres pesticides de ce type soulèvent divers problèmes de toxicité de l'environnement.

*** Lesurées** :L'**urée** ou **carbamide** (DCI) est un composé organique de formule chimique $CO(NH_2)_2$. C'est aussi le nom de la famille des dérivés de l'urée de formule générale $(R_1,R_2)N\text{-}CO\text{-}N(R_3,R_4)$.

L'urée naturelle est découverte en 1773 par Hilaire Rouelle. Formée dans le foie lors du cycle de l'urée, à partir de l'ammoniacqui provient de la dégradation terminale de trois acides aminés : l'arginine, la citrulline et l'ornithine, l'urée naturelle est éliminée par l'urine.

La plus importante utilisation actuelle se fait sous la forme d'engrais azotés.

L'urée est hydrolysée en ammoniac et en dioxyde de carbone dans le sol.

L'urée, qui contient 46 % d'azote, ne pourrait être utilisée comme engrais en raison de son caractère hygroscopique élevé. La présentation en granulés ou perles de calibre homogène est nécessaire pour la régularité de l'épandage.

Sur le plan agronomique, c'est une formulation intéressante car sa minéralisation est progressive. L'urée a tendance à acidifier les sols. On l'utilise généralement en couverture sur des cultures d'été.

D'un point de vue environnemental, son bilan carbone (si on tient compte des émissions en champs) est moins favorable que ceux d'autres engrais azotés (ammonitrate par exemple).

L'urée est un engrais azoté. C'est même le plus riche en azote (à part l'ammoniac) puisqu'il dose près de 50 % d'azote. Elle est utilisée principalement sous forme de solution (engrais liquide) épandue au printemps. Elle se présente aussi sous la forme de perles de 2 mm de diamètre.

*** Les organophosphorés** : Un **composé organophosphoré** est un type de composé organique comportant au moins un atome de phosphore lié directement à un carbone. Les composés d'intérêt biologique tels l'ADN, d'une importance capitale notamment en biochimie ne sont pas à proprement parler des composés organophosphorés : ils ne contiennent aucune liaison carbone -phosphore, et sont exclusivement des mono-, di- et triphosphates.

Fruits d'une recherche sur les gaz de combat entamée lors de la Seconde Guerre mondiale, les pesticides organophosphorés, comme le malathion, le Roundup se sont substitués, dans les années 1970, aux organochlorés, dont le chef de file, le DDT, faisait l'objet d'interdictions. Moins toxiques que le DDT et très efficaces, ils sont employés dans le monde entier.

Les composés organophosphorés se répartissent en différentes classes selon le degré d'oxydation du phosphore et la nature des substituants, notamment la présence d'un atome d'oxygène ou d'un autre chalcogène.

Les composés organophosphorés et organochlorés, les carbamates, et les triazines sont des composés constituant une famille de pesticides agissant sur l'enzyme acétylcholinestérase (la famille des carbamates agit également sur cette enzyme mais selon un mécanisme différent). Ils opèrent en bloquant irréversiblement l'acétylcholinestérase, essentielle aux transferts nerveux chez les insectes, les humains, ainsi que chez la plupart des animaux. La capacité à bloquer l'acétylcholinestérase (et donc la toxicité) peut varier de façon importante d'un composé à l'autre. Par exemple, le parathion, un des premiers organophosphorés, est beaucoup plus puissant que le malathion, un insecticide utilisé pour combattre la mouche du fruit méditerranéenne et les moustiques dans la vallée du Nil.

Les composés organophosphorés sont rapidement dégradés par le rayonnement solaire, dans l'air, et dans les sols, bien que de petites quantités puissent subsister et se retrouver dans la nourriture et l'eau. Le fait qu'ils se dégradent facilement fait de cette famille une alternative intéressante aux pesticides organochlorés persistants. Cependant, bien que les

organophosphorés se dégradent plus rapidement, ils sont plus toxiques, ce qui représente un risque pour les utilisateurs de ces composés.

Les composés organophosphorés les plus courants sont le parathion, le malathion, le methylparathion, le chlorpyrifos, le diazinon, le dichlorvos et le phosmet. Cette famille comprend un grand nombre de composés chimiques contenant du chlore et quelquefois d'autres éléments.

Les insecticides les plus puissants et les plus efficaces sont des organochlorés. On trouve dans cette famille le DDT, le chlordane, ou en encore le pentachlorophénol. Ils sont très persistants dans les sols, et ils se concentrent également dans les tissus biologiques. Beaucoup de composés de cette famille sont interdits en raison de leur neurotoxicité. Les carbamates présentent les mêmes caractéristiques que les organophosphorés, mais avec une toxicité moins importante.

Cette famille couvre un grand champ d'utilisation. La plupart sont utilisés comme herbicides sélectifs. Comme herbicides, les triazines peuvent être utilisés seules ou combinées avec d'autres composés afin d'augmenter leur efficacité. Le caractère sélectif des triazines vient du fait que certaines plantes peuvent métaboliser ces composés tandis que d'autres ne le peuvent pas. Les triazines comptent parmi les plus anciens herbicides, les recherches sur ces composés ayant commencé dans les années 30 du siècle précédent.

II. Le marché des pesticides :

Les pesticides font partie des substances susceptibles d'occasionner des risques à la fois pour la santé humaine et l'environnement (Observatoire Régional de la Santé de Bretagne, 2001).

Le marché mondial (environ 30 milliards d'euros) est globalement stable depuis quelques années (2000). Certains évènements climatiques récents (chaleur, sécheresse en Europe, puis en Océanie) influencent fortement ces chiffres. En Europe et en Amérique du Nord, les herbicides représentent 70 à 80% des produits utilisés pour deux raisons :

- La première, structurelle, pour maintenir des sols nus durant la période froide inculte en prévision des semailles à venir.
- La seconde, conjoncturelle, à cause de la forte augmentation des cultures de maïs (O.R.P., U.I.P.P.) (Fleischer G *et al.,* 1998).

Les usages non-agricoles représentent environs 12% du marché global (dont plus d'un tiers des usages pour les Etats-Unis).

En association avec la généralisation de l'usage d'engrais, les pesticides ont permis, depuis quarante ans, de tripler la productivité. Les pertes occasionnées aux cultures représentent pourtant encore près de 30% en Europe pour le maïs contre 50% en Afrique ; moins de 30% en Asie pour le riz contre plus de 50% en Afrique. Dans un contexte de sous-nutrition pour un milliard d'individus sur Terre et de projections démographiques moyennes visant à un accroissement de la population de 50% d'ici 2050, les pesticides restent comme l'une des opportunités agronomiques majeures. Le marché des pesticides est donc considéré comme disposant d'une marge de progression encore significative. Les entreprises de production de produits agropharmaceutiquesdominant le marché sont américaines et européennes.

Les pesticides les plus utilisés (en termes de quantité) sont les herbicides. Le composé le plus utilisé au monde est le glyphosate.

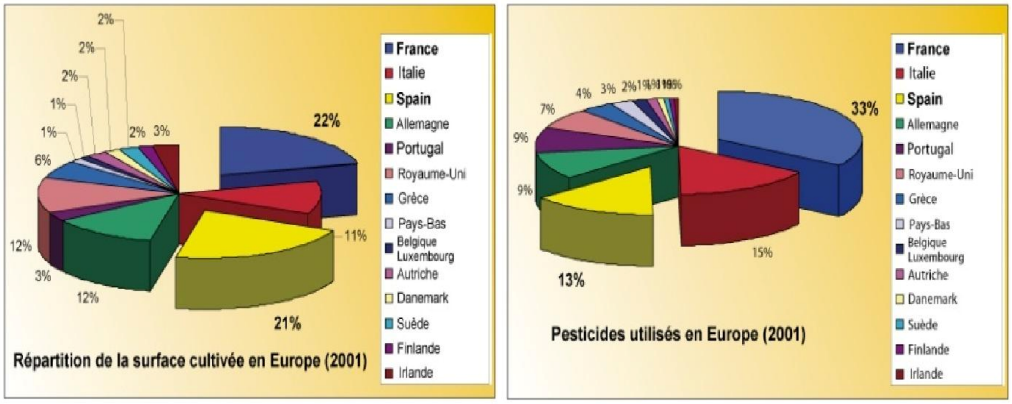


Figure 2: Répartition de la surface cultivée et pesticides utilisés en Europe

D'après : ORP (Observatoire des Résidus de Pesticides).

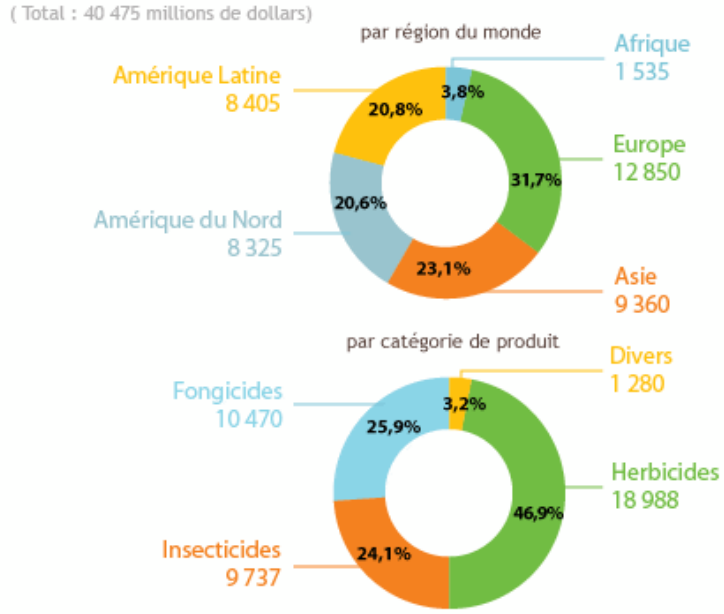


Figure 3: Le marché mondial des pesticides

(Source UIPP union des industries de la protection des plantes)

III. Les pesticides en Algérie :

Les pesticides seraient responsables du décès de 20.000 personnes environ chaque année, dans le monde. Actuellement, 25 groupes de pesticides, dont la plupart sont utilisés en Algérie, ont été déclarés substances cancérigènes. Dans notre pays, l'usage des insecticides, des fertilisants, des engrais, des détergents et autres produits phytosanitaires, se répand de plus en plus (Bounechada M, 2001).

L'Algérie est classée parmi les pays qui utilisent les plus grandes quantités de pesticides.Récemment dans notre pays, l'usage des pesticides ne cesse de se multiplier dans de nombreux domaines et en grandes quantités. Ainsi environ 400 produits phytosanitaires sont homologués en Algérie dont une quarantaine de variétés sont largement utilisées par les agriculteurs (ScheyerA, 2000). En 2009 l'Algérie a importé pour 67 millions USD de pesticides et en 2008, pour 77 millions USD contre 49,4 millions USD en 2007 (Bordjiba O, & Ketif A, 2009).

Les analyses des résidus de pesticides pour évaluer le degré de contamination des milieux naturels (les cultures, les eaux superficielles) ne sont pas faites systématiquement. Aucune analyse n'a été faite jusqu'à présent sur les fruits et légumes. Toutefois, quelques analyses effectuées sur des échantillons d'eau prélevés dans la région de Staouali (Alger) et

d'Annaba ont montré que dans plus de 30% des échantillons, la concentration de certaines molécules organochlorées : lindane ($C_6H_6Cl_6$), 2,4 et 4,4 DDT ($C_{14}H_9Cl_5$), 2,4 et 4,4 DDE « 2,4 et 4,4 Dichlorodiphényldichloroéthylène »($C_{14}H_8Cl_4$)et des organophosphorées : diazinon ($C_{12}H_{21}O_3PS$), dépasse les valeurs préconisées par l'O.M.S (Bordjiba O, & Ketif A, 2009).

En plus, selon le Cadastre national des déchets dangereux, il existe plus de 2.300 tonnes de pesticides périmés répartis sur 500 sites détenus majoritairement par les anciennes Entreprises nationales et usines de produits phytosanitaires (Onapsa, Asmidal...)(Bouziani M, 2007).

Uneenquêtea été réaliséeauprèsdesfellahsdelaChambred'Agricultured'Oranet de l'Institut de Protection des Végétaux de la Wilaya d'Oran montre que les pyréthrinoïdes, les organophosphorés et les carbamates sont les pesticides les plus utilisés en Algérie. Selon l'Institut National de Protection des Végétaux, la plusgrande quantité d'insecticides est utilisée dans la lutte antiacridienne.

Les services des urgences médicales d'Oran enregistrent plus de 200 cas d'intoxications par an par ingestion d'insecticides en milieu domestique.

Les empoisonnements par insecticides ou pesticides occupent la deuxième place en Algérie après les intoxications médicamenteuses (Benhamouche Z, 2010).

IV. Les pesticides et l'environnement :

Les pesticides ont été depuis près d'une cinquantaine d'années mis en évidence dans tous les compartiments environnementaux. Aussi bien, dans les eaux de rivières, les nappes phréatiques, l'air, les eaux de pluie, mais aussi dans les fruits, les légumes, les céréales et les produits d'origine animale(Fleischer G, 1998).

Les premiers effets des pesticides sur l'environnement ont été portés à la connaissance du grand public par Rachel Carson (1962) qui désignait par « dirtydozen » les treize matières actives les plus impliquées dans des mortalités d'oiseaux pêcheurs. Le retentissement de son ouvrage « Silentspring », changea le point de vue de la population sur les pesticides. Autrefois encline à voir en eux des molécules providentielles, celle-ci fut informée de leur nocivité potentielle.

Des cas ponctuels de mortalité de poissons, d'animaux domestiques et de bétail par simple boisson d'eaux contaminées ont été observés.Les effets des pesticides sur la santé dépendent toujours du type de pesticide. Certains pesticides comme l'organophosphate et les carbamates affectent le système nerveux. D'autres irritent la peau et les yeux. Quelques uns peuvent être cancérigènes et d'autres affectent le système hormonal dans le corps humain (FEPS les pesticides et la pollution de l'eau).

La toxicité des pesticides, et notamment des herbicides, est un sujet de préoccupation majeur.Cependant, les pesticides ne sont pas les seuls produits chimiques toxiques avec lesquels nousentrons en contact quotidiennement. Toute substance chimique, qu'elle soit naturelle ou synthétique,est toxique à un niveau d'exposition donné. La différence entre les toxicités aiguë et chroniquepar rapport à la dose sans effet nocif (NOEL – no observed effect level) est avant toutfonction de l'exposition au cours d'un temps donné et du mode d'action de la substance.

C'est grâce à leur mode d'action que la plupart des herbicides ont des effets utiles et efficacescar c'est lui qui permet de supprimer les plantes sensibles à des doses d'application relativementfaibles, alors que les animaux ne sont éliminés qu'à des doses beaucoup plus élevées. C'estégalement le mode d'action qui permet la sélectivité entre les espèces végétales. C'est principalementen raison de leur mode d'action que la plupart des herbicides sont relativement nontoxiques pour les mammifères, les oiseaux, les poissons, les insectes et les espèces aquatiquesnon végétales.

Les végétaux possèdent un certain nombre de systèmes métaboliques qui n'existent pas chezl'animal et qui constituent le plus souvent la cible du mode d'action de l'herbicide. Beaucoupd'herbicides forestiers font partie de la catégorie des régulateurs de croissance végétale (PGR : plant growth regulator ou PGRs) et notamment les produits suivants : asulame ; 2,4-D ; dalapon; dichlobénil ; fosamine ammonium ; propyzamide ; triclopyr. Ces régulateurs peuvent agir surplusieurs sites à la fois dans une plante et en perturber ainsi l'équilibre hormonal, en particulierpar mimétisme ou par inhibition d'une hormone de croissance végétale.

La toxicité pour les mammifères des herbicides à usage forestier est très faible. La plupart desherbicides utilisés en forêt sont bien moins toxiques que la vitamine D. Pour la plupart de cesherbicides, les doses sans effets nocifs se situent entre la dose létale 50 de la vitamine D (10 mg/kg/jour) et celle de la nicotine (moins de 5 mg/kg/jour). Les risques de cancer associés aux herbicides àusage forestier constituent l'un des principaux soucis de l'opinion publique. Théoriquement, n'importequel produit chimique pouvant être à l'origine du développement d'une tumeur est classécomme cancérigène, qu'il s'agisse de tumeur maligne ou bénigne.

Les données toxicologiques qui doiventêtre présentées au titre de l'évaluation des risques pour l'environnement portent sur des espèces représentatives des populations à risque. Outre les mammifères, les espècessuivantes sont également concernées : Abeille mellifère, Daphnie, Canard colvert, Colin deVirginie, Truite arc-en-ciel, Crapet arlequin, plusieurs espèces de Saumon dont le Saumon coho,Tête de boule, Huître, Crabe à signaux, Bouquet Mississipi, Crevette rouge, Algues vertes y comprisles *Chlorellasp* et *Selenastrumsp*., Algues bleues (et notamment *Anabaenasp*.), Diatomées ycompris *Navicullasp* (D'eau douce) et *Skeletonemasp* (D'eau de mer), et au moins un macrophyteaquatique sensible (généralement une *Lemnasp*). Les fabricants de produits chimiques quiprésentent des demandes d'homologation de pesticide ont une certaine marge quant au choix desespèces sur lesquelles ils effectuent leurs essais toxicologiques. De ce fait, les données disponiblespour les herbicides ne portent pas sur toutes les espèces et certaines espèces supplémentairesou de substitution sont parfois soumises aux essais à condition que la société en questionsoit en mesure de justifier de telles exceptions (Michael J L, 2002).

Pour la majorité des cas rapportésdans la littérature, les niveaux de contamination sont conformes aux normes américaines pourl'eau potable, autrement dit, même après pulvérisation directe des cours d'eau, les valeurs pourl'eau consommée sur place n'atteindront pas les limites toxiques. Michael *et al.,* ont analysé (Michael J L *et al.,* 1999) la dissipation de l'hexazinone dans les écosystèmes forestiers après application d'unedose trois fois plus élevée que la normale. Ils ont observé une contamination des cours d'eaupar l'hexazinone le jour même du traitement supérieure à la norme applicable à l'eau potable ; celle-ci a duré moins de 30 minutes et a rapidement diminué pour atteindre des niveaux prochesde la limite de détection le restant de la journée. Les concentrations maximales observées dansles cours d'eau à la suite du traitement ont duré entre 15 et 30 minutes.

IV. 1- Les voies d'exposition de la population aux pesticides :

Alors que les sources d'exposition professionnelle aux pesticides découlent directement de l'emploi qui en est fait (production, traitement des cultures ou des animaux, programmes de santé, etc.), la population générale est essentiellement exposée au travers de son alimentation et de son environnement. L'exposition par l'alimentation concerne certains aliments traités et l'eau dans une moindre mesure compte tenu des exigences de qualité de la réglementation. La contamination de l'environnement expose tout un chacun à des niveaux de pesticides variables et souvent difficiles à apprécier.

L'exposition aux pesticides se caractérise par une multiplicité des voies d'exposition, en effet ces substances peuvent pénétrer dans l'organisme par contact cutané, par ingestion et par inhalation. On distingue généralement deux types d'exposition :

IV. 1- 1- Les expositions primaires :

Elles concernent les personnes manipulant les produits, au moment de la préparation, de l'application mais aussi du nettoyage des appareils de traitement. Les populations concernées sont bien évidemment les agriculteurs et les professionnels mais tout un chacun est également exposé lors de l'utilisation de produits à usages domestiques ou d'entretien des jardins.

Les pesticides utilisés dans les champs ou à domicile sont trop souvent entreposés sans précaution particulière dans les habitations et les membres de la famille peuvent y avoir facilement accès. Ces substances toxiques peuvent, dans ces conditions, contaminer l'eau ou les aliments et polluer l'air ambiant. Plus grave encore, ils peuvent conduire à des expositions accidentelles des plus jeunes enfants. Ces accidents domestiques sont encore trop nombreux ! Les pesticides doivent être stockés sous clef, dans un endroit frais, sec et bien ventilé de préférence à l'extérieur des habitations.

En préalable à la mise sur le marché d'un produit phytopharmaceutique, la réglementation impose une évaluation des risques pour la santé des applicateurs à l'aide d'outils de modélisation, conduisant à des recommandations quant aux conditions d'utilisation du produit et aux mesures de protection individuelle à mettre en œuvre.

IV. 1- 2- Les expositions secondaires :

Elles concernent l'ensemble de la population, qui est exposée aux résidus de l'usage de ces produits, au travers de son alimentation et de son environnement.

Aujourd'hui, les mesures de contamination des sols et de l'air sont encore trop récentes et disparates pour permettre de renseigner correctement ces voies.

Dans ces conditions, il n'est pas possible de proposer une hiérarchisation des voies d'exposition aux pesticides.

Quelques populations ont été identifiées comme particulièrement à risque. Il s'agit de la femme enceinte exposée aux pesticides, l'enfant qu'elle porte est, lui aussi, exposé avant même sa naissance. Le bébé peut également être en contact avec des pesticides persistants et bio-accumulables par le lait maternel. D'où la nécessité de protéger la femme enceinte et la mère allaitante contre une exposition à ces contaminants.

De même, le jeune enfant est toujours attiré par son environnement immédiat. Il joue volontiers par terre et a tendance à mettre des choses dans sa bouche. Il risque donc d'absorber des doses non négligeables de pesticides provenant du sol, de la poussière ou de divers objets contaminés qu'on trouve en milieu rural, mais aussi urbain, à la maison ou au jardin.

V. Devenir des pesticides dans l'environnement

Les interactions entre les propriétés chimiques des herbicides et les conditions propres aux milieux concernés déterminent la persistance, la mobilité, ainsi que le potentiel de bioaccumulation. Chacun de ces aspects du destin écologiquedéfinit la probabilité pour un organisme d'être exposé. Les principales caractéristiques concernéessont les coefficients K_d et K_{ow}, l'hydrosolubilité, la dégradation hydrolytique et photolytique et les réactions d'oxydoréduction.

K_d, le coefficient de partage sol-eau, constitue la mesure du potentiel d'adsorption sur les particulesde sol d'un herbicide en solution aqueuse. Ce coefficient est souvent, mais pas systématiquement, lié à la quantité de carbone organique dans le sol (Wauchope R D *et al.,* 1992). La mesure du K_d peut se faire directement pour chaque sol mais de nombreux efforts ont néanmoins été faits pour se passer de la mesure directe. Dans la plupart des cas, c'est le K_{oc} qui est affiché pour les pesticides. Ce coefficient K_{oc} est calculé à partir de K_d en prenant

comme hypothèse que l'adsorption de pesticides sur les particules de sol dépend uniquement de la matière organique. On lui accorde ainsi une valeur "prévisible" basée exclusivement sur la solubilité. Cette valeur est utile à condition que sa précision soit admissible dans un ordre de grandeur donné. Cette hypothèse n'est toutefois pas valable pour les herbicides polaires, ionisants et hautement hydrosolubles, pour cela les chercheurs préfèrent le K_d qui est une estimation plus précise de l'affinité de chaque herbicide vis-à-vis du sol. Étant donné le très grand nombre de types de sols que l'on trouve dans le monde, on utilise volontiers des fourchettes de valeurs pour les K_d (tableau I). Pour certains de ces coefficients, les données sont limitatives et, dans ce cas, une valeur unique est affichée.

Tableau I : Propriétés chimiques de certains herbicides

Herbicide	K_d	K_{ow}	Solubilité aqueuse (mg/l) **	photolyse	hydrolyse	Demi-vie moyenne dans le sol (jours)
Asulame*	0,24 -1	0,31-1,8	5000	Oui	Oui	
2,4-D	0,14- 3,38	2,81	620	Oui	Oui	
Dalapon*	1	6	900000	Oui	Oui	
Dichlobénil*	0,295- 2 ,098	3,06	18	Oui	Non	-
Dichlorprop*	2	1000	350	Non	Non	-
Fluazifop-p-butyle*	67	31622	2	Non	Oui	-
Fosamine ammonium	0,095	0,00125	1790000	Non	Non	-
Glyphosate	62-175	0,0017	12000	Non	Non	29
Hexazinone	0,24-10,8	14,79	33000	Oui	Non	88
Imazapyr	0,06-3,02	1,3	15000	Oui	Non	46
Metsulfuron	1,4	0,01-1,0	9500	Oui	Oui	42
Oxyfluorfene*	1,160	29400	0,1	Oui	Non	-
Propyzamide*	3,2-10,1	1568	15	Oui	Non	-
Quizalofop éthyle*	6,2	15849	0,31	-	-	-
Sulfometuron	0,71	0,31	244	Oui	Oui	26
Triclopyr	0,165-0,975	<5(TEA)	435	Oui	Non	99

* Non homologué pour usage général forestier aux États-Unis,

** Hydrosolubilité à 20 ou 25°C et pH de 7 lorsque ce dernier est pertinent,

Sources: Grover (1977) ; Hay (1990) ; Kidd et James (1991); Pesticide Information Profiles (PIPs, http://ace,orst,edu/info/extoxnet/pips/searchindex,html) – fruit d'une collaboration entre l'Université de Californie Davis, la Oregon State University, la Michigan State University et CornellUniversity ; SERA (1999); les documents en ligne de la base de données USDA-ARS Pesticide PropertiesDatabase (http://wizard,arsusda,gov/acsl/ppdb2,html); et les documents RED (décision d'aptitude au renouvellement d'une homologation) de la US Environmental Protection Agency.

Des valeurs élevées de K_d indiquent une forte tendance à l'adsorption sur les particules de sol.Les herbicides à adsorption élevée sont moins disponibles et moins susceptibles d'être transféréshors de la station par le ruissellement en cas d'orage. Ces produits sont également généralement moinssujets aux déperditions par dégradation et par volatilité comparés à ceux qui ont une faiblevaleur K_d. Les valeurs élevées de K_d pour le fluazifop-p-butyle, le glyphosate et l'oxyfluorfene (tableau I) indiquent une forte adhérence de ces produits aux particules de sol et peu de mobilitéune fois au contact du sol.

K_{ow}, le coefficient de partage octanol-eau constitue lui aussi une mesure de l'hydrophobie desherbicides. Des valeurs élevées de K_{ow} indiquent l'existence d'un potentiel de stockage dans lestissus adipeux et par conséquent d'un plus grand potentiel de bioaccumulation. La bioaccumulationest une augmentation de la concentration d'une substance chimique dans un organisme,due à une absorption excessive par les tissus par rapport au taux métabolique et à la vitessed'excrétion. En règle générale, les pesticides persistants ayant des valeurs de K_{ow} supérieures à1 000 sont potentiellement biocumulables. L'oxyfluorfene a un potentiel faible à modéré debioaccumulation chez le Crapet arlequin (facteur de bioconcentration de 1 300) et le Poisson-chattacheté (facteur de bioconcentration jusqu'à 5 000) après une période d'exposition de 30 à40 jours dans des solutions contenant 10 ppb d'oxyfluorfene (EXTOXNET, 1996). L'hexazinone, dont la valeur K_{ow} est de 14,79, n'est pas biocumulable. Le fluazifop, dont le K_{ow} est très élevémais qui est peu soluble dans l'eau, rapidement hydrolysable et dont la demi-vie est d'environune semaine, ne risque pas de poser de problème de bioaccumulation.

La solubilité peut agir sur la persistance et la mobilité des herbicides dans les stations traitéesmais ne constitue généralement pas le facteur limitant lorsqu'elle est supérieure à 1 mg/l. Pourcertains herbicides, la solubilité et le K_{ow} sont fonction du pH. Lorsque le pH augmente, la solubilitédu sulfometuron et du metsulfuron, deux sulfonylurées, augmente parfois sensiblement, alors que le K_{ow} diminue. Par exemple, à un pH de 5 et une température de 25° C, l'hydrosolubilitédu sulfometuron est de 6,4 mg/l, alors qu'à un pH de 7, cette solubilité atteint 244 mg/l. Par conséquent, avec les pH couramment trouvés dans les forêts, la solubilité des herbicides qui figure au tableau I est telle qu'on peut estimer que tous, à l'exception de l'oxyfluorfene et du quizalofop éthyle, sont mobiles dans les sols forestiers. Au moyen du modèle GLEAMS, Micheal*et al.* (Micheal *et al.,*1996) avaient postulé que les

profils de mobilité hors station de l'imazapyr, de l'hexazinoneet du triclopyr seraient trèssemblables pour des applications sur sols identiques. Cependant, les coefficients de partage du sol indiquent que la mobilité du fluazifop-p-butyle, du glyphosate et, sur certains sols, de l'hexazinone et du propyzamide, est limitée par l'adsorption sur les particules de sol.

De plus, la persistance et la mobilité hors station des herbicides sont souvent limitées par l'hydrolyseet la photolyse (tableau I). Une hydrolyse rapide limite la persistance du butoxy éthyleester et des formulations triéthyle amine du triclopyr, les réduisant rapidement à de l'acidetriclopyr dans un intervalle allant de quelques heures à moins de deux jours. L'acide triclopyrsubit ensuite une dégradation supplémentaire par photolyse alors que l'hydrolyse n'agit plus surla forme acide. Il faut également tenir compte de la photolyse et l'hydrolyse lorsqu'on prélèvedes échantillons d'eau pour contrôler les résidus. Dans la majorité des cas, il est préférable degarder tous les échantillons au frais et à l'abri de la lumière. Dans le cas des sulfonylurées, letaux d'hydrolyse diminue lorsque le pH augmente et, par conséquent, il est nécessaire pour lessites où l'eau est acide de stabiliser la solution échantillon en ajustant le pH à un niveau neutreou supérieur.

Toutes ces propriétés chimiques jouent un rôle important dans le devenir dans l'environnementdes herbicides figurant au tableau I. Si on tient également compte de facteurs biotiquestels que l'activité microbienne, la persistance de ces produits est réduite à un niveau sensiblementplus bas que celui des composés chlorocarbonés plus anciens, objets de tant de méfianceaprès la publication du livre de Carson (Carson R, 1962). En réalité, la demi-vie dans le sol de ces composés estrelativement courte dans les conditions de terrain. Sont présentées au tableau I les données dedemi-vie disponibles à travers le monde pour les herbicides forestiers les plus largement utilisésaux États-Unis. Ces moyennes sont calculées à partir de données provenant de milieux et deconditions pédologiques très variés à travers le monde.

Lors d'un épandage, une grande partie du produit n'atteint pas l'organisme cible. Par exemple, sur une surface inclinée, le ruissellement des eaux de pluies peut entraîner le pesticide loin de sa cible. Il y a alors gaspillage du produit chimique, diminution de l'effet et risque de pollution du sol et des eaux.

De nombreux phénomènes influencent le devenir des pesticides dans l'environnement:

L'adsorption est un phénomène de surface par lequel les molécules se fixent aux particules du sol. La quantité de pesticide adsorbé varie selon le type de pesticide, la nature du sol, le pH du sol. Les pesticides s'adsorbent facilement sur des sols riches en argile ou en matière organique. Les pesticides adsorbés sont moins susceptibles de se vaporiser ou de migrer dans le sol. Ils sont aussi plus difficilement captés par les plantes.

La volatilisation est la transformation des solides ou des liquides en gaz. Ce processus peut disperser une grande partie du produit épandu dans l'atmosphère. Ce mouvement est appelé **dérive gazeuse**.

La dérive gazeuse de certains herbicides peut endommager des cultures voisines. Un temps chaud ou venteux accélère le phénomène de volatilisation. La dérive à l'épandage est le déplacement aérien des gouttelettes de produit phytosanitaire vaporisées lors du traitement des cultures. Ce phénomène peut entraîner les pesticides sur de grandes distances. La dérive à l'épandage dépend de : la taille des gouttelettes (plus celles-ci sont petites, plus elles seront entraînées), la vitesse du vent, la distance entre la buse de vaporisation et la plante. La dérive peut contaminer de grandes zones autour des sites traités.

Le ruissellement est l'entraînement des pesticides par l'eau sur des surfaces inclinées. Les pesticides étant soit mélangés à l'eau soit adsorbés sur les particules du sol érodé. L'infiltration est l'entraînement des pesticides par l'eau dans le sol.

L'infiltration peut se faire vers le bas, le haut ou horizontalement.

L'absorption est l'assimilation des pesticides par les plantes et les microorganismes. Une fois absorbé le composé peut être dégradé ou peut subsister dans l'organisme et être relâché dans l'environnement lorsque l'animal meurt ou que la plante se décompose.

De plus, lors des récoltes les résidus de pesticides peuvent également être déplacés.

I. Les modèles QSAR, QSPR:

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

Le principe des études QSAR (Quantative Structure- Activity Relationship), QSPR (Quantative Structure- Property Relationship) est d'établir une corrélation entre des données structurales de la molécule et leur activité biologique dans le cas d'étude QSAR ou leur propriété physico- chimique dans le cas d'une étude QSPR.

En fait, les premiers développements dans le sens de telles méthodologies sont plutôt anciens. Dès 1868, Crum-Brown et Fraser ont postulé l'existence de relations entre les activités physiologiques et les structures chimiques en reliant les changements d'activité biologique à des modifications structurales simples, ne disposant alors pas de moyen pour caractériser les structures chimiques en termes quantitatifs.

Hansch et Fujita établirent, en 1964, les premières corrélations entre les propriétés physico- chimiques (logP, pKa, paramètres stériques et électroniques) et l'activité biologique (activités enzymatiques, pharmacologiques) (Hansch C, & Fujita T, 1964). En 1971, ils réalisent une étude de relation structure- activité sur différentes familles d'antifongiques : benzoquinone, sels d'alkylpyridinium, imidazoles et phénols. Ils observent que quels que soient la famille et le champignon utilisé, l'activité antifongique dépend du logP (coefficient de partage octanol- eau) expérimental ou calculé (Hansch C, & Lien E J, 1971).

Ces études ont été extrapolées aux techniques séparatives en corrélant les propriétés physico- chimiques des analytes avec les temps de rétention obtenus expérimentalement (Tham S Y, & Agatonovic- Kustrin S, 2002).

Toutes ces études s'appuient sur le concept postulant que des structures similaires présentent des propriétés similaires.Ce type d'étude permet d'une part, d'expliquer les paramètres moléculaires impliqués dans l'activité biologique, une propriété physico-chimique et de prévoir d'autre part, l'influence de certaines modifications structurales dans l'activité biologique, une propriété physico- chimique d'un composé.

L'approche QSAR/QSPR procède de l'hypothèse d'une correspondance univoque entre n'importe quelle propriété physique, affinité chimique, ou activité biologique d'un composé chimique et sa structure moléculaire (Karelson M, 2000). Cette dernière peut être représentée par la composition chimique, la connectivité des atomes, la surface d'énergie potentielle, et la fonction d'onde électronique d'un composé.

II. Optimisation des molécules

II. 1. Généralités :

Les techniques de calcul qui peuvent fournir la valeur de l'énergie d'une géométrie, aussi particulière que l'état fondamental, appartiennent à plusieurs catégories :

- Méthodes *ab initio*,
- Méthodes semi- empiriques,
- Méthodes empiriques,
- Mécanique moléculaire.

Concernant les deux premières méthodes, elles sont fondées sur l'évaluation des interactions électroniques complètes ou partiellement négligées. Le terme *ab initio* est réservé aux calculs déduits directement des principes théoriques, sans faire intervenir de données expérimentales. Deux méthodes fondamentales sont proposées pour la résolution de l'équation de Schrödinger à partir des principes de base. La théorie des orbitales moléculaires (OM) tend à établir une expression pour la fonction d'onde ψ, alors que dans la théorie de la fonctionnelle de la densité (DFT), la distribution de la densité électronique (ρ) joue ce rôle. Le fondement de la DFT est associé à un théorème dû à Hohenberg et Kohn (Hohenberg P, & Kohn W, 1964) qui ont démontré que toutes les propriétés d'un système dans un état fondamental non dégénéré sont complètement déterminées par sa densité électronique.

Le type le plus courant de calcul *ab initio*, ou calcul Hartree Fock (HF), repose sur l'approximation principale du champ central. Le calcul variationnel mis en œuvre conduit à des énergies supérieures aux énergies réelles (Théorème de Eckart) et tendent vers une valeur limite appelée limite de Hartree Fock. La seconde approximation dans les calculs HF consiste à décrire la fonction d'onde par une « fonction utile » qui est connue exactement pour quelques systèmes mono- électroniques. Les fonctions les plus souvent utilisées sont des combinaisons linéaires d'orbitales de type Slater (e^{-ax}) ou d'orbitales gaussiennes ($e^{(-ax^2)}$),

dont les abréviations sont, respectivement, STO (pour Slater Type Orbitals) et GTO (pour Gaussian Type Orbitals). La fonction d'onde est obtenue à partir de combinaisons linéaires d'orbitales, ou plus souvent à partir combinaisons de fonctions d'un ensemble de base. A cause de cette approximation, la plupart des calculs HF conduisent à des énergies supérieures à la limite HF. L'ensemble exact de fonction de base utilisé est souvent spécifié par une abréviation du genre STO- 3G ou 6- 311++g**.

L'utilisation de bases de fonctions gaussiennes permet de calculer toutes les intégrales de la méthode sans autres approximations que celles inhérentes à la méthode elle- même.

Réservées initialement au traitement de petites molécules (une dizaine d'atomes), les méthodes *ab initio* ont été étendues, ces dernières décades, à des systèmes de quelques centaines d'atomes, comme conséquence de l'augmentation de la puissance des ordinateurs (hardware et software).

Une approximation sur l'hamiltonien est considérée comme une méthode semi-empirique.

Les méthodes semi- empiriques sont moins contraignantes en moyens de calculs. De plus, l'incorporation de paramètres déduits des données expérimentales dans certaines de ces méthodes permet de prédire quelques propriétés avec une meilleure précision que celle obtenue avec les méthodes *ab initio* les plus élaborées.

Les méthodes de champ de force ne demandent pas de temps excessifs de calcul pour donner des informations sur l'énergie de la molécule étudiée. La mécanique moléculaire (MM), appelée parfois : calcul par champ de force empirique, (empirical Force Field, EFF, en anglais), permet le calcul de la structure et de l'énergie d'entités moléculaires (Allinger N L, 1976, Niketic S R, & Rasmussen K, 1977, Burkert U, & Allinger N L, 1982). D'une part, les distributions électroniques ne sont pas explicitement détaillées (à quelque exception près), d'autre part, la recherche de l'énergie minimale par optimisation de la géométrie joue un rôle primordial.

L'énergie de la molécule est exprimée sous la forme d'une somme de contributions associées aux écarts de la structure par rapport à des paramètres structuraux de référence. Les variables de calcul sont alors les coordonnées internes du système : longueur de liaison, angles de valence, angles dièdres et distances entre les atomes non liés. Un calcul de MM aboutit à une disposition des noyaux telle que la somme de toutes les contributions énergétiques est minimisée ; ses résultats concernent surtout la géométrie et l'énergie de système (LomasJ S, 1986).

II. 2. Méthodes semi- empiriques utilisées

Les méthodes AM1 et PM3 utilisées étant des re- paramétrisations de la méthode MNDO, nous présenterons ces trois méthodes, en rappelant au préalable le cadre des équations (*ab initio*) HFR (Hartree – Fock - Roothaan) sur lequel elles sont basées et les approximations supplémentaires auxquelles il est fait recours.

- **Le cadre Hartree – Fock–Roothaan**

Les méthodes *ab initio* utilisent l'équation de Schrödinger électronique obtenue après séparation des mouvements électroniques et nucléaires (approximation de Born-Oppenheimer) (Kolos W, & Wolniewiez L, 1964, Suteliffe B T,).
Dans la méthode Hartree – Fock la fonction d'onde ψ d'un système à N électrons est représentée par un déterminant de Slater ψ_0 de spin orbitales φ unique. Les spin orbitales consistent en des produits d'orbitales moléculaires (OM) φ et de fonction de spin (α ou β),

$\phi_a = \varphi_a\alpha$, $\bar{\phi} = \varphi_a\beta$.

On représentera ψ_0 par :

$$\psi_0 = \left|\phi_1\bar{\phi}_1\ \phi_2\bar{\phi}_2\ \ldots..\ \phi_M\bar{\phi}_M\right\rangle \qquad (1)$$

Pour un système à couches complètes comportant N électrons (auquel cas $M=\frac{N}{2}$).

Chaque OM est développée sous forme d'une combinaison linéaire de fonctions de base, appelées conventionnellement orbitales atomiques (OM – CLOA), combinaison linéaire d'orbitales atomiques), quoiqu'elles ne soient pas généralement, solutions du problème HF atomique.

$$\phi_a = \sum_{\mu}^{m} c_{\mu a}\chi_{\mu} \qquad (2)$$

En tenant compte de (1), on obtient après multiplication à gauche par une fonction spécifique, intégration et application du principe variationnel, un système d'équations linéaires, ou équations de Roothaan – Hall (pour un système à couches complètes) (Roothan C C J, 1951, Hall G G, 1951). Signalons que la résolution des équations de Roothaan – Hall fournit un total de m (= nombre de fonctions de base) orbitales moléculaires (OM) dont n sont

occupées et (m – n) libres ou virtuelles. Celles- ci sont orthogonales à toutes les orbitales occupées, mais n'ont pas d'interprétation physique directe exceptée comme affinité électronique (via le théorème de Koopmans (Koopmans T A, 1933). Elles servent dans la description des états excités. L'équation (3) condense, sous forme matricielle, les équations de Roothaan – Hall.

$$\mathbf{F}\,\mathbf{C} = \mathbf{S}\,\mathbf{C}\,\boldsymbol{\varepsilon} \tag{3}$$

où :

- La matrice **F**de Fock est l'opérateur hamiltonnien effectif,
- **C** est la matrice des coefficients des OM, $c_{a\mu}$,
- **S** est la matrice de recouvrement,
- et ε une matrice diagonale comportant les énergies orbitalaires.

La matrice de Fock, **F**, comporte toutes les informations relatives au système quanto-mécanique, c'est – à - dire toutes les interactions prises en compte dans les calculs. Sa formulation *ab initio* est la suivante :

$$F_{\mu\nu} = H_{\mu\nu} + J_{\mu\nu} - \frac{1}{2}K_{\mu\nu}$$

$$F_{\mu\nu} = H_{\mu\nu} + \sum_{\rho}^{n}\sum_{\sigma}^{m} P_{\rho\sigma}\left[\langle\mu\nu/\rho\sigma\rangle - \frac{1}{2}\langle\mu\sigma/\rho\nu\rangle\right] \tag{4}$$

Avec :

$$H_{\mu\nu} = \int \chi_{\mu}^{*}(1)\,\hat{h}\,\chi_{\nu}(1)\,d\tau_1 \tag{5}$$

$$\langle\mu\nu/\rho\sigma\rangle = \int\int \chi_{\mu}^{*}(1)\,\chi_{\nu}(1)\frac{1}{r_{12}}\,\chi_{\rho}^{*}(2)\,\chi_{\nu}(2)\,d\tau_1 d\tau_2 \tag{6}$$

et

$$P_{\rho\sigma} = 2\sum_{a}^{m} C_{\rho a}^{*} C_{\nu a} \tag{7}$$

où μ, ν, ρ et σ désignent des orbitales atomiques, et $H_{\mu\nu}$ des intégrales mono-électroniques représentant les valeurs moyennes de l'opérateur associé à l'énergie cinétique et l'opérateur énergie potentielle d'interaction noyau – électron $\left(\hat{V}_{en}\right)$. Les $\langle\mu\nu/\rho\sigma\rangle$ sont des intégrales de répulsion bi- électroniques représentant $\hat{V}_{ee}$ (opérateur d'interaction entre les électrons eux – mêmes), et les $P_{\rho\nu}$ sont les éléments de la matrice densité **P**.

$J_{\mu\nu}$ et $K_{\mu\nu}$ sont les représentations matricielles des opérateurs coulombien $\hat{J}$ et d'échange $\hat{K}$ respectivement.

L'énergie électronique (E_{el}) peut être exprimée au moyen des valeurs propres ε_a :

$$E_{el} = 2\sum_{a}^{m}\varepsilon_a - \frac{1}{2}\sum_{\mu\nu}^{m} P_{\mu\nu}\left(J_{\mu\nu} - \frac{1}{2}K_{\mu\nu}\right) \qquad (8)$$

Comme la matrice de Fock dépend des coefficients des orbitales, les équations de Roothaan doivent être résolues de façon itérative en utilisant la procédure du champ auto- cohérent ou SCF (pour : Self Consistent Field) (Blinder S M, 1965).

Une étape importante de la procédure SCF est la conversion de l'équation générale aux valeurs propres (3) en une équation ordinaire par une transformation orthogonale (méthode d'orthogonalisation de Löwdin (Lowdin P O, 1950)).

$$\mathbf{F}^{\lambda}\,\mathbf{C}^{\lambda} = \mathbf{S}^{-1/2}\,\mathbf{F}$$

Avec :

$$\mathbf{F}^{\lambda} = \mathbf{S}^{-1/2}\,\mathbf{F}\,\mathbf{S}^{-1/2} \qquad (9)$$

et

$$\mathbf{C}^{\lambda} = \mathbf{S}^{1/2}\,\mathbf{C}$$

Notons que $\mathbf{S}^{-1/2}$ qui est obtenue à partir de la matrice de recouvrement **S** qui n'est jamais singulière, n'est jamais singulière non plus.

Signalons que les approximations électronique et CLOA (utilisation d'un nombre limité d'orbitales atomiques) et un problème de corrélation limitent la méthode HFR. On dépasse ces limitations par l'utilisation de fonctions corrélées ou en faisant intervenir l'interaction de configuration.

- **Les méthodes semi- empiriques**

Dans les méthodes semi- empiriques on simplifie l'approche Hartree – Fock – Roothaan.

1) Dans la construction de ψ_0 : seuls les électrons de valence sont traités de façon explicite en utilisant un ensemble de base minimal. Ce qui signifie que les atomes H sont décrits par une fonction 1s, les éléments Li à F par un ensemble {2s, 2p}, les éléments Na à Cl par un ensemble {3s, 3p}, Ca, K, et Zn à Br avec un ensemble {4s, 4p}, Se – Cu avec un ensemble de base {4s, 4p, 3d} ; etc…

On tient compte des électrons de cœur soit en corrigeant la charge nucléaire, soit en introduisant des fonctions pour modéliser les répulsions simultanées entre noyaux d'une part et entre électrons de cœur d'autre part.

2) dans la construction de $\mathbf{F}^{\lambda}$ on néglige une grande part des interactions, en particulier dans la partie bi- électronique $\langle \mu\nu/\rho\sigma \rangle$. Toutes les intégrales mettant en jeu des orbitales atomiques centrées sur plus de 2 noyaux sont négligées. Certaines classes d'intégrales sont remplacées par des paramètres. C'est le cas, en particulier, des intégrales mono- électroniques bi- centres $\mathbf{H}_{\mu\nu}$ qui sont, pour une large part, responsables de la liaison chimique.

La façon d'introduire ces simplifications dans le modèle permet de distinguer entre les différentes méthodes.

Une autre façon de réduire les intégrales bi- électroniques est l'approximation du recouvrement différentiel nul (RDN) dans laquelle on néglige tous les produits des fonctions de base dépendant des coordonnées d'un même électron localisé sur des atomes différents. Cela signifie que tous les produits des fonctions orbitales atomiques $\chi_\mu \chi_\nu$ sont posés égaux à zéro et l'intégrale de recouvrement se réduit à $S_{\mu\nu} = \delta_{\mu\nu}$ ($\delta_{\mu\nu}$ est le symbole de Kronecker ; $\delta_{\mu\nu} = 0$ si $\mu \neq \nu$ et $\delta_{\mu\nu} = 1$ si $\mu = \nu$).

Dans l'approximation RDN, toutes les intégrales tri et tétra- centres s'annulent ce qui transforme la matrice de recouvrement en une matrice unité. Les intégrales mono-électroniques tri- centres sont égales à zéro. Toutes les intégrales bi- électroniques tri et tétra-centres sont négligées.

Les paramètres sont imposés pour compenser les approximations. Ainsi toutes les intégrales restantes sont remplacées par des paramètres convenables ajustés sur des grandeurs fournies par l'expérience.

Toutes les méthodes semi- empiriques modernes sont basées sur l'approche MNDO (ModifiedNeglect of DifferentialOverlap) (Dewar M J S, Thiel W, 1977), dans laquelle des paramètres sont assignés aux différents types d'atomes puis ajustés de telle sorte à reproduire certaines propriétés comme les chaleurs de formation, les variables géométriques, les moments dipolaires et les énergies de première ionisation.

Les paramètres sont conçus séparément pour des classes de composés tels que les hydrocarbures, les systèmes CHO, les systèmes CHN, etc…

Les méthodes AM1 (Tremaine L M *et al.,* 1984) et PM3 (DewarM J S *et al.,* 1985) appartiennent aux dernières versions de la méthode MNDO.

Dans la méthode MNDO les paramètres associés aux intégrales bi- électroniques mono- centres sont basés sur des données spectroscopiques relatives aux atomes isolés et l'évaluation des autres intégrales bi- électroniques repose sur les interactions multipole – multipole de l'électrostatique classique. Dans cette méthode, des composés contenant H, Li, Be, B, C, N, O, F, Al, Si, Ge, Sn, Pb, P, S, Cl, Br, I, Zn, et Hg ont été paramétrés.

L'hamiltonien associé aux électrons de valence est donné par :

$$\hat{H}_{val} = \sum_{i=1}^{n(val)} \left[-\frac{1}{2}\nabla_i^2 + V(i) \right] + \sum_{i=1}^{n(val)} \sum_{j>i} \frac{1}{r_{ij}} \qquad (10)$$

Qui se simplifie en :

$$\hat{H}_{val} = \sum_{i=1}^{n(val)} \hat{H}_{val}^{c}(i) + \sum_{i=1}^{n(val)} \sum_{j>i} \frac{1}{r_{ij}} \qquad (11)$$

où :

$$\hat{H}^{c}_{val}(i) = \left[-\frac{1}{2}\nabla_i^2 + V(i) \right] \qquad (12)$$

n(val) désigne le nombre d'électrons de valence du système, V(i) est l'énergie potentielle de l'électron i dans le champ des noyaux et des électrons de cœur, $\hat{H}^{c}_{val}$ (i) est la contribution mono – électronique à $\hat{H}_{val}$.

Les élements de la matrice de Fock sont calculés à l'aide de l'équation :

$$F_{val.rs} = \hat{H}^{c}_{val.r\,s} + \sum_{t=1}^{b}\sum_{u=1}^{b} P_{tu} \left[\langle rs/tu \rangle - \frac{1}{2}\langle ru/ts \rangle \right] \qquad (13)$$

Dans la méthode MNDO les éléments de la matrice de Fock peuvent être calculés comme suit.

Les éléments de la matrice de cœur (intégrale de résonance de cœur) $H^{c}_{\mu_A\mu_B} = \left\langle \mu_A(1) \left| \hat{H}^{c}_{(1)} \right| \mu_A(1) \right\rangle$, avec des orbitales atomiques centrées sur les atomes A et B sont donnés par :

$$H^{c}_{\mu_A\mu_B} = \frac{1}{2}\left(\beta_{\mu_A} + \beta_{\nu_B}\right) S_{\mu_A\nu_B}; \quad A \neq B \qquad (14)$$

où les β sont les paramètres de chaque orbitale. Par exemple, le carbone avec les orbitales atomiques de valence 2s 2p, centrées sur le même atome de carbone, aura les paramètres β_{C2s} et β_{C2p}.

Les éléments de la matrice de cœur à partir d'orbitales atomiques différentes centrées sur le même atome sont fournis par l'équation (15) :

$$H^{c}(1) = -\frac{1}{2}\nabla_1^2 + V(1)$$

où V(1) est l'énergie potentielle de l'électron de valence 1 dans le champ du cœur. Décomposant V(1) en contributions individuelles de cœurs atomiques, il vient :

$$H^{c}(1) = -\frac{1}{2}\nabla_1^2 + V_A(1) + \sum_{B\neq A} V_B(1) \qquad (15)$$

Ainsi :

$$H^{c}_{\mu_A \nu_B} = \left\langle \mu_A \left| -\frac{1}{2}\nabla_1^2 + V_A \right| \nu_A \right\rangle + \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle \quad (16)$$

Des considérations de la théorie de groupes permettent d'annuler $\left\langle \mu_A \left| -\frac{1}{2}\nabla_1^2 + V_A \right| \nu_A \right\rangle$, de telle sorte que :

$$H^{c}_{\mu_A \nu_B} = \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle \quad (17)$$

Si l'on considère que l'électron 1 interagit avec un point du cœur de charge C_B, alors :

$$V_B = - \frac{C_B}{r_{1B}} \quad (18)$$

$$\left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle = - C_B \left\langle \mu_A \left| \frac{1}{r_{1B}} \right| \nu_A \right\rangle \quad (19)$$

Dans la méthode MNDO, $\left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle = - C_B \left\langle \mu_A \nu_A \left| s_B s_B \right. \right\rangle$, où s_B est l'orbitale de valence centrée sur l'atome B :

$$H^{c}_{\mu_A \nu_B} = \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle = - C_B \sum_{B \neq A} \left\langle \mu_A \nu_A \left| s_B s_B \right. \right\rangle ; \ \mu_A \neq \nu_A \quad (20)$$

Les éléments de la matrice de cœur : $H^{c}_{\mu_A \mu_A} = \left\langle \mu_A(1) \left| \hat{H}^{c} \right| \mu_A(1) \right\rangle$ sont calculés en utilisant la relation :

$$H^{c}_{\mu_A \mu_A} = \left\langle \mu_A \left| -\frac{1}{2}\nabla_1^2 + V_A \right| \nu_A \right\rangle + \sum_{B \neq A} \left\langle \mu_A \left| \nu_B \right| \nu_A \right\rangle \quad (21)$$

$U^{c}_{\mu_A \mu_A} = \left\langle \mu_A \left| -\frac{1}{2}\nabla_1^2 + V_A \right| \nu_A \right\rangle$ est évalué à partir de paramètres tirés de spectres atomiques (les paramètres utilisés pour l'atome de carbone : U_{ss} et U_{pp}). Donc :

$$H^{c}_{\mu_A \nu_A} = U_{\mu_A \mu_B} \sum_{B \neq A} C_B \left\langle \mu_A \nu_A \left| s_B s_B \right. \right\rangle \quad (22)$$

L'évaluation de $\left\langle \mu_A \nu_A \left| s_B s_B \right. \right\rangle$ est réalisée comme suit :

1) Toutes les intégrales tri et tétra- centres sont annulées dans la méthode RDN.
2) Les intégrales de répulsion électroniques mono- centres sont soit des intégrales coulombienne $g_{\mu\nu} = \langle \mu_A \mu_A | \nu_A \nu_A \rangle$, soit des intégrales d'échange $h_{\mu\nu} = \langle \mu_A \nu_A | \mu_A \nu_A \rangle$.

Pour l'atome de carbone, par exemple, les intégrales sont g_{ss}, g_{sp}, g_{pp}, $g_{pp'}$, h_{sp} et $h_{pp'}$, p et p' étant portées par des axes différents.

3) Les intégrales de répulsion bi- centres sont calculées à partir des valeurs d'une intégrale mono- centre et la distance inter- nucléaire en utilisant une procédure d'expansion multipole (Dewar M J S, Thiel W, 1977).
4) Le terme de répulsion cœur- cœur est donné par :

$$V_{CC} = \sum_{B>A} \sum_{A} \left[C_A C_B \left(s_A s_B / s_B s_B \right) + f_{AB} \right] \qquad (23)$$

où

$$f_{AB} = f_{AB}^{MNDO} = \left[C_A C_B \left(s_A s_B / s_B s_B \right) \left(e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}} \right) \right] \qquad (24)$$

α_A et α_B sont les paramètres des atomes A et B. Pour les paires O-H et N-H, par exemple, on aura :

$$f_{AH}^{MNDO} = \left[\left(C_A C_H \left(s_A s_H \right) s_H s_H \right) \left(R_{AH} e^{-\alpha_A R_{AH}} + e^{-\alpha_H R_{AH}} \right) \right] \alpha_A \alpha_H \qquad (25)$$

où A désigne soit N soit O.

Dans la méthode MNDO, les paramètres suivants doivent être optimisés :

1) Les intégrales mono- électroniques mono- centres U_{ss} et U_{pp}.
2) L'exposant ξ de la STO. Pour la MNDO $\xi_s = \xi_p$.
3) β_s et β_p. la méthode MNDO suppose que $\beta_s = \beta_p$.

Dans la méthode AM1, $\xi_s \neq \xi_p$

Des composés comportant différents atomes (H, B, Al, C, Si, Ge, Sn, N, P, O, S, F, Cl, Br, I, Zn, et Hg) ont été paramétrés dans AM1.

On a :

$$f_{AB}^{AM1} = f_{AB}^{MNDO} + \frac{C_A C_B}{R_{AB}}\left[\sum_k a_{kA}\exp\left[-b_{kA}\left(R_{AB}-C_{BA}\right)^2\right]\right] + \frac{C_A C_B}{R_{AB}}\left[\sum_k a_{kB}\exp\left[-b_{kB}\left(R_{AB}-C_{kB}\right)^2\right]\right] \quad (26)$$

Stewart a re- paramétré les valeurs pour générer la série PM. Celle qui dérive de AM1 est connue sous l'appellation PM3 (ParametricMethod 3).

Dans la méthode PM3, les intégrales de répulsion mono- centres sont paramétrées par optimisation. La fonction de répulsion de cœur contient seulement deux fonctions gaussiennes par atome. Des composés comportant des atomes parmi : H, C, Si, Ge, Sn, Pb, N, P, As, Sb, Bi, O, S, Se, Te, F, Cl, Br, I, Al, Ga, In, Be, Mg, Zn, Cd et Hg ont été paramétrés dans PM3.

II. – 3. Champ de force

II. 3- 1- Définition

La mécanique moléculaire est une méthode d'analyse conformationnelle basée sur l'utilisation de champs de forces empiriques et la minimisation d'énergie.

Dans un sens général, la mécanique moléculaire traite les atomes (ou les noyaux) d'une molécule comme des masses ou des sphères reliées par des ressorts de différentes forces représentant les liaisons.

Les interactions entre particules (de type atomique) sont traitées à l'aide de fonctions de potentiel tirées de la mécanique classique : fonctions de potentiel individuelles pour décrire les différents types d'interactions.

Les fonctions d'énergie potentielle comportent des paramètres empiriques décrivant des interactions entre des ensembles d'atomes. La parametrisation est faite à partir de données expérimentales (RMN, RX, calculs *ab initio*) sur le plus grand ensemble possible de molécules. Le choix des données expérimentales est important et le modèle obtenu en dépend étroitement. Les constantes sont ajustées pour rendre l'expression de l'énergie potentielle, E, la plus générale possible.

Les fonctions de potentiel et les paramètres exploités pour l'évaluation des interactions sont désignés par "champ de force".

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

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

II. 3. 2. Quelques exemples

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

-MM2, MM3, et MM4 : introduit par Allinger*et al*.(Alliger N L,1977, Burkert U, Alliger N L,, 1982- 1986, Alliger N L*et al.,*1989, Alliger N L *et al.,* 1996), largement utilisé pour le traitement de petites molécules.

-AMBER : (Assisted Method Building and Energy Refinement) introduit par Cornell *et al.,*(Mc kerell A D*et al.,* 1998), très largement utilisé dans le traitement des protéines et des acides nucléiques.

-CHARMM : (Chemistry at HarvadMacrmolecular Mechanics) développé par Mackerall, Karplus*et al.,* (Alliger N L *et al.,* 1989) qui est largement utilisé pour la simulation de petites molécules et jusqu'aux complexes solvatés de grandes macromolécules biologiques.

-MMFF : (MerckMolecular Force Field) développé par Halgren (HalgrenT A, 1996[1,2], Halgren T A, Nackbar R B 1996), il est similaire à MM3 dans la forme, mais en diffère par son application focalisée sur les processus de condensation de phases (en dynamique moléculaire). Il reproduit l'exactitude MM3 pour les petites molécules, et est applicable aux protéines et autres systèmes d'importance biologique.

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

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

II. 4- Représentation simple d'un champ de force

Beaucoup de champs de force utilisés actuellement pour la modélisation moléculaire peuvent s'interpréter en termes d'une représentation relativement simple à quatre

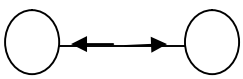
composantes des forces intra et intermoléculaires internes au système considéré. Les pénalités énergétiques sont associées aux écarts des liaisons et des angles par apport à leurs valeurs de 'référence' ou 'd'équilibre', il y a une fonction qui décrit la façon dont l'énergie change lors de la rotation des liaisons, et finalement le champ de force contient des termes décrivant l'interaction entre des parties non liées du système. Des champs de forces plus sophistiqués peuvent comporter des termes additionnels, mais ils présentent invariablement ces quatre composantes. Un fait intéressant lié à cette représentation est qu'on peut imputer les variations à des coordonnées internes spécifiques telles que les longueurs et les angles de liaisons, la rotation des liaisons ou les mouvements des atomes relativement les un aux autres. Ce qui permet de comprendre facilement comment les changements des paramètres du champ de force affectent ses performances, et aident également dans le processus de paramètrisation.

Pour des molécules seules ou des ensembles d'atomes et/ ou de molécules, un tel champ de force a la forme fonctionnelle suivante :

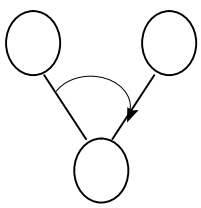
$$\mathbf{V}\left(r^N\right)_{liaisons} = \frac{k_i}{2}\left(l_i - l_{i,0}\right)^2 + \sum_{angles}\frac{k_i}{2}\left(\theta_i - \theta_{i,0}\right)^2 + \sum_{torsions}\left(1 - \cos\left(n\omega - \gamma\right)\right)$$
$$+\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\varepsilon_0\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right) \qquad (27)$$

$\mathbf{V}(r^N)$ représente l'énergie potentielle qui est fonction des positions (**r**) des N particules (habituellement les atomes).

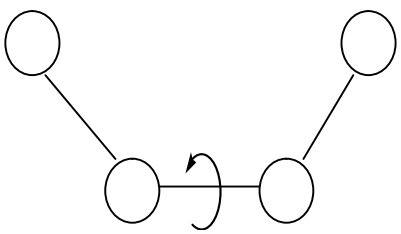
Les diverses contributions sont représentées schématiquement sur la figure suivante :



Elongation de liaison



Flexion angulaire



Rotation autour d'une liaison (torsion)

Interactions non liées (électrostatique)

Interactions non liées (van der Waals)

Figure 4: Représentation schématique des quatre contributions d'un champ de force de MM : élongation de liaison, flexion angulaire.

Le premier terme de l'équation (27) modélise l'interaction entre paires d'atomes liés, représentée dans ce cas par un potentiel harmonique qui fournit la variation de l'énergie lorsque la longueur de liaison l_i dévie de sa valeur de référence (à l'équilibre) $l_{i,0}$. Le second terme est une sommation sur tous les angles de valence de la molécule, encore modélisé par un potentiel harmonique (un angle de valence est l'angle formé par trois atomes A- B- C où A et C sont toux deux liés à B). Le troisième terme dans l'équation (27) renseigne sur la variation d'énergie lorsqu'une liaison tourne. La quatrième contribution est le terme de non liaison, qui est calculé entre toutes les paires d'atomes (i et j) localisés sur différentes molécules ou appartenant à la même molécule mais séparés par au moins 3 liaisons (c'est- à – dire avec une relation l, n où $n \geq 4$). Dans un champ de force simple le terme de non liaison comprend un terme de potentiel coulombien pour les interactions électrostatiques et le potentiel de Lennard- Jones pour les interactions de van der Waals.

- **Exemple de calcul :** énergie d'une conformation du propane

A titre d'illustration nous montrons comment la relation (27) peut être utilisée pour calculer l'énergie de conformation du propane (Fig. 5).

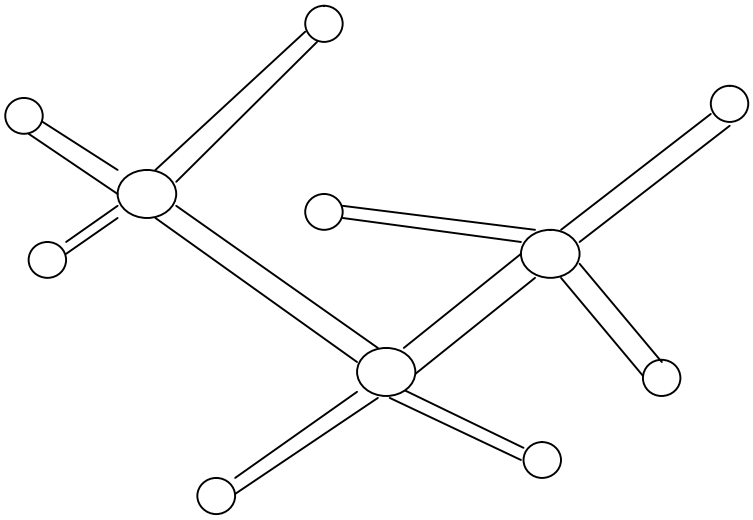


Figure 5: Un modèle de champ de force typique pour le propane contient 10 termes d'élongation de liaison, 18 termes de flexion angulaire, 18 termes de torsion et 27 interactions de non liaison.

Le propane possède 10 liaisons : 2 liaisons C-C et 8 liaisons C-H. Les liaisons C-C sont symétriques et équivalentes, mais les liaisons C-H appartiennent à 2 classes, un groupe comprend les 2H liés au carbone central du méthylène (CH_2) et un groupe correspondant aux 6 hydrogènes liés aux carbones des groupements méthyles.

Dans certains champs de force compliqués des paramètres différents seront utilisés pour ces deux types de liaison C-H, mais dans la plupart des champs de force les paramètres de liaison (k_i et $l_{i,0}$) seront utilisés pour les 8 liaisons C-H. Il y a 18 angles de valence différents pour le propane, comprenant un angle C-C-C, 10 angles C-C-H et 7 angles H-C-H. il est à noter que tous les angles sont pris en compte dans le modèle de champ de force quoique certains d'entre eux peuvent ne pas être indépendants des autres. Il y a 18 termes de torsion : 12 de type H-C-C-H et 6 du type H-C-C-C. chacun d'eux est modélisé par un développement en série de cosinus présentant des minima pour les conformations trans et gauche. Finalement, il y a 27 termes de non- liaison à calculer, impliquant 21 interactions H-H et 6 interactions H-C. la contribution électrostatique sera obtenue en appliquant la loi de Coulomb aux charges atomiques partielles et la contribution de van der Waals en utilisant un potentiel de Lennard-Jones avec des paramètres ε_0 et σ appropriés. Un assez grand nombre de termes sont ainsi inclus dans le modèle de champ de force, même pour une molécule aussi simple que le propane. Même ainsi, le nombre de termes (73) est beaucoup moindre que le nombre d'intégrales qui seraient impliquées dans un calcul quanto- mécanique équivalent.

II. 5- Champ de force MM2 et MM+

II. 5- 1- Champ de force MM2

* Elongation des liaisons : MM2 a été paramétré pour ajuster les distances moyennes calculées à partir du mouvement vibrationnel à température ambiante à celles obtenues par diffraction électronique.

Pour mieux reproduire la courbe de Morse, MM2 fait intervenir un terme quadratique et un autre cubique :

$$v(l) = \frac{k}{2}(l - l_0)^2\left[1 - k'(l - l_0)\right] \qquad (28)$$

* Variation des angles : les déviations des angles de leurs valeurs de références sont souvent écrites en utilisant la loi de Hooke ou potentiel harmonique :

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \qquad (29)$$

Comme pour les termes d'élongation des liaisons, la précision du champ de force peut être augmentée par l'incorporation de termes d'ordres supérieurs. MM2 contient un terme d'ordre 4 en plus du terme quadratique.

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2\left[1 - k'(\theta - \theta_0)\right] \qquad (30)$$

* Torsion des angles dièdres : correspond à la rotation d'une liaison selon l'angle dièdre ω formé par 4 atomes. Le champ de force MM2 utilise 3 termes d'une série de Fourier :

$$v(\omega) = \frac{v_1}{2}(1 + \cos\omega) + \frac{v_2}{2}(1 - \cos 2\omega) + \frac{v_3}{2}(1 + \cos 3\omega) \qquad (31)$$

Une interprétation physique a été attribuée à chacun des 3 termes à partir de l'analyse de calcul *ab initio* effectués sur des hydrocarbures fluorés simple.

* Angle dièdre impropre ou déviation extra- planaire : Examinons comment la cyclobutanone serait modélisée en utilisant uniquement un champ de force comportant des termes standards pour l'élongation des liaisons et les variations angulaires du type de ceux apparaissant dans l'équation (31). La structure d'équilibre obtenue avec un tel champ de force

sera caractérisée par la localisation de l'atome d'oxygène hors du plan formé par l'atome de carbone contigu (à l'oxygène) et les 2 carbones qui lui sont liés (Fig.6).

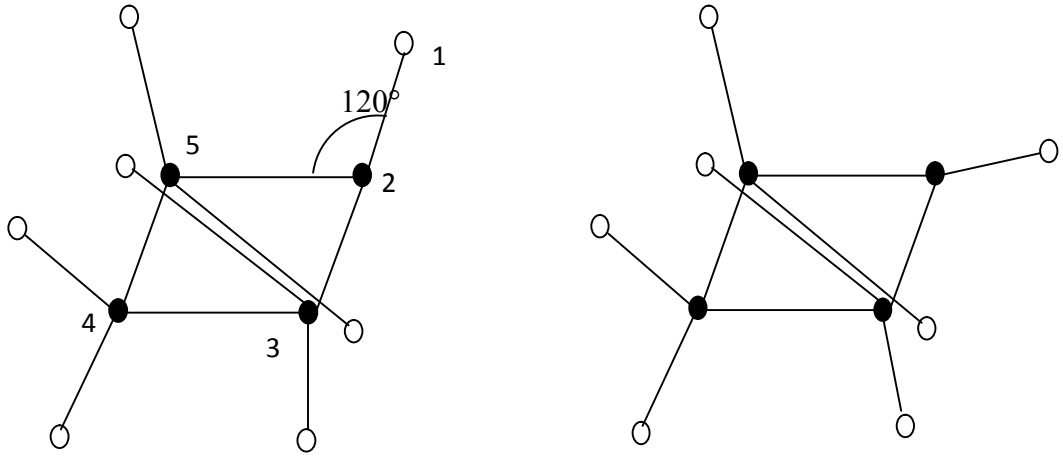


Figure 6: Sous un terme extra- planaire, l'atome d'oxygène de la cyclobutanone est prévu de se situer hors du plan du cycle / gauche plutôt que dans le plan.

Dans cette configuration, les angles vers l'oxygène adoptent des valeurs proches de la valeur de référence 120°. Expérimentalement, il est établi que l'atome d'oxygène demeure dans le plan du cyclobutane bien que les angles C-C=O soient grands (133°). Ceci parce que l'énergie de liaison π, qui est maximalisée dans l'arrangement coplanaire, sera plus réduite si l'atome d'oxygène est dévié hors du plan. Pour réaliser la géométrie désirée il est nécessaire d'ajouter un (ou des) terme (s) additionnel (s) dans le champ de force qui maintienne (nt) le carbone sp^2 et les 3 atomes qui lui sont liés dans le même plan. La plus simple façon de le réaliser est d'utiliser un terme de variation angulaire extra- planaire.

Il y a plusieurs façons d'incorporer des termes de variation extra- planaire dans un champ de force. Une approche possible est de traiter les 4 atomes comme un angle de torsion impropre pour lequel les 4 atomes (Fig.7) ne sont pas liés dans la séquence 1- 2- 3- 4. Une façon de définir, dans ce cas, une torsion impropre consiste à impliquer les atomes 1- 5- 3- 2 de la figure.

Un potentiel de torsion de la forme suivante :

$$v(\omega) = k\left(1 - \cos 2\varpi\right) \qquad (32)$$

Peut être utilisé pour maintenir l'angle de rotation impropre à 0° ou 180°.

Il existe diverses autres façons pour inclure la contribution de la variation d'angle extra-planaire. Par exemple, une définition qui est la plus proche de la notion de « variation extra-planaire » implique le calcul de l'angle entre une liaison à partir de l'atome central et le plan défini par l'atome central et les 2 autres atomes (Fig.7). La valeur 0° correspond aux 4 atomes coplanaires. Une troisième approche consiste à calculer la hauteur de l'atome central au-dessus du plan défini par les 3 autres atomes (Fig.7). Avec ces deux définitions la déviation de la coordonnée extra- planaire (angle ou distance) peut être modélisée à l'aide d'un potentiel harmonique de la forme :

$$v(\theta) = \frac{k}{2}\theta^2 \quad ; \quad v(h) = \frac{k}{2}h^2 \tag{33}$$

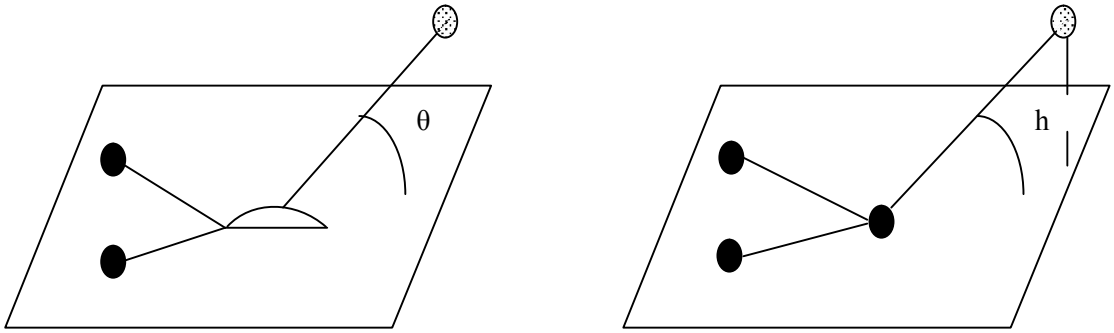


Figure 7: Deux façons pour modéliser les contributions de la variation d'angle extra- planaire.

***** Termes de croisement : les termes de croisement permettent de tenir compte des interactions entre l'élongation des liaisons, la variation des angles et la torsion des angles dièdres. Par exemple, si les liaisons C-O et O-H de l'angle C-O-H sont étirées, alors la distance entre les atomes terminaux (C et H) est augmentée, ce qui rend plus facile la diminution de l'angle COH. Pareillement, la diminution de l'angle COH tend a être accompagnée d'une augmentation des longueurs des liaisons O-H et C-O. Pour tenir compte de cette interaction, on peut ajouter un terme de croisement « élongation- variation angulaire ». (stretch- bend) de la forme :

$$v_{\Delta\theta} = \frac{1}{2} k_{12}(\Delta l_1 + \Delta l_2)\,\Delta\theta \tag{34}$$

avec $\Delta l_1 = l_1 - l_{1\,0}$; $\Delta l_2 = l_2 - l_{2\,0}$ et $\Delta\theta = \theta - \theta_0$

$l_{1\,0}$, $l_{2\,0}$ et θ_0 représententles valeurs de références pour l_1, l_2 et θ respectivement.

Les termes de croisement les plus utilisés sont (Fig 8) :

* élongation- élongation et élongation- variation angulaire, pour deux liaisons à un même atome ;

* élongation- torsion angle dièdre, variation angulaire- torsion angle dièdre et variation angulaire- variation angulaire pour 2 angles avec un atome central commun.

Le champ de force MM2 fait intervenir uniquement un terme de croisement élongation-variation angulaire.

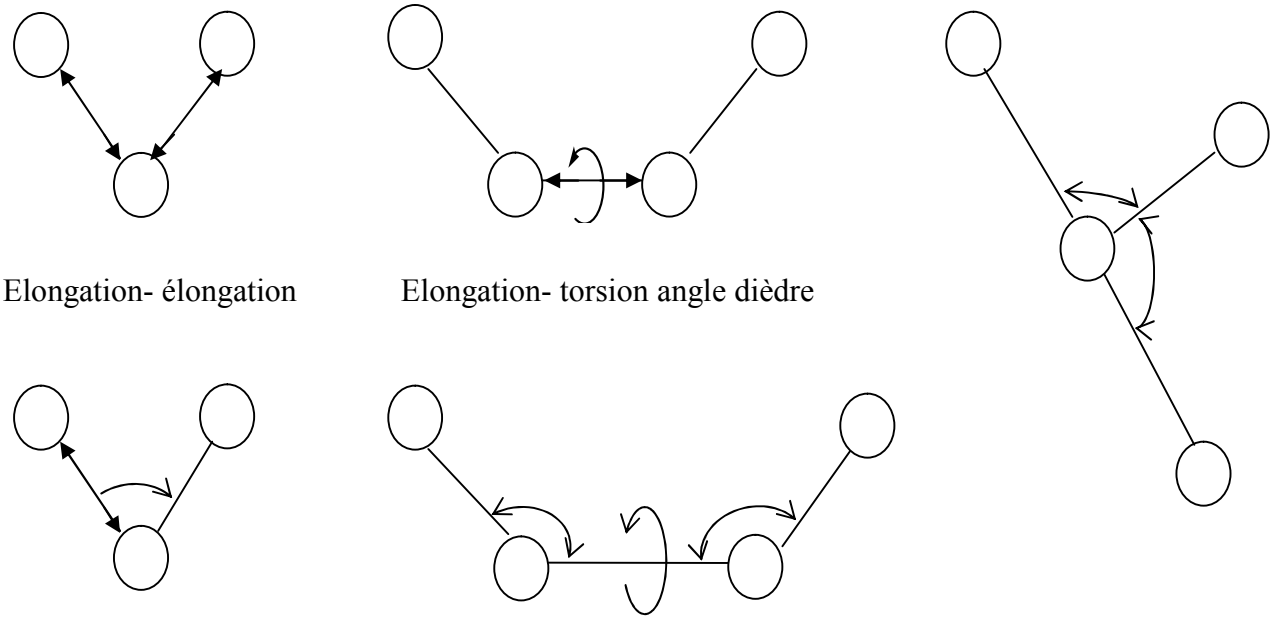


Figure 8: Illustration schématique des termes de croisement supposés les plus importants pour les champs de forces.

* Interactions électrostatiques : Le terme électrostatique V_{es} est généralement défini par la somme des interactions électrostatiques entre toutes les paires d'atomes exceptées les paires 1, 2 et 1, 3 : $v_{es} = \sum l \geq 4\ v_{es,ij}$, où les atomes i, j vérifient la relation ($l \geq 4$).

V_{es} est généralement calculé en affectant des charges atomiques partielles à chaque atome et en appliquant un potentiel coulombien.

MM2 n'utilise pas cette procédure mais associe un moment dipolaire avec chaque liaison puis calcule v_{es}comme somme des énergies potentielles d'interactions entre moments de

liaisons. Chaque moment dipolaire étant localisé au centre, et dirigé le long de cette liaison. Les valeurs des moments de liaisons de certains types de liaisons ont été choisies pour ajuster les moments dipolaires expérimentaux de petites molécules.

L'équation (35) d'écrit un modèle de charge, basé essentiellement sur les dipôles centrés, utilisé dans MM2 (Alliger N L, 1977).

$$v_{es} = \frac{\mu_i \mu_j}{kr^3}(\cos \chi - 3 \cos \alpha_i \cos \alpha_j) \qquad (35)$$

χ et α_i, α_j désignent respectivement les angles entre les dipôles et les angles entre chaque dipôle et le vecteur de connexion.

* Interactions de van der Waals : la plupart des champs de force utilisent le potentiel 12- 6 de Lennard Jones.

MM2 utilise le potentiel de Buckingham, avec un terme attractif proportionnel à r^{-6} et un terme répulsif proportionnel à $e^{-\alpha r}$ où α est un paramètre :

$$v_{v d w} = A\, e^{-\alpha r} - \frac{B}{r^6} \qquad (36)$$

* Liaisons conjuguées : MM2 utilise une procédure générale qui consiste à effectuer des calculs semi- empiriques sur les électrons π pour en tirer les ordres de liaisons, qui sont ensuite utilisés pour affecter des longueurs de liaisons et des constantes de foce de référence pour les liaisons conjuguées.

II. 5- 2- Champ de force MM+

C'est une extension du Champ de force MM2, avec l'ajout de quelques paramètres additionnels. MM+ est un champ de force robuste, il a l'aptitude de prendre en considération les paramètres négligés dans d'autres champs de force et peut donc s'appliquer pour des molécules plus complexes tels que les composés inorganiques.

Le tableau suivant (Ramachandran *et al.,* 2008) compare les trois techniques computationnelles majeures évoquées.

Tableau II: Etude comparative des techniques *ab initio*, semi- empirique et mécanique moléculaire (Ramachandran K I*et al.,* 2008)

ab initio	**Semi- empirique**	**Mécanique moléculaire**
- Prise en compte de tous les électrons. - Limité à quelques dizaines d'atomes. Nécessite un super calculateur - Peut être appliquée à des composés inorganiques, organométalliques, et aux fragments moléculaires (composants catalytiques d'enzymes). - Vide, solvatation implicite. - Applicable à l'état fondamental, et aux états de transition et excité.	- Ignore certains électrons (simplification). - Limité à quelques centaines d'atomes. - Peut être appliquée à des composés inorganiques, organiques, organométalliques et de petits oligomères (peptides, nucléotides, saccharides). - Vide, solvatation implicite. - Applicable à l'état fondamental, et aux états de transition et excité.	- Ignore tous les électrons. Seuls les noyaux sont considérés. - Molécules contenant des milliers d'atomes - Peut être appliquée aux composés inorganiques, organiques, oligonucléotides, peptides, saccharides, métallo- organiques et inorganiques. - Vide, solvatation implicite ou explicite. - Applicable uniquement à l'état fondamental.

III. Calcul des descripteurs moléculaires

Différents descripteurs moléculaires physico- chimiques reflétant la structure peuvent être déterminés empiriquement où en utilisant des méthodes théoriques et computationnelles de différentes complexités. Il est à souligner que la connaissance de la constitution chimique exacte et/ou de la structure moléculaire tridimensionnelle des composés chimiques étudiés est un pré- requis à l'application de l'approche QSAR/QSPR.

Le succès de l'approche QSAR/QSPR dépend de façon critique de la définition précise et de l'utilisation appropriée des descripteurs moléculaires. On distingue, arbitrairement, **les descripteurs moléculaires empiriques**et les **descripteursmoléculaires théoriques**.

Les descripteurs empiriques peuvent être divisés en deux classes générales (tableau III), la première reflète les interactions électroniques intramoléculaires **(descripteurs structurels)** alors que la seconde tient compte des interactions intermoléculaires dans les milieux condensés tels que les liquides et les solutions **(descripteurs de solvatation)**.

Tableau III: Classification d'ensemble des descripteurs moléculaires empiriques

Classe	**Sous- classe**
Descripteurs structurels	- Constantes d'induction - Constantes de résonance - Constantes stérique
Descripteurs de solvatation	- Echelles de polarité - Echelles de polarisabilité - Echelles d'acidité - Echelles de basicité - Echelles mixtes

Les descripteurs structurels les plus répandus ont été définis pour quantifier les propriétés d'induction, l'effet mésomère ou de résonance, et les effets stériques des composés chimiques. Les descripteurs de solvatation reflètent les interactions du soluté avec la masse du solvant environnant **(effets de solvant macroscopiques** ou **non spécifiques)**, et les liaisons spécifiques, souvent des liaisons hydrogène entre le soluté et les molécules individuelles de

solvant **(effets de solvant spécifiques** ou **microscopiques)**. Les effets de solvant macroscopiques sont quantifiés en utilisant diverses échelles de polarité et de polarisabilité. Les descripteurs des effets de solvant microscopiques impliquent les échelles générales d'acidité et de basicité. Certaines échelles empiriques d'effets de solvant (échelles mixtes) peuvent impliquer en même temps ces deux effets macroscopique et microscopique. Le coefficient de partage octanol/ eau, log P, est le représentant typique de tels descripteurs.

Les descripteurs moléculaires théoriques peuvent, conventionnellement, être répartis en un certain nombre de classes, selon leur complexité ou leur méthode de calcul. Les descripteurs théoriques les plus simples sont des **descripteurs constitutionnels**ou des **descripteurs physico- chimiques** qui peuvent être construits à partir de l'information sur la composition chimique du composé considéré, ils caractérisent généralement la structure bidimensionnelle de la molécule. Certains reflètent la composition moléculaire du composé soit les nombres, absolus et relatifs, des différents types d'atomes et de liaisons chimiques, la masse molaire, et le nombre de différents cycles dans le composé représentent quelques descripteurs constitutionnels typiques(Bosque R *et al.,* 2003). Certains représentent la surface accessible au solvant (nommée « Connolly Accessible Surface »), le volume de solvant couvert par cette surface (« Connolly Solvent-Excluded Volume ») (Connolly M L, 1985), le caractère hydrophile ou lipophile de la molécule généralement évalué à partir du coefficient de partage octanol/ eau représenté par le logP (Viswanadhan V N *et al.,* 1989).

Les descripteurs, **ou indices**, **topologiques** décrivent la connectivité des atomes dans la molécule. On a avancé (Karelson M, 2000) que les indices topologiques pouvaient encoder des interactions moléculaires subtiles et non pas seulement renseigner sur le degré de ramification des liaisons chimiques ou la distribution de la masse spécifique dans la molécule.

Les descripteurs géométriques sont obtenus à partir de la structure tri-dimensionnelle des molécules définie par les coordonnées des noyaux atomiques et la grosseur de la molécule représentée, par exemple, par le rayon atomique de van der Waals. Les molécules de la plupart des composés chimiques possèdent une certaine flexibilité conformationnelle et les surfaces de potentiels moléculaires respectives possèdent de multiples minima locaux. Selon la structure de la molécule, le nombre de ces minima peut être très grand et, par conséquent, il est plutôt difficile de trouver le minimum d'énergie global pour des conditions expérimentales établies.

Evidemment, les descripteurs géométriques peuvent varier de façon significative selon la conformation utilisée dans le calcul de ces descripteurs. Dans une certaine mesure, **les descripteurs théoriques liés à la distribution de charge** peuvent également dépendre de la conformation(Bosque R *et al.,* 2003). Ces descripteurs sont basés sur la structure tri-dimensionnelle et la distribution des charges dans la molécule. Ces dernières peuvent se présenter comme charges atomiques partielles obtenues à partir d'un schéma empirique ou en utilisant des fonctions plus sophistiquées basées sur la fonction d'onde de la molécule calculée par la chimie quantique.

Un certain nombre de **descripteurs quanto- chimiques basés sur les OM** ont été employés dans le développement d'équations QSAR/QSPR. Les plus utilisés sont les énergies des OM frontières, c'est-à-dire, l'énergie calculée de la plus basse orbitale moléculaire inoccupée (ε_{LUMO}), et l'énergie de la plus haute orbitale moléculaire occupée (ε_{HOMO}), et la différence entre ces énergies. De même, différents indices de réactivité déduits de la théorie de la superdélocalisabilité de Fukui ou d'autres constructions théoriques ont gagné en popularité parmi les chercheurs.

Tous les descripteurs théoriques ne peuvent être strictement classés selon le schéma présenté dans le tableau IV. Par exemple, les indices topographiques sont déduits de l'information contenant à la fois la topologie et la géométrie des molécules. **Les indices électrotopologiques** sont fondés sur la topologie et la distribution de charge alors que les aires de surfaces partielles chargées sont des descripteurs qui encodent à la fois la distribution de charge et la géométrie des molécules. De tels descripteurs peuvent être classés comme **descripteurs moléculaires mixtes ou combinés**.

Les descripteurs moléculaires peuvent être définis pour tout le système moléculaire étudié ou pour n'importe laquelle de ses parties (fragments). Par exemple, la majorité des descripteurs empiriques structurels sont reliés à des fragments moléculaires appelés substituants. En conséquence, les molécules d'une série congénère de composés chimiques sont divisées formellement en deux ou plusieurs fragments qui correspondent à une unité structurale constante Y (c'est-à-dire le centre de réaction) et à des unités structurales variables Xi (les substituants). Les relations QSAR/QSPR sont ainsi présentées comme suit :

$$P = P_0^{(Y)} + \sum_i \sum_k \alpha_{ik}^{(Y)} D_{ik}^{(X)} \qquad (37)$$

Où $P_0^{(Y)}$ est l'ordonnée à l'origine correspondant au fragment moléculaire constant Y, les $D_{ik}^{(X)}$ sont les descripteurs moléculaires de type k pour les fragments variables X_i, et les $\alpha_{ik}^{(Y)}$ sont les coefficients de développement caractéristiques d'une série donnée de composés X_iY.

Tableau IV: Classification générale des descripteurs moléculaires théoriques

Classe	Sous- classe
Descripteurs constitutionnels	- Dénombrement des atomes ou des liaisons. - Descripteurs basés sur les masses atomiques.
Descripteurs topologiques	- Indices topologiques (connectivité). - Descripteurs théoriques d'information. - Descripteurs topochimiques.
Descripteurs géométriques	- Descripteurs liés à la distance. - Descripteurs liés à l'aire de la surface. - Descripteurs liés au volume. - Descripteurs du champ stérique moléculaire.
Descripteurs liés à la distribution de charge	- Charges atomiques partielles. - Moments électriques moléculaires - Polarisabilités moléculaires. - Descripteurs du champ électrique moléculaire.
Descripteurs liés aux orbitales moléculaires	- Energie des OM frontières - Ordres de liaison - Indices de réactivité de Fukui.
Descripteurs température dépendants	- Fonctions thermodynamiques. - Descripteurs facteurs de Boltzmann pondérés.
Descripteurs de solvatation	- Energie électrostatique de solvatation. - Energie de dispersion de solvatation. - Enthalpie libre de formation de cavité. - Descripteurs de liaison hydrogène. - Entropie de solvatation. - Descripteurs d'énergie de solvatation linéaire théorique.
Descripteurs mixtes	- Descripteurs topographiques. - Descripteurs électrotopologiques. - Descripteurs de la charge partielle de l'aire de la surface.

La plupart des descripteurs théoriques qui apparaissent dans le tableau IV peuvent être calculés soit pour la molécule entière soit pour un fragment moléculaire pré- défini.

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 (Hyperchem^{TM}, 2000) pour représenter les molécules puis, à l'aide de la méthode semi-empirique AM1, 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 1600 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, est éliminé automatiquement celui qui présente les plus hautes corrélations croisées avec les autres descripteurs.

IV. Méthodes de sélection des ensembles de calibrage et de test :

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

-les échantillons retenus doivent présenter une variabilité maximale;

- la plage de variation des valeurs doit être la plus grande possible, mais limitée aux valeurs rencontrées dans la pratique;

- les échantillons doivent être uniformément répartis.

Dans La littérature, cette tâche a été exécutée en utilisant beaucoup et différentes méthodes de sélection d'échantillons, chacune avec ses avantages et ses inconvénients.Plusieurs méthodes de sélection d'échantillons (algorithme de Kennard-Stone, algorithme DUPLEX, sélection aléatoire des échantillons, OPTISIM, Répartition uniforme des échantillons sur la variable dépendante, etc...) peuvent être utilisées (Bouveresse D J R, 2004).

IV. 1. Choix aléatoire :

L'échantillonnage aléatoire simple (au hasard) est la méthode la plus courante pour le fractionnement des données dans le développement des modèles, où les données sont sélectionnées avec une probabilité uniforme. L'échantillonnage au hasard simple estfacile à réaliser et peut être efficacement exécuté dans un seul passage sur les données en utilisant des algorithmes tels que l'algorithme de Knuth (Knuth D E, 1997). Cependant, le problème avec cette approche est qu'il ya une chance que la scission de données souffre de la variance, ou de partialité, en particulier lorsque les données ne sont pas réparties uniformément (Tourassi G A, 2001).

IV. 2. Algorithme DUPLEX :

Une version améliorée appelée DUPLEX a été proposée par Snee (Snee R D, 1977); il est largement utilisé dans le domaine de la chimiométrie, y compris plusieurs applications ANN (Despagne F & Massart D L, 1998, Sprevak D *et al.,* 2000). Cependant, la complexité de calcul de cet algorithme peut interdire son utilisation sur de grands ensembles de données.Par ailleurs, selon un travail récent de Ren *et al.* (Ren Y Y *et al.,* 2007), DUPLEX est l'une des meilleures méthodes pour diviser les données en un ensemble d'apprentissage et un ensemble de test, qui mesure la distance entre tous les échantillons par la distance euclidienne.

Cet algorithme commence avec la liste des n observations, les ℓ régresseurs étant standardisés à l'unité selon :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j\sqrt{n-1}} \; i = 1,..\,n \;\; ; \;\; j = 1, .., \ell \qquad (37)$$

Où

s_j : Ecart-type du jème régresseur.

$\bar{x}_j$: Moyenne du jème régresseur.

x_{ij} : Valeur du régresseur j pour la i ème observation.

n : Nombre d'observations.

Les régresseurs standardisés sont alors orthonormalisés en factorisant le produit à gauche de la matrice $\mathbf{Z} = (z_{ij})$ par sa transposée $\mathbf{Z'}$, sous la forme :

$$\mathbf{Z'.Z} = \mathbf{T'.T} \tag{38}$$

$\mathbf{T}$ est une matrice ($\ell \times \ell$) triangulaire supérieure unique, dont les éléments peuvent être obtenus par la méthode de Cholesky (Graybill F A, 1976). On opère alors la transformation :

$$\mathbf{W} = \mathbf{Z.T^{-1}} \tag{39}$$

qui conduit à un nouvel ensemble de variables w orthogonales et de variance unité. Celles-ci sont utilisées pour calculer la distance euclidienne, entre les C_n^2 paires de points. Les 2 points les plus éloignés sont sélectionnés pour l'ensemble de calibrage, puis parmi les points restants, les 2 plus éloignés sont sélectionnés pour la validation (ensemble de test). Puis parmi les points restants, le plus éloigné des points de calibrage précédemment sélectionnés est sélectionné pour le calibrage. Puis parmi les points restants, le plus éloigné des points de validation précédemment sélectionnés est sélectionné pour la validation. Puis l'algorithme continue à placer les points restants, alternativement dans l'ensemble de calibrage et dans l'ensemble de validation, jusqu'à ce que les n points soient affectés.Les ensembles de calibrage et de validation n'étant pas forcément de même taille, l'algorithme DUPLEX peut séparer les données dans n'importe quel rapport souhaité. De telles séparations sont réalisées en utilisant l'algorithme jusqu'à ce que l'ensemble de validation contienne le nombre de points requis, puis en versant les points non assignés dans l'ensemble de calibrage.L'utilisation de l'algorithme DUPLEX suppose que le nombre d'observations, n, est tel que : $n \geq 2\ \ell + 25$, ℓ désignant le nombre de régresseurs ; l'ensemble de validation devant contenir 15 éléments au minimum.

Par conséquent, il garantit que la composition de l'ensemble de calibrage et de l'ensemble de test ne présente pas, en même temps, un déséquilibre des deux ensembles de données (Jin C *et al.,* 2008).

V. Développement de modèles QSAR/QSPR

V. 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 (TodeschiniR*et al.,* 2004), en maximisant Q^2_{LOO}.

V. 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 (TodeschiniR*et al.,* 2005). 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 de la 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 dans de grands espaces 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. Le premier modèle, et le plus largement utilisé, consiste en une équation (multi) linéaire obtenue par régression des données expérimentales en fonction d'un ensemble de descripteurs pré- sélectionnés (ou d'un seul), en utilisant la méthode des moindres carrés ordinaires (MCO). Dans quelques cas, les modèles physiques ou chimiques connus du phénomène étudié laissent prévoir certaines formes mathématiques non linéaires (exponentielles ou logarithmiques) de la dépendance entre les données expérimentales et les descripteurs moléculaires. Les modèles QSAR/QSPR peuvent alors être établis à l'aide de la technique de régression par les moindres carrés non linéaires. D'autres modèles ont été développés en utilisant l'analyse factorielle ou l'analyse en composantes principales. L'intérêt de ces méthodes est qu'elles évacuent le problème de multicolinéarité 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 (RLM) et d'analyse en

composantes principales (ACP) est la technique de régression par les moindres carrés partiels (MCP ou PLS) (Kowalski B *et al.,* 1982, Erikson L *et al.,* 2001, Wold S *et al.,* 1984, Wold S, 1984, Gelada P & Kowalski B R, 1986, Höskuldsson A,1988).

On a également appliqué les méthodes modernes de l'intelligence artificielle au développement de modèles QSAR/QSPR (Burns J A & Whiteside G M, 1993, Anker L S & Jurs P C, 1992, Aoyama T *et al.,* 1990, Andrea T A, 1991, Jurs P C, 1996). Ces méthodes comprennent: les réseaux de neurones (RNA), les algorithmes génétiques (GA), et d'autres méthodes globales d'optimisation.

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

V. 2. 1. La régression linéaire multiple :

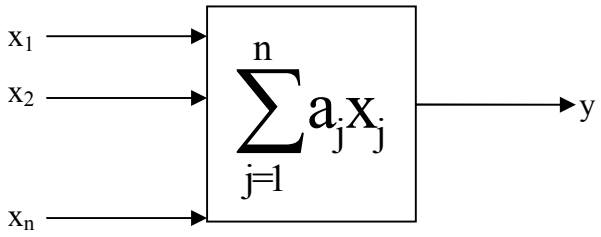
L'étude d'un phénomène peut, le plus souvent, être schématisé de la manière suivante: on s'intéresse à une grandeur y, que nous appellerons par la suite réponse ou variable expliquée, qui dépend d'un certain nombre de variables x_1; x_2; ….x_n que nous appellerons facteurs ou variables explicatives.

La régression est une des méthodes les plus connues et les plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une liaison entre une variable quantitative et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de régression simple en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs autres variables, on parlera de régression multiple. La mise en œuvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle (Chouquet C, 2010).

La régression multi-linéaire (MLR, pour Multiple LinearRegression) (Lejeune M, 2004) est la méthode la plus simple et la plus communément employée pour le développement de modèles prédictifs. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante y (ici, la propriété) et une série de n variables indépendantes xi (ici, les descripteurs). L'objectif est d'obtenir une équation de la forme suivante :

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n \qquad (40)$$

où a_i sont les coefficients de la régression.



La détermination de l'équation (41) se fait alors à partir d'une base de données de p échantillons pour laquelle à la fois les variables dépendantes et la variable indépendante sont connues. Il s'agit donc de considérer un système de p équations.

$$\begin{aligned} \hat{y}_1 &= a_0 + a_1 x_{1,1} + a_2 x_{2,1} + \ldots + a_n x_{n,1} + \varepsilon_1 \\ \hat{y}_2 &= a_0 + a_1 x_{1,2} + a_2 x_{2,2} + \ldots + a_n x_{n,2} + \varepsilon_2 \\ \hat{y}_p &= a_0 + a_1 x_{1,p} + a_2 x_{2,p} + \ldots + a_n x_{n,p} + \varepsilon_p \end{aligned} \qquad (41)$$

où les résidus ε_i représentent l'erreur du modèle, constituée par l'incertitude sur la variable dépendante y_i d'une part, sur les variables indépendantes x_i d'autre part, mais aussi par les informations contenues dans les variables indépendantes mais non exprimées via les variables dépendantes.

Ce système d'équations peut être écrit sous la forme matricielle suivante :

$$\begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{n,1} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{1,p} & \ldots & x_{1,p} \end{pmatrix} \begin{pmatrix} a_0 \\ . \\ . \\ . \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_p \end{pmatrix} \qquad (42)$$

soit de manière condensée :

$$Y = X\,A + \varepsilon \qquad (43)$$

La méthode consiste alors à choisir les coefficients du vecteur A en faisant en sorte de minimiser la somme des carrés des écarts entre les valeurs prédites et les valeurs réelles sur l'intégralité de la base de données et ceci sous couvert de certaines hypothèses de départ.

En premier lieu, les variables indépendantes xi, comme leur nom l'indique, sont supposées indépendantes entre elles et leur incertitude est négligeable. Ensuite, les différents échantillons yi sont supposés indépendants entre eux et suivent une distribution normale. L'erreur ε est elle-même supposée suivre une distribution normale, centrée en 0. Enfin, par nature, la dépendance de y vis-à-vis des xi est supposée linéaire.

La valeur prédite de la variable dépendante est alors :

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_{1,i} + \ldots + \hat{a}_n x_{n,i} \tag{44}$$

Les résidus peuvent donc être définis comme la différence entre les valeurs prédites et observées de y.

$$\varepsilon_i = y_i - \hat{y}_i \tag{45}$$

Il s'agit alors de trouver les coefficients $\hat{a}_i$ afin de minimiser la somme des carrés de ces résidus pour l'intégralité de la base de données.

$$\min [\sum(\varepsilon_i)^2] = \min [\sum(y_i - \hat{y}_i)^2] = \min [\sum(y_i - \hat{a}_0 - \hat{a}_1 x1_{,i} - \ldots - \hat{a}_n x_{n,i})^2]$$

$$= \min (Y - X\hat{A})^T (Y - X\hat{A}) \tag{46}$$

Les coefficients peuvent être obtenus à partir de l'équation matricielle suivante :

$$\hat{A} = (X^T X)^{-1} X^T Y \tag{47}$$

Bien entendu, la régression multi-linéaire souffre de certains désavantages. Le principal découle de sa linéarité. Elle est donc défaillante pour la mise en évidence de dépendances non-linéaires. Cela dit, elle n'en reste pas moins une méthode simple et efficace dans la plupart des cas.

De plus, pour peu que les variables indépendantes soient choisies de manière raisonnée, les équations obtenues peuvent être interprétées d'un point de vue phénoménologique (Fayet G, 2010).

V. 2. 2. Analyse en composantes principales (ACP):

L'analyse en composantes principales est une des techniques les plus anciennes et les plus connues de l'analyse multi- variée (Malinowsky E R & Howery D G, 1980, Strouf O, 1986, Jollcliffe I T, 1986, Meloum M *et al.,* 1992, Escofier B & Pages J, 1998, Lebart L *et al.,* 2004). L'ACP a été « inventée » en 1901 par Karl Pearson (Chatfield C & Collins A J, 1980). Actuellement, l'ACP est utilisée 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 :

$$p_i = \sum_{j=1}^{v} c_{ij}\, x_j \qquad (48)$$

ou p_i est la ième 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 précisement, 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 nonencore prises en compte par les précédents axes principaux. Chaque composante principale correspond à un axe dans l'espace de dimenssion 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 (Wold H, 1966). 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 **D** avec v lignes et s colonnes. La matrice **Z** de variance-covariance est alors :

$$\mathbf{Z} = \mathbf{D}^{T}\mathbf{D} \tag{49}$$

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 : $|\mathbf{Z}-\lambda\mathbf{I}| = 0$, 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 ième composante propre tient compte d'une proportion : $\lambda_i \,|\sum_{j=1}^{v} \lambda_j$ de la variance totale des données.

V. 2. 3. 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.

V. 2.4. La régression PLS

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) mais devient très populaire en chimie grâce au fils d'Herman, Svante. Cette méthode connait un très grand succès dans le domaine de la chimie, particulièrement dans les applications concernant des données de chromatographie ou de spectrographie. De plus SvanteWold, NounaKettanech- 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).

La régression PLS est une techniquerécente 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 (Tenenhaus M, 1998).

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 (Vancolen S, 2004).

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 (Tenenhaus M, 1998), Pagès et Tenenhaus (Pagès J & Tenenhaus M, 2001). 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) (Tenenhaus M, 1998) 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) Yi – dans le cas d'une analyse QSPR, une ou des propriété(s) - dont la variance est expliquée par un nombre de variables indépendantes Xi – les descripteurs moléculaires.

La PLS repose sur une projection des Xi 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 Yi.

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, SIMCA-P+, 2005).

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 (GauchiJP, 1995). On trouve donc des utilisations de cette méthode pour d'autres types d'applications telles que le traitement d'images (Mesquita D P *et al.,* 2009).

V. 2. 5. Méthode des réseaux de neurones artificiels :

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

L'origine des réseaux de neurones peut être attribuée à McCulloh et Pitts en 1943 (McCulloh W S & Pitts W, 1943) ; ils proposent un modèle mathématique décrivant le fonctionnement d'un neurone biologique. Dans les années 80, Hopfield suscite à nouveau l'intérêt des scientifiques en proposant des neurones associatifs (Hopfield J J, 1982).

Un réseau de neurones est composé d'unités de calculs « le neurone artificiel » (Figure 9) disposées en couches et reliées entr' elles pour échanger de l'information (Hinton G, 1992). Le réseau contient trois types de couches ; les couches d'entrée, les couches cachées et les couches de sortie. L'information circule des neurones d'entrée vers les neurones de sortie sans retour arrière possible via des fonctions de transfert ou d'activation. Les valeurs d'entrée sont multipliées par leur poids correspondant et additionnées pour obtenir la somme S.

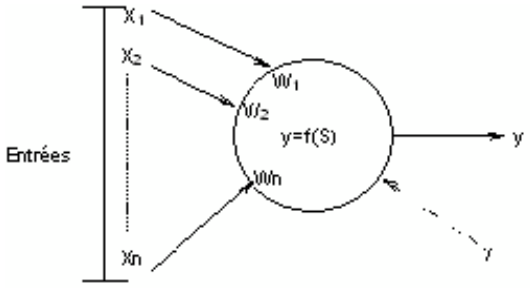


Figure 9: Le neurone artificiel générique.

Cette somme devient l'argument de la fonction d'activation, ces fonctions de transfert existent essentiellement sous trois formes : les fonctions linéaires, les fonctions seuils et les fonctions sigmoïdes. Ces dernières sont généralement les plus utilisées car elles représentent un bon compromis entre les fonctions seuils et linéaires (figure 10).

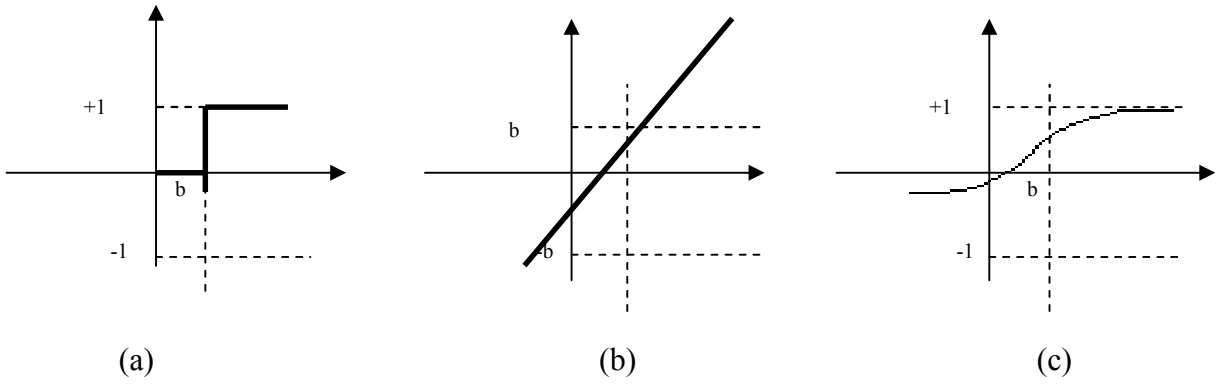


Figure 10 : Fonction de transfert (a) seuil, (b) linéaire et (c) sigmoïde du neurone.

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.

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 (Rumelbart D E *et al.,* 1988).

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 (Rumelbart D E *et al.,* 1988), 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 présentations distribuées.

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 :

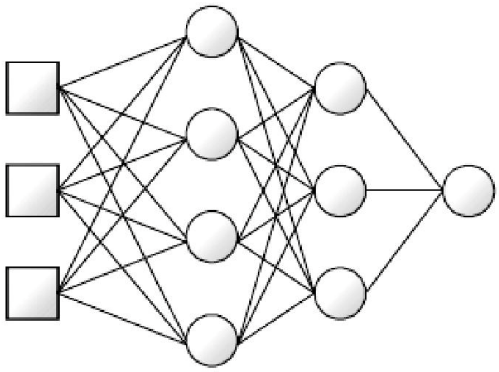
- Le réseau de Hopfield (Hopfield J J, 1982)(et sa version incluant l'apprentissage, la machine de Boltzmann) : est un réseau avec des sorties binaires où tous les neurones sont interconnectés avec des poids symétriques. C'est-à-dire que le poids du neurone N_i au neurone N_j est égal au poids du neurone 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 (Kohonen T, 1995) : sont utilisées pour faire des classifications automatiques des vecteurs d'entrées.
- Les réseaux à fonction radiale que l'on nomme aussi RBF (pour " Radial Basic Functions ") : 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 ou perceptron multicouches PMC : Les réseaux multicouches (PMC) sont les réseaux les plus puissants des réseaux de neurones qui utilisent l'apprentissage supervisé. Ces réseaux (figure 11) se composent des entrées, une couche de sortie et zéro ou plusieurs couches cachées (RumelbartD E*et al.,* 1988). 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 11: Structure générale du perceptron multicouche : shéma de principe

Il a été prouvé (Hecht N R, 1990) 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.

V. 2. 6. Machinesà vecteurs supports SVM

La régression par Machines à Vecteurs de Support (SVR) (Vladimir N & Vapnik V, 1995) consiste à trouver la fonction f(x) qui a au plusune déviation " par rapport aux exemples d'apprentissage (xi; yi), pour i = 1,…, N, et qui est la plus platepossible. Cela revient à ne pas considérer les erreurs inférieures à ε et à interdire celles supérieures à ε (Smola A J & Schölkopf B, 2004). Maximiser la platitude de la fonction permet de minimiser la complexité du modèle qui influe sur ses performances en généralisation. En effet, la théorie de l'apprentissage (Vladimir N & Vapnik V, 1995) permet de borner l'erreur de généralisation par unesomme de deux termes : l'un dépendant de la complexitédu modèle et l'autre dépendant de l'erreur sur les données d'apprentissage (Cristianini N & Taylor J S, 2000). Les méthodes SVMs sont basées sur le contrôle de la complexité du modèle lors del'apprentissage.

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

l'erreur de généralisationqui peut être évaluée sur un jeu indépendant de données de validation ou par validation croisée (Christopher M, 1995). Cela implique de réaliser l'apprentissage pour différentes valeurs et d'estimer leur performances. Dans le cas d'une estimation de l'erreur de généralisation par validation croisée, cette procédure peut se révéler très coûteuse en temps de calcul.

V. 3. Evaluation d'un modèle QSAR/ QSPR

Nous avons utilisé le logiciel de modélisation moléculaire HyperChem 6.03 (HyperChem^{TM} Release 6.03 for Windows, 2000) pour représenter les molécules puis, à l'aide de la méthode semi-empirique AM1, PM3 (Dewar M J S *et al.,* 1985), on a obtenu les géométries finales. Tous les calculs ont été menés dans le cadre du formalisme RHF (Levine I N, 2000) sans interaction de configuration. Les structures moléculaires ont été prèoptimisées à l'aide de l'algorithme Polak- Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,001kcal/mol. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique Dragon version 5.3 (Todeschini R *et al.,* 2006) pour le calcul de 1200 descripteurs appartenant à différentes classes. Les descripteurs d'un même groupe, à valeur constante (écarts types inférieurs à 0,0001) ont été exclus. Pour un seuil de corrélation de $R \geq 0,95$ entre deux descripteurs ; celui qui présente le plus de corrélations avec les autres variables, est exclu.

La MLR et la PLS ont été utilisées comme techniques linéaires, alors que les machines à vecteurs supports (SVM) et les réseaux de neurones artificiels (RNA) ont été employés comme techniques non linéaires pour la construction des modèles QSPR. La SVM est une nouvelle méthode de classification et de régression proposée par Vapnik (Vapnik V, 1995).

L'analyse de régression linéaire multiple (MLR) et la sélection des variables ont été effectuées par le logiciel MobyDigs (Todeschni R *et al.,* 2009) en utilisant la méthode ordinaire de régression des moindres carrés (OLS) et l'algorithme génétique(GA-VSS) pour la sélection des sous-ensembles (Leardi R *et al.,* 1992).

Nous avons appliqué la PLS sur le même ensemble de pesticide et les mêmes sous ensembles de calibrage et de validation utilisé pour la modélisation de la pression de vapeur par la RLM.

Les modèles ont été justifiés par le R^2, le R^2 ajusté, les valeurs de la validation croisée de Q^2 par leave-one-out (LOO), les valeurs de ratio F et l'erreur s standard.

La robustesse des modèles et leur prédictivité ont été évalués par les deux Q^2_{LOO} et Q^2 bootstrap. Dans cette dernière procédure K des groupes n-dimensionnelles sont générés par une sélection répétée au hasard des objets n- de l'ensemble de données d'origine.

Le premier modèle obtenu sur les objets sélectionnés est utilisé pour prédire les valeurs de l'échantillon exclu, et Q^2 est calculée pour chaque modèle. On répète l'opération 8000 fois.

Le modèle proposé a également été vérifiée pour la fiabilité et la robustesse par un test de permutation: les nouveaux modèles sont recalculés pour une réponse enregistrée de façon aléatoire (Y- scrambling) en utilisant la même matrice variable indépendante d'origine. Après avoir répété ce test plusieurs fois (100 fois dans ce travail), il est prévu d'obtenir de nouveaux modèles qui ont nettement inférieur R^2 et Q^2 que le modèle original. Si cette condition n'est pas vérifiée le modèle original n'est pas acceptable, car il a était due à une corrélation de hasard ou une redondance structurelle dans l'ensemble de calibrage.

L'obtention d'un modèle robuste ne donne pas des informations réelles sur son pouvoir de prédiction. Ceci est évalué en prédisant les composés inclus dans l'ensemble de test.

L'analyse des résidus présente un intérêt à plusieurs égards. Elle permet en effet de vérifier, a posteriori, la validité du modèle utilisé, en ce qui concerne, d'une part la forme de celui- ci (linéarité ou non linéarité de la relation, par exemple) et d'autre part, certaines hypothèses plus spécifiques, telles que l'égalité des variances résiduelles, la normalité des résidus ou l'absence d'auto- corrélation.

Pour minimiser l'influence des erreurs de détermination des valeurs explicatives (ou régresseurs) sur la précision des résultats de la régression 4 à 5 données (variables dépendantes, ou encore observations) doivent, à la limite, être associées à chaque variable explicative. Le nombre de degrés de liberté final (n-p-1) doit être (Tomassone R *et al.,* 1983) tel que :

$$n- p- 1 \geq 10 \qquad (50)$$

n étant la dimension de l'échantillon, et p le nombre de variables explicatives entrant dans la construction du modèle.

Pour les modèles à plus de deux 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 obéissent (TomassoneR*et al.,*1983)à la condition FIV< 5.

Deux paramètres statistiques sont couramment utilisés pour l'évaluation de la qualité du modèle :

- Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (51)$$

Où $\bar{y}$ est la valeur moyenne des valeurs observées.

- La racine de l'écart quadratique moyen de prédiction :

$$\sigma_N = EQMP = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_{(i)})^2} = \sqrt{\frac{PRESS}{n}} \quad (52)$$

Il est intéressant de considérer, également, la racine de l'écart quadratique moyen calculé sur les ensembles de calibrage (EQMC), et sur l'ensemble de validation externe (EQMPext) :

$$EQMC = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \quad (53)$$

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}\left(y_i - \hat{y}_{(i)}\right)^2}{n_{ext}}} \quad (54)$$

La validation croisée par « leave – one - out » (LOO) (Wehrens R *et al.,* 2000) consiste à recalculer le modèle sur (n-1) observations, et à utiliser le modèle ainsi obtenu pour calculer la grandeur d'intérêt du composé écarté, notée $\hat{y}_{(i)}$. On répète le procédé pour chacune des grandeurs d'intérêt. La somme des carrés des erreurs de prédiction, désignée par le symbole PRESS (eq. (53)), est une mesure de la dispersion des estimations. On l'utilise pour définir le coefficient de prédiction (Wehrens R *et al.,* 2000):

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \quad (55)$$

Contrairement à R^2 qui augmente avec le nombre de paramètres du modèle, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO}. Une valeur de $Q^2_{LOO} > 0,5$ est considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente (Eriksson L *et al.,* 2003).

Si de petites valeurs de Q^2_{LOO} indiquent des modèles peu robustes, caractérisés par de faibles capacités prédictives internes, le contraire n'est pas nécessairement vrai. En fait, si une forte valeur de Q^2_{LOO} 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.

Evidemment, on peut être amené à écarter 2, 3 ou un plus grand nombre d'éléments à la fois, ce qui conduit aux procédures LMO (leave – many- out).

La validation interne peut être également réalisée en utilisant la technique du bootstrap : Q^2_{boot} (bootstrapping). Elle consiste à simuler m échantillons de même taille n que l'échantillon initial. Ils sont obtenus par tirage au hasard avec remise parmi les n individus observés au départ, ceux-ci ayant tous la même probabilité 1/n d'être choisis (Wehrens R *et al.,* 2000, DraperN R & Smith H, 1998). Contrairement aux validations croisées par LOO et LMO, les méthodes de bootstrap sont plus efficaces et plus stables.

Une validation externe supplémentaire selon (Golbraikh et Tropsha, 2002) est appliquée uniquement à l'ensemble de test. Selon les critères recommandés de Tropsha et al, un modèle de QSPR prédictive, doit assister aux conditions suivantes:

1) $Q^2_{EXT} > 0.5$ (56-a)

2) $R^2 > 0.6$ (56-b)

3) $(R^2 - R_0^2)/R^2 < 0.1$ and $0.85 < k < 1.15$ (56-c)

$(R^2 - R'^2_0)/R^2 < 0.1$ and $0.85 < k' < 1.15$ (56-d)

Ou

$$R = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \qquad (57\text{-a})$$

$$R_0^2 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \bar{y})^2} \qquad (57\text{-b})$$

$$R_0'^2 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \qquad (57\text{-c})$$

$$k = \frac{\sum (y_i \tilde{y}_i)}{\sum (\tilde{y}_i)^2} \qquad (57\text{-d})$$

$$k' = \frac{\sum (y_i \tilde{y}_i)}{\sum (y_i)^2} \qquad (57\text{-e})$$

où R est le coefficient de corrélation entre les valeurs calculées et expérimentales dans l'ensemble de test; R_0^2 (valeurs calculées par rapport aux observées) et $R_0'^2$ (valeurs observées par rapport aux calculées) sont les coefficients de détermination; k et k 'sont les pentes des droites de régressions passant par l'origine pour les valeurs calculées par rapport aux valeurs observées et observées par rapport aux calculées, respectivement; $y_i^{r_0}$ et $\tilde{y}_i^{r_0}$ sont tels que définis respectivement par : $y_i^{r_0} = k\,\tilde{y}$ et, $\tilde{y}_i^{r_0} = k'\,y$; et les sommations sont sur tous les échantillons dans l'ensemble de test.

La raison d'utiliser et exiger des valeurs de k qui sont proches de 1 est que lorsque les propriétés réelles par rapport prédites sont comparés, un ajustement précis est nécessaire, non seulement une corrélation.

V. 4. Domaine d'application:

Le domaine d'application (AD) (Tropsha A *et al.,* 2003 & Shen M *et al.,* 2004) est une région théorique dans l'espace définie par les descripteurs du modèle et la réponse modélisée, pour lequel un modèle QSPR donné devrait faire des prédictions fiables. Dans ce travail, l'AD structurel a été vérifié par l'approche des leviers (h_{ii}) (Weisberg S, 2005).

Un avertissement sur l'effet de levier important d'un échantillon est, en général, donné pour un h¨>= h*=3(m + 1)/n où n est le nombre total d'échantillons dans l'ensemble de calibrage et m le nombre de descripteurs impliqués dans la corrélation.

La présence de valeurs aberrantes en réponse (valeurs aberrantes en **Y**) et les composés structurellement influents (valeurs aberrantes en **X**) a été vérifiée par le diagramme de Williams (SCAN- Software for Chemometric Analysis- 1995. version 1.1), le tracé des résidus standardisés en fonction des valeurs des leviers.

I. Modélisation de la constante de Henry :

I. 1. Introduction :

La constante de Henry joue un rôle important dans le comportement des pesticides. Elle peut être déterminante sur la manière dont les pesticides vont migrer, et donc de fait sur les conséquences d'une pollution.

Différents modèles permettent de prévoir ce paramètre environnemental important. Les méthodes incrémentielles sont basées sur des caractéristiques structurelles comme le type d'atome, le type de liaison et l'environnement structurel local (Mackay D *et al.,* 2000). Les stratégies QSPR (pour Quantitative Structure /Property Relationship) mettent en jeu des propriétés physico-chimiques, des descripteurs structuraux comme les indices de connectivité, et des descripteurs reflétant la structure électronique (Mackay D *et al.,* 2000, Dearden J C & Schüürmann G, 2003). Signalons en plus, pour la solubilité hydrique et la constante de Henry, la possibilité d'utiliser des modèles basés sur la structure moléculaire et les modèles quantiques de solvatation, via l'enthalpie libre de solvatation ΔGs. Les résultats de ces modèles révèlent des différences substantielles dans les domaines d'application et dans les capacités de prédiction (Mackay D *et al.,* 2000, Dearden J C & Schüürmann G, 2003, Estrada E *et al.,* 2004.).

Dans ce chapitre nous avons appliqué la méthodologie QSPR, dans l'approche hybride algorithme génétique / régression linéaire multiple (AG/RLM), pour modéliser la constante de Henry d'une série d'herbicides.

Les données prélevées dans la littérature (Hansen O C, 2004) ont été, au préalable, séparées aléatoirement (commande Show Advanced du logiciel de traitement des données MINITAB (MINITAB, Release 13.31, Statistical software, 2000)en un ensemble de calibrage pour la sélection des descripteurs par algorithme génétique (Leardi R *et al.,* 1992) et le calcul du modèle QSPR, et un ensemble de validation uniquement utilisé pour la validation statistique externe.

La qualité de l'ajustement, ainsi que la robustesse du modèle, et ses capacités prédictives (interne et externe) ont été examinées. Enfin, le domaine d'application (DA) a été discuté à l'aide du diagramme de Williams (Eriksson L *et al.,* 2003, Tropsha A *et al.,* 2003).

I. 2. Résultats et discussion

I. 2.1. Modélisation de la constante de Henry d'un enssemble : triazines-carbamates

Les valeurs expérimentales de la cte H des 27 pesticides ont été séparées aléatoirement en deux sous-ensembles respectivement de 20éléments pour la sélection des variables explicatives puis le calcul du modèle, et de 7éléments pour la validation externe.

L'application de l'AG- VSS a conduit à plusieurs bons modèles pour la prédiction du logarithme de la constante de Henry sur la base de différents ensembles de descripteurs moléculaires.

Le meilleur modèle à quatre dimensions a été construit à l'aide des descripteurs : GATS4v, RDF015m, TPSA(NO), et F05[C-C].

L'équation du modèle optimale a la forme suivante :

log cte H = 6,86(±0,9466)– 2,28(±0,5369)**GATS4v**– 0,844(±0,0873)**RDF015m** – 0,127(±0,0068)**TPSA(NO)** + 0,753(±0,0733)**F05[C-C]** (58)

Ici, GATS 4v est l'autocorrélation Geary - lag 4 / pondéré par le volume atomique de van der Waals ; RDF015 m est la fonction radial de distribution - 1,5 / pondérée par les masses atomiques; TPSA(NO) est surface polaire topologique à l'aide des contributions polaire N, O etF05[C-C] est la fréquence de C-C a une distance topologique 1.

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

Tous les paramètres statistiques pertinents du modèle proposé sont présentés ci-après

R^2= 96,14 Q^2_{LOO} = 92,81 Q^2_{LMO} (20%) = 89, 64 Q^2_{BOOT} = 87,22

EQMP = 0,575 EQMC= 0,421 K_{xx}= 26,88 K_{xy} = 34,09

n= 20S= 0,486 F= 93,41

n_{ext}= $7Q^2_{ext}$ = 88,03 $EQMP_{ext}$ = 0,741

Les paramètres statistiques montrent que les quatre descripteurspermettent de corréler les constantes de Henry des 20 pesticides appartenant à deux classes chimiques. En effet, la valeur du coefficient de détermination signifie que 96,14% de la variabilité de log H, peut être expliquée par ces quatres descripteurs. La grande valeur de Fisher indique que le modèle est très significatif.

Les valeurs de la constante de Henry expérimentales calculées, et prédites pour l'ensemble de validation, ainsi que les valeurs des leviers et des erreurs standardisées sont regroupées dans le tableau V

Tableau V : Valeurs de log H expérimentales, calculée, prédites, h_{ii}, et e_{istd}

N°	**Composé**	**log H_{Exp}**	**log $H_{Calc,}$ Pred**	**hii**	**e_{istd}**
1	Butylate	0,44	0,55	0,264	0,35
2	Carbetamide	-4,84	-5,38	0,32	-2,01
3	Chlorpropham	-2,54	-3,43	0,265	-2,92
4	Cycloate*	0,09	-0,62	0,288	-1,73
5	Desmedipham	-5,16	-5,02	0,234	0,43
6	Eptc	0,19	-0,04	0,21	-0,68
7	Pebulate*	0,32	-0,58	0,2	-2,07
8	Phenmedipham	-5,24	-4,83	0,193	1,15
9	Prosulocarb	-1,02	-0,45	0,228	1,74
10	Thiobencarb	-1,4	-0,92	0,206	1,4
11	Triallate	-0,24	-1,31	0,307	-3,81
12	Vernolate*	0,32	-0,78	0,191	-2,5
13	Vinclozolin*	-0,62	-1,04	0,138	-0,93
14	Ametryn	-3,16	-3,4	0,107	-0,59
15	Atrazine	-3,34	-3,31	0,115	0,08
16	Cyanazine*	-6,73	-5,86	0,149	1,94
17	Desmetryn*	-3,28	-3,78	0,243	-1,17
18	Dipropetryn	-2,91	-2,5	0,168	1,12
19	Hexazinone	-7,38	-7,39	0,852	-0,26
20	Metamitron	-6,23	-6,03	0,28	0,68
21	Metribuzin	-6,73	-6,48	0,386	1,08
22	Prometon	-3,36	-3,31	0,343	0,2
23	Prometryn	-3,04	-2,97	0,081	0,17
24	Propazine	-3,22	-2,9	0,082	0,75
25	Simazine	-3,47	-3,65	0,276	-0,6
26	Terbutylazine*	-3,22	-2,81	0,086	0,88
27	Terbutryn	-3,04	-2,92	0,083	0,28

*composés de validation

La qualité de l'ajustement a été vérifié par le graphe des valeurs calculées et prédites du logarithme de la constante de Henry en fonction des celles expérimentales. Le graphe présenté dans la figure 12, fait ressortir une faible dispersion autour de la première bissectrice.

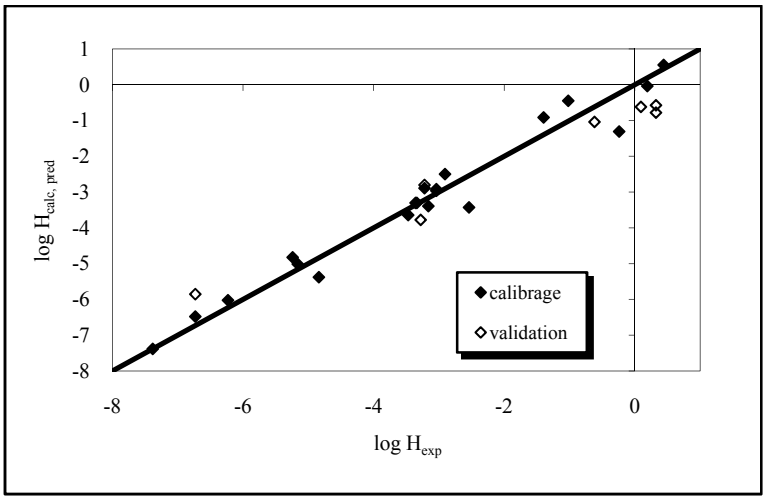


Figure 12 : Graphe des valeurs log cte H calculées prédites en fonction des valeurs observées

Le domaine d'application a été discuté à l'aide du diagramme de Williams (figure 13) qui représente les résidus de prédiction standardisés en fonction des valeurs des leviers (h_i).

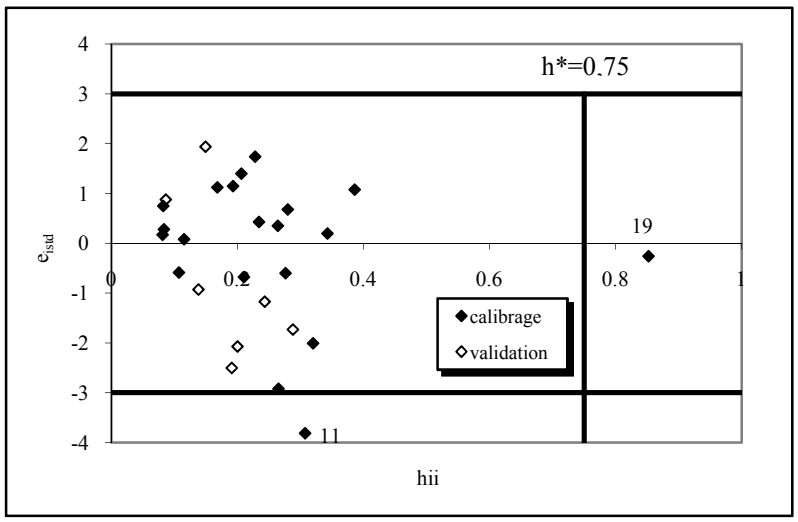


Figure 13 : Diagramme de Williams

Comme on peut le voir sur le diagramme de Williams, un point influent de l'ensemble de calibrage (hexazinone) et au même temps il représente un point aberran.

Le diagramme de Williams (Figure 13) fait ressortir parmi les éléments de l'ensemble de calibrage un point aberrant et un point influent.

La figure suivante (Figure 14) permet de comparer les résultats obtenus pour les modèles randomisés (cercles noircis) au modèle réel de départ (cercle vide).

Il est clair que les statistiques obtenues pour les vecteurs modifiés du logarithme de la constante de Henry sont plus petites que celles du modèle QSPR réel, et on obtient aussi des $Q^2 < 0$. Ce qui permet de s'assurer que le modèle établi a une base réelle et qu'il n'est donc pas fortuit.

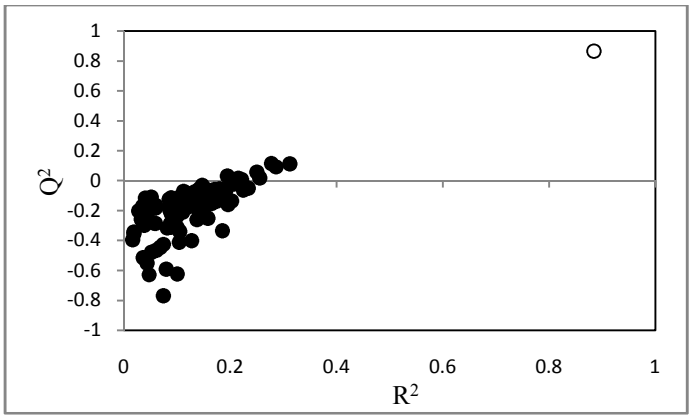


Figure 14 : test de randomisation

I. 3. Conclusion

La constante de Henry étant une mesure de l'affinité relative d'un composé pour la phase vapeur et l'eau, et l'état gazeux étant proche de l'état idéal, H dépendra essentiellement des interactions dans la phase aqueuse.

Le modèlemulti-linéaire présentéconduit au meilleur modèle à tous les points de vue : qualité de l'ajustement, robustesses interne et externe, capacité prédictive.

La modélisation de la constante de Henry d'un mélange hétérogène d'herbicides nous a conduit à un bon modèle acceptable ($Q^2 > 0,7$) à tous les points de vue: qualité de l'ajustement, robustesse…, le diagramme de Williams fait ressortir deux points de l'ensemble de calibrage hors du domaine d'application, un quiprésente un bras de levier important ($h_i > h^*$) et un autre représente une erreur standardisée hors les limites ± 3.

II. Modélisation de la solubilité aqueuse

II. 1. Introduction :

L'utilisation massive de produits agrochimiques, connus génériquement comme pesticides (Price N R & Watkins R W, 2003), a permis une réduction significative des fléaux agricoles, et par conséquent, une augmentation de la productivité. D'autre part, l'utilisation massive de ces produits a un coût environnemental (en raison de leur toxicité, persistance ou de leur tendance à la bioaccumulation), qu'il est nécessaire de connaître pour concilier la productivité et la protection de l'environnement (Stevens J T& Breckenridge C B, 2001).

La solubilité dans l'eau est une propriété physicochimique importante, ayant de nombreuses applications dans la modélisation des effets environnementaux des produits chimiques (Mackay D, 20003). C'est une mesure directe de l'hydrophobie, à savoir la tendance de l'eau à éliminer la substance de la solution. Bien que la détermination expérimentale de la solubilité ne soit pas difficile, il y a quelques justifications pour développer des modèles qui peuvent la prédire. Ceci est particulièrement important dans les études de l'environnement où les composés sont toxiques, cancérigènes, ou indésirables pour une raison ou une autre.

Une vaste série d'études pour la prédiction de la solubilité aqueuse a été rapportée dans la littérature (Lipinski C A *et al.,* 2001, Jorgensen W L & Duffy E M, 2002, Kartizky A R *et al.,* 2010, Skyner R E *et al.,* 2015, Ruelle P & Kesselring U W, 1997, Deeb O & Goodarzi M, 2010, Ran Y *et al.,* 2002). Les méthodes reportées peuvent être classées en trois types:

1 - Corrélation de la solubilité avec des données expérimentales telles que le point de fusion (MP) et log P (logarithme du coefficient de partage octanol / eau). Cependant, cette approche est de peu d'utilité, car elle exige une connaissance du point de fusion expérimental pour les composés virtuels. Le point de fusion qui est un indice clé des interactions cohérentes dans le solide est difficile à estimer.

2 - Estimation de la solubilité par des méthodes de contribution de groupes, La méthode de contribution de groupes permet le calcul approximatif de la solubilité en additionnant les valeurs fragmentales associées à des unités sous-structurelles des composés. Les inconvénients de la méthode de contribution de groupes sont les suivants: a / les groupes

inclus doivent être définis à l'avance et donc la solubilité d'un nouveau composé contenant de nouveaux groupes ne peut être estimée;

b/ les différents effets d'un groupe dans différents environnements chimiques ne sont pas considérés.

3 - La corrélation entre la solubilité et des descripteurs calculés à partir de la structure moléculaire. Cette troisième approche s'est révélée particulièrement efficace pour la prédiction de la solubilité, car elle n'a pas besoin de descripteurs expérimentaux et peut donc également être appliquée à des collections de composés virtuels.

Le but de ce travail est de développer un modèle QSPR robuste en vue de prédire la solubilité aqueuse pour un ensemble diversifié de produits agrochimiques (comprenant 26 acides, 25 urées, 13 triazines et 13 carbamates) en utilisant les descripteurs moléculaires théoriques calculés a l'aide du logiciel DRAGON (Todeschini R *et al,,* 2005).

II. 2. Résultats et discussion

II. 2. 1. Résultats du modèle RLM:

Le processus de dissolution est l'établissement d'un équilibre entre la phase du soluté et de sa solution aqueuse saturée. La solubilité aqueuse est presque exclusivement dépendante des forces intermoléculaires qui existent entre les molécules de soluté et les molécules d'eau. Les interactions: soluté- soluté, eau- soluté et eau- eau déterminent la quantité de composé dissout dans l'eau. Les interactions supplémentaires soluté- soluté sont associées à l'énergie du réseau, à l'état cristallin.

La solubilité d'un composé est donc affectée par de nombreux facteurs: l'état du soluté, le degré aromatique et aliphatique relatif des molécules, la taille et la forme des molécules, la polarité, les effets stériques et la capacité de certains groupes à participer à des liaisons hydrogènes.

Afin de prédire avec précision la solubilité, tous ces facteurs liés à la solubilité doivent être représentés numériquement par des descripteurs rattachés à la structure de la molécule.

Le modèle optimal à six descripteurs obtenu a pour équation:

log S = - 2,80 - 1,27 $\mathbf{E_{HOMO}}$ - 0,182 **Mor02v** - 17,2 **G2e** - 9,56 **HATS7v** + 4,76 **RTu+** - 0,0821 **AlogP2** (59)

R^2= 88,95 R^2_{adj}= 87,65 Q^2_{LOO} = 85,47 Q^2_{EXT} =85,11 Q^2_{BOOT} = 83,23

n= 58 n_{ext}=19s = 0,52 F = 68,42 K_{xx}= 37,68 K_{xy} = 45,67

Ici, E_{HOMO} est l'énergie de la plus haute orbitale moléculaire occupée (Clare B W, 1994 & Huang Q G *et al.,* 1996); Mor 02 v est le 3D- MoRSE- signal 02 / pondéré par le volume atomique de van der Waals (Gasteiger J *et al.,* 1996, Schuur J *et al.,* 1996); G2E est la deuxième composante de symétrie directionnelleWHIM / pondérée par l'électronégativité atomique de Sanderson (Todeschini R *et al.,* 1994, Todeschini *et al.,* 1995); HATS7v est l'effet de levier d'autocorrélation de distance topologique 7 / pondéré par le volume atomique de van der Waals (Consonni V *et al.,* 2002 , Consonni V *et al.,* 2002); RTu + est l'indice maximal R / non pondéré (Consonni V *et al.,* 2002, Consonni V *et al.,* 2002); AlogP2 est le carré du coefficient de partage octanol-eau de Ghose-Crippen-Viswanadhan (Ghose A K & Crippen G M, 1986, Viswanadhan V N *et al.,* 1993).

De plus amples informations concernant ces descripteurs peuvent être trouvées dans le guide de l'utilisateur du logiciel « Dragon » et dans les références indiquées.

La valeur de R^2indique que 88,95 (%) de variation totale est expliquée par le modèle, alors que la valeur élevée du rapport de la variance expliquée par le modèle à la variance résiduelle (F= 68,42 ; p = 0,000) montre que l'équation (59) permet une très bonne prédiction des n (=58) valeurs de log S de l'ensemble de calibrage (erreur standard s = 0,52), la valeur élevée de Q^2_{LOO} , qui diffère peu de celle deR^2, renseigne sur la robustesse du modèle.

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

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

$$y = (y_1, y_2, ...57, 58) \underline{RND}\ y_{RND} = (y_8, y_5, ...y_{27}, y_3) \quad (60)$$

et utilisées comme sources d'observations pour des modèles QSPR dans les conditions optimales établies.

La figure (15) qui représente le graphes des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés au modèle réel de départ.

Il est clair que les statistiques obtenues pour les vecteurs modifiés de log S sont plus petites que celles du modèle QSPR réel, ce qui permet d'affirmer que le modèle proposé n'est pas aléatoire.

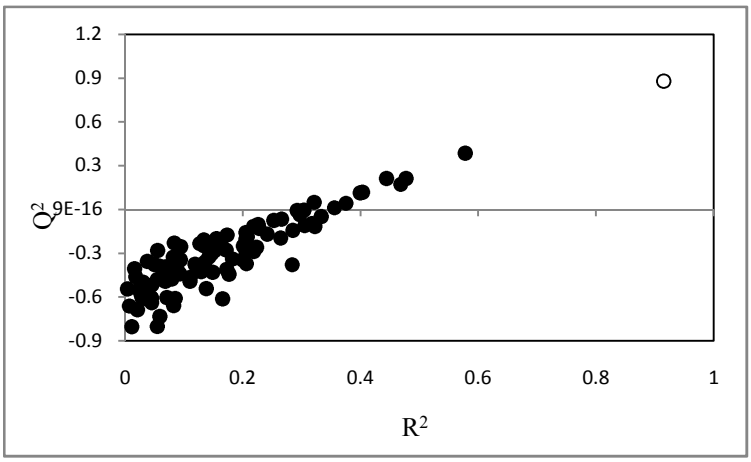


Figure 15 : Test de randomisation

Certains paramètres statistiques importants (comme indiqué dans le tableau VI) ont été utilisés pour évaluer les descripteurs concernés. La valeur de t d'un descripteur mesure la signification statistique de son coefficient de régression. Les valeurs de t absolues élevées indiquées dans le tableau VI expriment que les coefficients de régression des descripteurs impliqués dans le modèle MLR sont significativement plus grands que l'écart-type. La probabilité de t d'un descripteur donnesa signification statistique lorsqu'il est combiné avec d'autres descripteurs dans un modèle QSPR global (c-à-d: les interactions entre descripteurs). Les descripteurs avec des valeurs de la probabilité de t inférieures à 0,05 sont généralement considérés comme statistiquement significatifs dans un modèle particulier, ce qui signifie que leurs influences sur la variable réponse ne sont pas dues au hasard (Ramsey F L & Schafer D W, 2002). Les petites valeurs des probabilités de t sont liées aux descripteurs les plus importants. Les valeurs des probabilités de t des six descripteurs sont très faibles, ce qui indique que tous sont des descripteurs très significatifs. Les valeurs des VIF suggèrent que ces descripteurs sont faiblement corrélés les uns avec les autres. Ainsi, le modèle peut être considéré comme une équation de régression optimale.

Les valeurs du log S calculées à partir de l'équation (59) pour lesensembles de calibrage et de validation sont montrées dans le Tableau I (annexe) et représentées dans la Figure 16. La distribution des erreurs pour l'ensemble de données est donnée par la figure 17. Comme les erreurs sont réparties sur les deux côtés de la ligne zéro, on peut conclure qu'il n'y a pas d'erreur systématique dans le modèle développé.

Tableau VI : Caractéristiques des descripteurs sélectionnés dans le meilleur modèle MLR

Descripteur	type	x	Dx	t	probabilité-t	VIF
Constante		-2,801	2,545	-1,1	0,276	
E_{HOMO}	Descripteur quanto-chimique	-1,267	0,245	-5,18	0,000	1,1
Mor02v	Descripteur 3D- MoRSE	-0,182	0,031	-5,78	0,000	4
G2e	Indice WHIM	-17,202	4,131	-4,16	0,000	2,1
HATS7v	Descripteur GETAWAY	-9,561	1,492	-6,41	0,000	1,2
RTu+	Descripteur GETAWAY	4,762	1,909	2,49	0,000	3,2
AlogP2	Propriété moléculaire	-0,082	0,014	-5,96	0,000	2,1

Les valeurs expérimentales de la solubilité aqueuse, calculées, et prédites pour l'ensemble de validation, ainsi que les valeurs des leviers et des erreurs standardisées sont regroupées dans le tableau VII.

Tableau VII : Valeurs de log S expérimentales, calculées, prédites, leviers, et résidus standardisés de prédictions

N°	Composé	log S_{Exp}	log $S_{Calc, Pred}$	h_{ii}	e_{istd},
1	Ametryn	2,27	1,85	0,053	-0,88
2	Atrazine	1,52	1,71	0,1	0,43
3	Butylate	1,64	1,75	0,138	0,26
4	Carbetamide	3,54	3,68	0,12	0,31
5	Chorimuronethyl	3,08	2,79	0,073	-0,62
6	Clopyralide	5,16	4,95	0,153	-0,5
7	Clopyralidolamine	5,75	5,71	0,215	-0,1
8	Cyanazine	2,23	1,81	0,107	-0,97
9	2,4-D	2,95	3,44	0,192	1,31
10	2,4-DB	1,66	1,5	0,047	-0,33
11	2,4-D dimethylammonium	5,9	5,05	0,216	-2,34
12	2,4-D-methyl	2	2,76	0,046	1,56
13	2,3,6-TBA	3,89	3,86	0,204	-0,06
14	2,4,5-T	2,18	1,97	0,2	-0,57
15	Desmetryn	2,76	2,1	0,059	-1,4
16	Dichlorprop	2,54	2,55	0,095	0,02
17	Dipropetryn	1,2	0,93	0,067	-0,57
18	Diuron	1,62	1,4	0,093	-0,49
19	Fenoxapropethyl	-0,1	-0,66	0,093	-1,26
20	Fluazifopbutyl	0	-0,18	0,156	-0,46
21	FluazifopP butyl	0,3	-0,18	0,156	-1,2
22	Fluometuron	2,04	2,38	0,048	0,69
23	Fluoxypyrmeptyl	-1,05	-0,59	0,142	1,1
24	Glyphosate	4,08	4,29	0,459	1,01
25	Haloxyfop	1,64	1,81	0,103	0,4
26	Haloxyfopethoxyethyl	0,11	0,74	0,124	1,46
27	Isoproturon	1,81	2,38	0,126	1,33
28	Lenacil	0,78	1,61	0,113	1,92
29	Linuron	1,88	2,47	0,039	1,22
30	MCPA	2,87	3,03	0,049	0,34
31	Mecoprop	2,87	2,63	0,118	-0,54
32	MecopropP	2,93	2,63	0,118	-0,69
33	Metamitron	3,23	2,72	0,15	-1,25
34	Methabenzthiazuron	1,77	2,56	0,044	1,63
35	Metsulfuronmethyl	3,98	3,53	0,17	-1,14
36	Pebulate	2	1,78	0,06	-0,47
37	Phenmedipham	0,67	1,41	0,144	1,8
38	Picloram	2,63	3,27	0,264	1,93
39	Prometon	2,86	1,94	0,065	-1,95
40	Prometryn	1,52	1,49	0,054	-0,06

N°	Composé	log S_{Exp}	log $S_{Calc, Pred}$	h_{ii}	e_{istd},
41	Propazine	0,93	1,22	0,084	0,63
42	Prosulocarb	1,12	0,9	0,038	-0,45
43	Prosulfuron	3,6	2,91	0,156	-1,72
44	Quizalofopethyl	-0,51	-0,14	0,097	0,83
45	Rimsulfuron	3,86	3,32	0,145	-1,31
46	Tebuthiuron	3,4	3,68	0,143	0,68
47	Terbutryn	1,34	1,75	0,077	0,88
48	Thifensulfuronmethyl	3,8	4,13	0,107	0,76
49	Thiobencarb	1,45	1,26	0,053	-0,4
50	Triallate	0,6	0,68	0,112	0,19
51	Triasulfuron	2,91	2,4	0,089	-1,14
52	Triclopyr	3,91	3,17	0,198	-1,97
53	Triclopyrbutotyl	1,36	0,74	0,073	-1,33
54	Triflusulfuronmethyl	2,04	2,19	0,058	0,31
55	Vinclozolin	0,53	1,47	0,129	2,23
56	Propaquizafop	-0,2	-0,11	0,135	0,22
57	Ethametsulfuronmethyl	1,7	2,81	0,048	2,29
58	Desmedipham	0,95	0,27	0,284 *	-2,17
59	Bensulfuronmethyl*	2,08	2,86	0,057	1,55
60	Bromacil*	2,85	3,55	0,077	1,41
61	Chloroxuron*	0,4	0,85	0,112	0,92
62	Chlorpropham*	1,95	1,88	0,074	-0,14
63	Chlorsulfuron*	4,45	3,48	0,112	-1,96
64	Cycloate*	1,98	1,8	0,059	-0,35
65	DichlorpropP*	2,77	2,55	0,094	-0,44
66	Difenxuron*	1,3	1,14	0,233	-0,36
67	EPTC*	2,54	2,35	0,054	-0,38
68	MCPB*	1,64	1,45	0,064	-0,39
69	Metoxuron*	2,83	2,1	0,171	-1,54
70	Metribuzin*	3,09	2,33	0,116	-1,55
71	Napropemide*	1,87	1,26	0,079	-1,22
72	Primisulfuronmethyl*	2,39	2,93	0,154	1,14
73	Simazine*	0,79	1,94	0,106	2,34
74	Terbacil*	2,85	3,48	0,091	1,28
75	Terbutylazine*	0,93	0,81	0,156	-0,24
76	Tribenuronmethyl*	3,18	3,44	0,105	0,53
77	Vernolate*	2,03	2,33	0,045	0,59

* Composés de validation

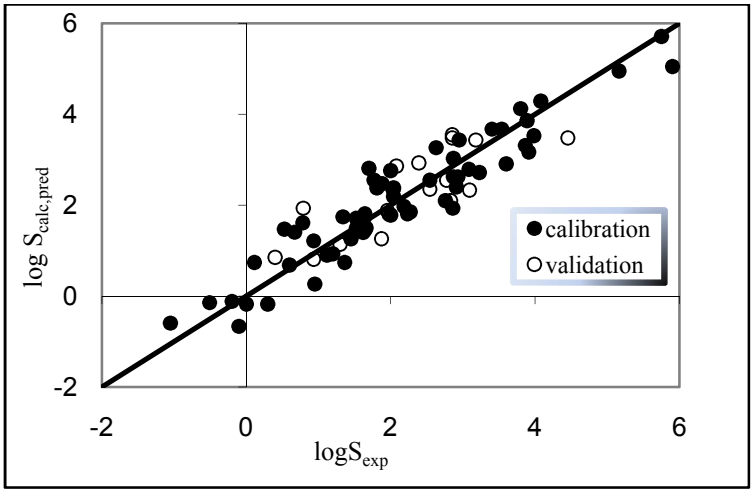


Figure 16 : Graphe des valeurs log Scalculées en fonction des valeurs observées

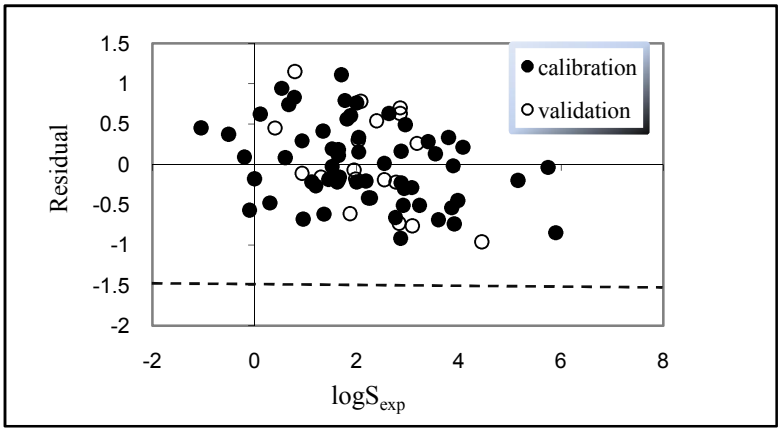


Figure 17 : Graphe des résidus enfonctiondes valeursexpérimentales de logS

II. 2. 2.Contribution des descripteurs et interprétation:

En se basant sur une procédure décrite dans la littérature (Zheng F *et al.,* 2006, Guha R & Jurs P C, 2005), les contributions relatives des six descripteurs du modèle ont été déterminées.Elles diminuent selon l'ordre suivant: HATS7v (17,91%)> Mor2v (17,67%) > HOMO (16,94%)> AlogP2 (16,80%)> G2E (15,76%)> RTu + (14,89%). Il convient de noter que la différence de contribution entre deux descripteurs utilisés dans le modèle n'est pas significative, ce qui indique que tous les descripteurs sont indispensables pour générer le modèle prédictif.

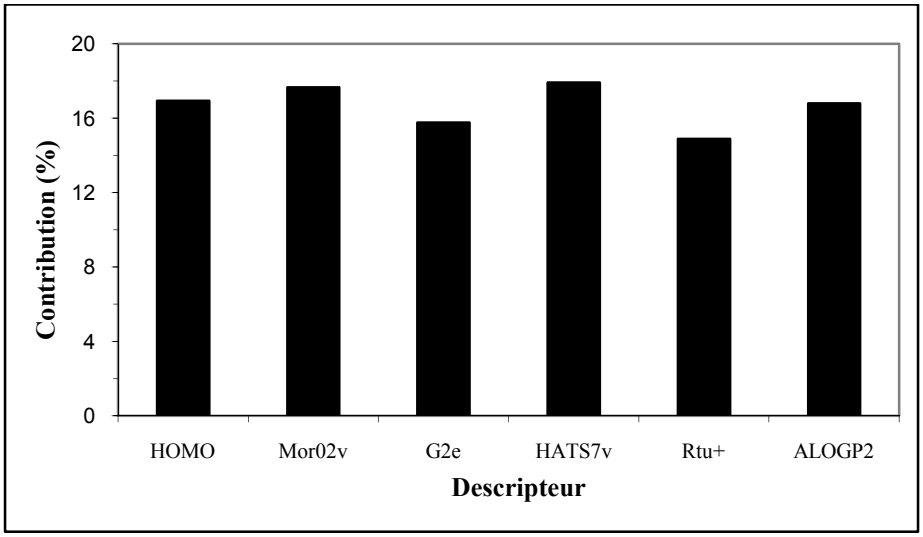


Figure 18 : Contributions des descripteurs dumodel MLR

L'importance du volume atomique de van der Waals sur les valeurs de log S est évidente, étant donné que les descripteurs pondérés par ce volume expliquent 35,58% des contributions (17,91% des HATS7v et 17,67% des Mor2v). Le premier descripteur important est HATS7v, qui a une corrélation négative relativement élevée avec les valeurs expérimentales de log S (R = -0,328). Le coefficient négatif de HATS7v indique que les produits agrochimiques avec des valeurs élevées pour ce descripteur auraient des valeurs de log S petites.

Le deuxième descripteur important est Mor02v, un descripteur 3D- MoRSE, qui a un coefficient de corrélation négatif plus petite avec les valeurs de log S expérimentales (R = -0,787). Les descripteurs 3D- MORSE sont les représentations moléculaires 3D de la structure à base de descripteur de diffraction électronique (Gasteiger J *et al.,*1996, Schuur J *et al.,* 1996), qui sont calculés en additionnant les poids atomiques vues par des fonctions de diffractions angulaires différentes. Les descripteurs 3D- MORSE sont calculés à partir de l'expression suivante:

$$Morsw = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i w_j (\sin(s.r_{ij}) / s.r_{ij}) \qquad (61)$$

où s est l'angle de diffraction, n_{AT} le nombre d'atomes, r_{ij} la distance interatomique entre les $i^{ème}$ et $j^{ème}$ atomes, w est une propriété atomique, pouvant être le numéro atomique, la masse, le volume de van der Waals, l'électronégativité de Sanderson ou la polarisabilité. Le

coefficient de Mor02v est négatif, ce qui indique qu'une augmentation de Mor02v se traduirait par une diminution des valeurs de log S.

Par conséquent, comme prévu, les volumes atomiques ont un effet spécifique sur les valeurs de log S: une augmentation de Mor02v (ou HATS7v) se traduirait par une diminution des valeurs de log S.

Le coefficient de partage octanol-eau de Ghose- Gippen-Viswanadhan (AlogP) (Ghose A K & Crippen G M, 1986, Viswanadhan V N *et al.,* 1993) est calculé à partir d'une équation de régression basée sur le caractère hydrophobe de la molécule. Il reflète à la fois les interactions du soluté avec le solvant l'entourant (effets macroscopiques ou non spécifiques du solvant) et la liaison spécifique entre le soluté et les molécules individuelles du solvant (effets microscopiques ou spécifiques du solvant). Lorsque le carré de ce descripteur augmente (AlogP2), log S diminue.

L'énergie de la plus haute orbitale moléculaire occupée (E_{HOMO}) est une mesure de la nucléophilie d'une molécule. Il faut expliquer les différences dans la tendance des solutés à prendre part dans les interactions de transfert de charge, c'est-à-dire la capacité des molécules de soluté à être donneuses d'électrons aux molécules d'eau. D'après le théorème de KOOPMANS (Guha R& Jurs P C, 2005), l'énergie HOMO est directement liée au potentiel d'ionisation IP ($-E_{HOMO}$ = IP), à condition que le processus d'ionisation soit représenté de façon adéquate par l'élimination d'un électron à partir d'une orbitale sans modification des fonctions d'ondes des autres électrons. Le descripteur et son coefficient dans le modèle sont négatifs, de sorte que la contribution de E_{HOMO} est positive.

L'importance de la forme axiale et la symétrie de la molécule sur les valeurs de logS est évidente en raison de la présence de G2e. Dans les calculs, les électronégativités atomiques de Sanderson pour chaque atome ont été utilisées, car elles peuvent déterminer, avec d'autres propriétés atomiques, les propriétés macroscopiques d'un composé. Le signe positif de G2e signifie que l'augmentation de ce descripteur diminue le log S.

RTu +, comme HATS7v, est un descripteur GETAWAY qui est en corrélation assez étroite avec les valeurs de log S expérimentales (r = 0,490). Les descripteurs ~~de~~ GETAWAY (Consonni V *et al.,*2002, Consonni V *et al.,*2002) ont été proposés comme descripteurs de structure chimique dérivés d'une nouvelle représentation de la structure moléculaire, la

matrice d'influence moléculaire. Ces descripteurs, basés sur l'autocorrélation spatiale, codent des informations sur l'espace moléculaire. Dans une certaine mesure, ils tiennent compte aussi des informations sur la taille et la forme moléculaire ainsi que des propriétés atomiques spécifiques.

HATS7v et RTu + sont calculés par les équations, (62) et (63), respectivement,

$$HATSk(w) = \sum_{i=1}^{A}\sum_{j^3i}^{A}(w_i.h_i)(w_j.h_j)\delta(d_{ij};k) \text{ pour } \quad k = 0,1,2,3, ,,, D \qquad (62)$$

$$RTu+ = \max_{ij}\left(\frac{\sqrt{h_i h_j}}{r_{ij}}.w_i.w_j.\delta(d_{ij};k)\right) k = 0, 1, 2,3, ,,, D \qquad (63)$$

Où A est le nombre d'atomes, w est un schéma de pondération atomique, d_{ij} est la distance topologique, δ (k, d_{ij}) est la fonction delta de Dirac (δ = 1 si d_{ij} = k, zéro autrement), r_{ij} est la distance interatomique, D est le diamètre topologique de la molécule qui est la distance topologique maximale dans la molécule.

Le coefficient de RTu + est positif, ce qui signifie que les pesticides avec des valeurs plus élevées pour ce descripteur auraient de plus grandes valeurs de log S.

II. 2. 3.Domaine d'application du modèle RLM:

Avant qu'un modèle QSPR ne soit mis en service pour le criblage de composés, son domaine d'application doit être défini, pour que les prédictions des composés qui tombent dans ce domaine puissent être considérés comme fiables.

Le domaine d'application du modèle MLR a été analysé dans le cadre du diagramme de Williams (représenté en Fig.19). Il est clair que l'observation 35 de l'ensemble de calibrage avec un effet de levier supérieur à la valeur critique h*= 0,36 est un composé structurellement influent. La suppression de l'observation 35modifie légèrement R^2 entre les valeurs expérimentales de log S et les descripteurs sélectionnés à 0,8866 (Q^2= 0,8485) et augmente l'erreur standard à 0,524, tandis que l'utilisation d'une grande énergie issue d'une nouvelle géométrie de conformation pour cette observation modifie négativement le modèle calculé.

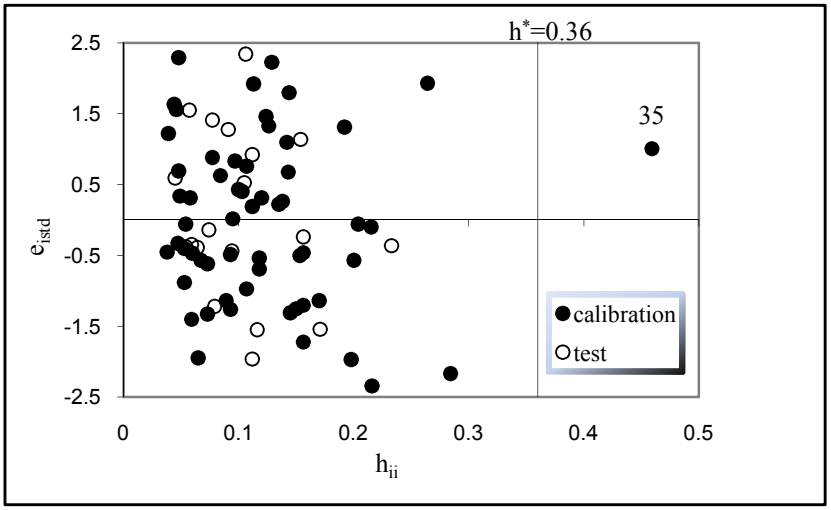


Figure 19 : Diagramme de Williams

II. 3. Conclusion:

Dans cette étude, le procédé QSPR a été appliqué à la prédiction de la solubilité aqueuse de différents types de pesticides. Un modèle linéaire à six descripteurs a été développé par l'approche hybride GA / MLR avec des valeurs de R^2=88,95 et s = 0,52 unité de log pour l'ensemble de calibrage. Les descripteurs sélectionnés expriment de nombreux facteurs qui influent sur la solubilité aqueuse: la taille et la forme moléculaires, les propriétés atomiques spécifiques, des effets macroscopiques et microscopiques et la tendance des solutés à prendre part à des interactions de transfert de charge. Plusieurs techniques de validation, y compris la validation croisée par leave-one-out et bootstrap, test de randomisation et validation externe ont illustré la fiabilité du modèle proposé. Tous les descripteurs peuvent être directement calculés à partir de la structure moléculaire du composé, ainsi le modèle proposé est prédictif et pourrait être utilisé pour estimer la solubilité des pesticides. Dans ce cas, le domaine d'application sera un outil précieux pour filtrer les structures chimiques "dissemblables".

III. Modélisation du coefficient de partage octanol/eau

III. 1. Introduction

La détection de certains pesticides, principalement des insecticides, dans les sources d'eau a indiqué un manque de compréhension de leurs propriétés et a été le stimulus pour de nombreuses études sur leur sort et leur transport (Deeb et Goudarzi, 2010).

L'impact du danger potentiel des produits chimiques non testés, un défi auquel sont confrontés les organismes de réglementation nationaux et internationaux, (Zeeman *et al.,* 1995; Walker, 2003; Bradbury *et al.,* 2003; Commission européenne, 2001) peut être mesuré par des études expérimentales, mais cette approche est à la fois très coûteuse et prend beaucoup de temps (Toussaint *et al.,* 1995).

Les propriétés physicochimiques d'un composé chimique organique jouent un rôle important dans la détermination de sa distribution et de son devenir dans l'environnement. La pression de vapeur (PV), la solubilité aqueuse (S) et le coefficient de partage n-octanol / l'eau (K_{ow}) sont des propriétés physicochimiques clés qui peuvent être utilisées pour l'évaluation de la partition dans l'environnement et le transport de substances organiques (Xu *et al.,* 2008).

Le coefficient de partage n-octanol / eau est le rapport de la concentration d'un produit chimique dans le n-octanol à celle dans l'eau du système à deux phases à l'équilibre. Le logarithme de ce coefficient de partage, log P, est le paramètre qui détermine la lipophilie d'une molécule, et il a trouvé une large application dans la prédiction des activités biologiques et effets toxicologiques (Katritzky *et al.,* 2010).

Le but est de trouver un modèle statistique pour la prédiction du coefficient n-octanol / eau (Kow) de composés organophosphorés. A cet effet, la relation entre les descripteurs moléculaires (Todeshini *et al.,* 2006) reliés aux facteurs expérimentalement constatés comme affectant le K_{ow} des composés a été établie. Le modèle QSPR a été construit en utilisant les méthodes de la régression linéaire multiple (MLR), et les machines à vecteurs support (SVM), et sa performance validée. Le modèle obtenu montre quels descripteurs jouent un rôle important dans la variation K_{ow} de ces pesticides.

III. 2. Résultats et discussion

III. 2. 1. La régression linéaire multiple

Les 43 pesticides d'intérêt (Tableau I- Annex) sont des produits chimiques organophosphorés très polluants.Les valeurs mesurées de leur Kow respectifs prélevées dans la littérature (Hansen O C, 2004) variant de plusieurs ordres de grandeurs (de 10^{-1} à 9, 10^{5}, environ) nous exploitterons les valeurs de leurs logarithmes pour diminuer l'intervalle de variation qui s'étend alors sur [-0,89 ; 5,96].

L'application de la procédure GA-VSS conduit à plusieurs modèles pour la prédiction de log K_{ow} de ces produits chimiques organophosphorés en fonction de différents ensembles de descripteurs moléculaires.

Pour s'assurer que la qualité du modèle n'est ni surestimée ni sous-estimée, 10 choix aléatoires ont été effectuées et les statistiques de chaque modèle ont été calculées. Les 10 groupes de validation ont ainsi été sélectionnés de manière aléatoire avec un nombre constant de composés pour l'ensemble de calibrage et l'ensemble de validation (34 molécules de calibrage et 9 composés de validation). L'analyse MLR a été effectuée sur chaque groupe de calibrage alors que la prédictivité de chaque modèle MLR a été évaluée en utilisant le groupe correspondant de validation. Le tableau XII donne les paramètres obtenus pour les analyses de MLR pour les différentes divisions réalisées.

La séparation de la base de données offrant le comportement médian, à savoir les paramètres statistiques les plus proches de la valeur moyenne a été choisie. Ainsi, le modèle à développer correspond au modèle N°6.

La performance du modèle est décrite aux moyens des paramètres liés à la capacité prédictive du modèle (Q^2_{LOO}, Q^2_{LMO}) et la capacité d'ajustement (R^2). Les déviations standards des erreurs de prédiction (SDEP) et de calcul (SDEC) sont également rapportées.

Tableau VIII: Les statistiques des 10 modèles

	R^2	R^2_{cv}	R^2_{ext}
1	85,45	81,10	90,45
2	86,20	82,94	94,83
3	86,60	83,37	90,56
4	86,66	83,08	95,44
5	87,35	84,20	90,79
6	87,87	85,35	85,08
7	88,44	85,46	85,97
8	88,87	86,20	88,26
9	89,72	87,32	89,06
10	90,02	87,79	84,14
$<R^2>^{(a)}$	87,71	84,68	89,45

(a) Valeur moyenne

Le meilleur modèle obtenu en utilisant 34 organophosphorés de calibrage [log Kow: (-0,89; 5,96)] est un modèle à trois dimensions avec une capacité prédictive élevée:

$$\mathbf{logK_{ow}} = -0,649 + 0,869\ \mathbf{GATS7m} - 0,0456\ \mathbf{TPSA\ (NO)} + 0,196\mathbf{polarizability} \quad (64)$$

n_{tr}= 34 R^2= 87,87 % Q^2_{LOO}= 85,35 % Q^2_{BOOT}= 82,97 % Q^2_{LMO}= 83,95 %

n_{test}= 9 Q^2_{ext}= 85,08 % SDEP= 0,689 SDEC= 0,627 $SDEP_{ext}$= 0,696

K_{XX}= 31,42 K_{XY}= 52,00 s= 0,668 F= 72,41

Les paramètres d'ajustement et de validation rapportés ont, comme prévu, des valeurs élevées indiquant que le modèle a une très bonne performance prédictive et que les descripteurs impliqués décrivent très bien le coefficient de partage K_{ow}.

Les valeurs absolues de t élevées indiquées dans le tableau IX expriment que les coefficients de régression des descripteurs impliqués dans le modèle MLR sont significativement plus grand que l'écart-type.Les valeurs de la probabilité de t pour chaque descripteur sont très faibles, ce qui indique que chacun des descripteurs est très significatif. Les modèles ne seront pas acceptés s'ils incluent des descripteurs avec des VIF> 5 (Holder *et al.,* 2003). La matrice de corrélation, comme indiqué dans le tableau X, suggère que ces descripteurs sont faiblement corrélés entre eux. Ainsi, le modèle peut être considéré comme une équation de régression optimale.

Tableau IX : Caractéristique des descripteurs sélectionnés pour le modèle MLR

Descripteur	x	Dx	t	Probabilité-t	VIF
Constante	-0,649	0,6409	-1,01	0,319	
GATS7m	0,8693	0,2780	3,13	0,004	1,6
TPSA(NO)	-0,0456	0,0068	-6,67	0	1,2
Polarisability	0,196	0,0220	8,87	0	1,5

Tableau X : Matrice de corrélation

	log Kow	GATS7m	TPSA(NO)
GATS7m	0,474 0,005		
TPSA(NO)	-0,404 0,018	0,311 0,074	
Polarisability	0,835 0,000	0,529 0,001	-0,032 0,856

Les valeurs de log Kow expérimentales, calculées, et prédites pour l'ensemble de validation, ainsi que les valeurs des leviers et des erreurs standardisées sont regroupées dans le tableau XI.

Tableau XI : Valeurs de log K_{ow} expérimentales, calculées, prédites, leviers et résidus standardisés de prédictions

N°	Composé	log Kow_{Exp},	log $Kow_{Calc, Pred}$	h_{ii}	e_{istd},
1	Naled	1,38	1,24	0,083	-0,24
2	Trichlorfon	0,51	0,09	0,115	-0,75
3	Thiometon	3,46	3,47	0,13	0,03
4	Phorate	3,56	3,4	0,138	-0,29
5	Disulfoton	3,95	4,71	0,187	1,55
6	Terbufos	4,48	4,12	0,166	-0,7
7	Methamidophos	-0,8	-1	0,209	-0,43
8	Dichlorvos	1,16	0,18	0,13	-1,81
9	Oxydometonmethyl	-0,74	0,47	0,094	2,1
10	Ethion	5,07	5,41	0,105	0,6
11	Dimethoate	0,7	1,05	0,086	0,61
12	Mevinphos	0,13	-0,28	0,133	-0,76
13	Malathion	2,75	2,44	0,093	-0,55
14	Phosphamidon	0,79	2	0,052	1,96
15	Propetamphos	3	3,24	0,125	0,44
16	Formothion	-0,56	1,31	0,058	3,06
17	Fonofos	3,94	3,93	0,188	-0,03
18	Fenthion	4,09	4,13	0,098	0,07
19	Sulprofos	5,48	5,23	0,17	-0,49
20	Cyanophos	2,65	2,47	0,154	-0,34
21	Chlorpyrifos	4,7	4,2	0,078	-0,85
22	Etrimfos	3,3	2,76	0,067	-0,9
23	Diazinon	3,74	3,38	0,046	-0,58
24	Pirimiphosmethyl	4,2	3,87	0,121	-0,61
25	Pirimiphosethyl	4,85	4,37	0,082	-0,82
26	Fenitrothion	3,43	2,26	0,088	-2,01
27	Phoxim	3,38	2,87	0,054	-0,83
28	Methidathion	2,2	2,4	0,058	0,32
29	Temephos	5,96	5,84	0,321	-0,32
30	Phosalone	4,3	3,81	0,122	-0,89
31	Azamethiphos	1,05	0,8	0,218	-0,55
32	Phosmet	2,78	3,23	0,044	0,72
33	Azinephosmethyl	2,56	3,15	0,083	1
34	Azinephoseyhyl	3,18	4,09	0,107	1,61
35	Ethoprophos*	3,59	2,66	0,103	-1,46
36	Acéphate*	-0,89	-0,41	0,14	0,78
37	Dicrotophos*	-0,49	0,46	0,091	1,5
38	Isazofos*	3,82	3,14	0,057	-1,05
39	Profenofos*	4,44	4,74	0,143	0,49

N°	Composé	log $Kow_{Exp,}$	log $Kow_{Calc, Pred}$	h_{ii}	$e_{istd,}$
40	Chlorpyrifosmethyl*	4,24	3,59	0,13	-1,05
41	Tetrachlorvinphos*	3,53	3,56	0,036	0,04
42	Chlorfenvinphos*	3,95	3,89	0,042	-0,1
43	Isofenphos*	4,12	5,29	0,118	1,86

* Composés de validation

La figure (20) reproduit, pour la totalité des données, les valeurs calculées de log Kow en fonction des valeurs observées.

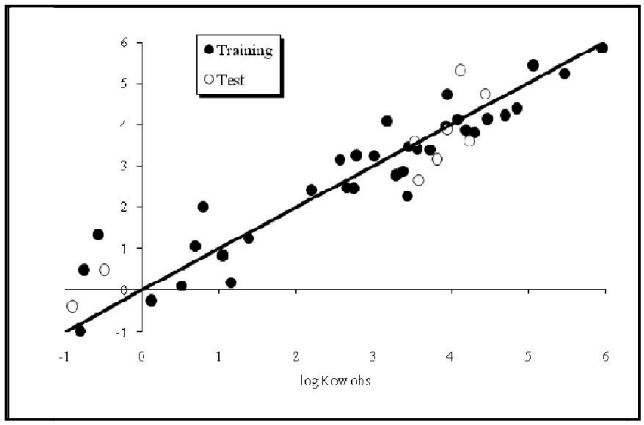


Figure 20 : Graphe des valeurs calculées de log K_{ow} en fonction des valeurs observées

La dispersion des points est acceptable, bien qu'il y ait un point un peu éloigné du reste (Formothion). Ce composé chimique dont le réside standardisé (e_{istd}) est supérieur à 3 unités d'écart- type est un point aberrant. C'est ce qui ressort du diagramme de Williams de la figure 21 qui montre que le formothion est un objet aberrant pour la détermination des paramètres du modèle. Tous les points présentent un effet de levier inférieur à la valeur critique (h*= 0,35) représentée par la ligne droite verticale.

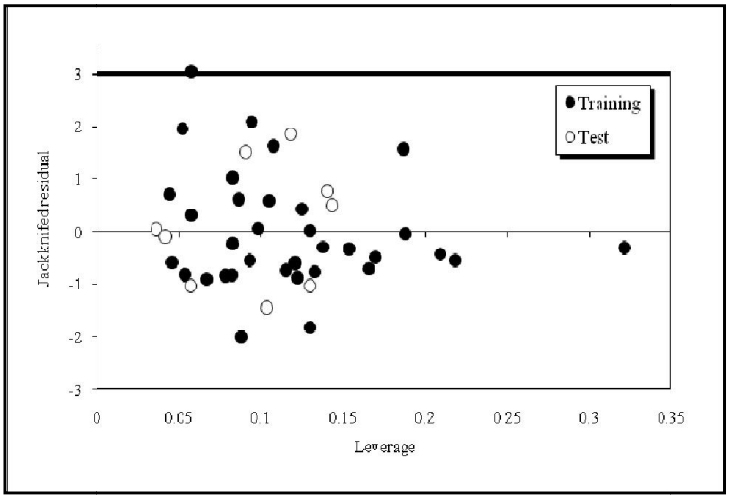


Figure 21 : Diagramme deWilliams

Toutes les erreurs étant réparties des deux côtés de la ligne $e_{istd} = 0$, on peut conclure qu'il n'y a pas d'erreur systématique dans le développement du modèle.

La figure 22 qui représente le diagramme des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (cercles noirs) avec le modèle de départ (cercle vide). Il est clair que les statistiques obtenues pour les vecteurs modifiés du log K_{ow} sont inférieures à ceux du modèle réel; les Q^2 sont inférieurs à 20, et pour la majeure partie, on obtient même des $Q^2<0$. Ceci permet d'assurer que le modèle établi a une base réelle et n'est pas dû au hasard.

L'utilité du rapport de la concentration d'un soluté entre l'eau et le n-octanol en tant que modèle pour son transport entre les phases dans un système physique ou biologique est reconnu depuis longtemps (Leo *et al.,* 1971, Leo, 1981). Il est exprimé en $P_{oct} = C_o / C_w = K_{ow}$ ce rapport est essentiellement indépendant de la concentration.

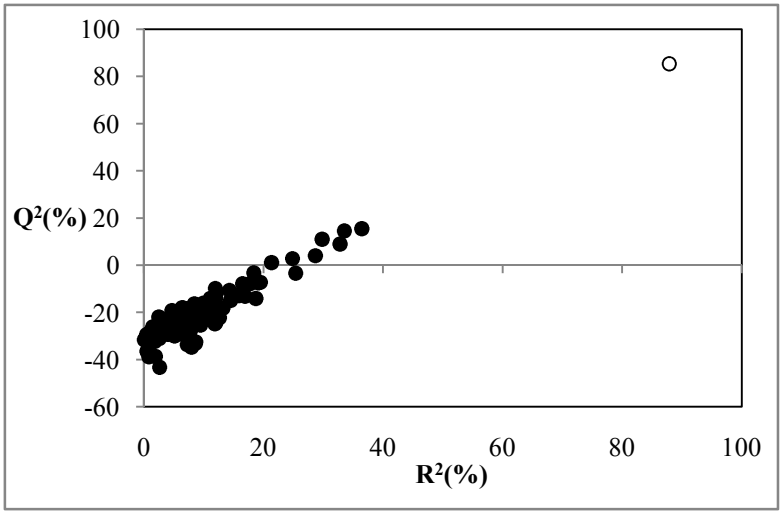


Figure 22 : Test de randomisation

Les plus clairs aperçus sur les forces de solvatation qui déterminent les coefficients de partage proviennent de travaux initiés par Kamlet, Taft et leurs collègues, ce qui est désigné par l'approche solvatochromique (Kamlet *et al.,* 1988, Abraham *et al.,* 1994). Pour le coefficient de partage octanol / eau, les paramètres importants du soluté sont: la taille (V), polarité / polarisabilité (Π), et l'intensité d'accepteur de liaison hydrogène(β), Ceci est exprimé comme suit:

$$\log P_{oct} = aV + b\Pi + c\beta + d \quad (65)$$

La valeur du coefficient de polarisabilité (+ 0,196) dans l'équation (64), indique la contribution positive de ce descripteur à la valeur de K_{ow}.

Aucune méthode entièrement satisfaisante n'existe à ce jour pour calculer le paramètre de polarité / polarisabilité, Π, appliqué à l'équilibre du soluté entre l'eau et l'octanol. L'excès de réfractivité molaire du soluté (par rapport à un alcane de taille égale) peut être estimé séparément de la polarisabilité / dipolarité (Leo, 1981) et semble une approche intéressante à ce problème, mais il a besoin d'une vérification plus poussée. Le moment dipolaire de la molécule entière a été utilisé en tant que paramètre de polarité (Bodor et Huang, 1992), mais il y a de bonnes raisons de croire qu'il a au mieux une valeur marginale. Le carré du moment dipolaire a également été utilisé (Leahy *et al.,* 1992), et ceci, au moins, a une certaine base théorique (Kirkwood, 1934).

III. 2. 2. Machine à vecteur support

Après la mise en place du modèle MLR, une régression SVM a été utilisée pour développer un modèle sur les composés de l'ensemble de calibrage, sur la base du même sous-ensemble de descripteurs.

Dans notre travail, le modèle SVM utilise la fonction de base radiale (RBF). Avec une procédure de réglage fin, nous avons essayé d'obtenir la plus faible racine de l'erreur quadratique moyenne (RMSE) liée au meilleur paramètre de régression en utilisant le leave-one- out (LOO) en tenant compte du RMSE de l'ensemble de test. Les valeurs optimales obtenues pour les paramètres SVM et les résultats de la régression sont présentés dans le tableau XII.

Tableau XII : Paramètres et résultats du modèle SVM

C	γ	E	R^2	Q^2_{loo}	Q^2_{ext}	RMSE	$RMSE_{ext}$
10	0,5	0,2	91,06%	87,47%	92,09%	0,563	0,564

Les log K_{ow} observés et prédits de l'ensemble de calibrage et de l'ensemble de validation sont présentés dans la figure 23. Les valeurs calculées sont, en général, en bon accord avec les valeurs expérimentales correspondantes.

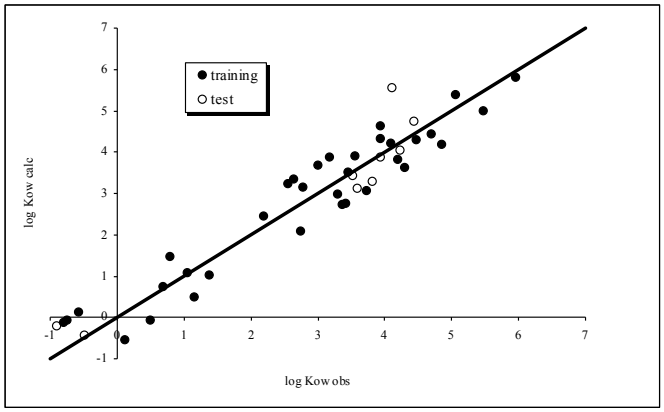


Figure 23 : Graphe des valeurs calculées, prédites en fonction des valeursexpérimentales

Le tableau XIII condense les résultats des modèles linéaire (MLR) et non linéaire (SVM). Ce dernier semble plus puissant du fait d'une qualité statistique plus élevée en sus d'une faible erreur de prédiction.

Tableau XIII : Comparaison entre les résultats des modèles MLR et SVM

Méthode	Calibrage n= 34				Validation n= 9					
	R^2	Q^2	Q^2_{ext}	RMSE	R^2_{test}	Q^2_{ext}	$(R^2-R^2_0)/ R^2<0,1$	$(R^2-R'^2_0)/ R^2<0,1$	0,85 < k< 1,15	0,85< k'< 1,15
MLR	87,87%	85,35%	85,08%	0,627	87,57%	87,40%	-0,1417	-0,1381	0,9928	0,9680
SVM	91,06%	87,47%	92,06%	0,563	91,23%	90,78%	-0,0907	-0,0957	0,962	1,0114

Les paramètres statistiques obtenus pour l'ensemble de test (Wang et al,, 2003), démontrent la puissance de la prédictivité des modèles.

III. 3. Conclusion:

Deux méthodes MLR et SVM ont été utilisées pour prédire le coefficient de partage octanol / eau (Kow) de pesticides phosphorés. Le modèle OLS a été développé par une sélection par algorithme génétique de descripteurs moléculaires théoriques parmi un large éventail obtenu avec plusieurs logiciels. Les données ont été séparées au hasard en deux sous-ensembles de 34 éléments pour le calibrage et 9 pour la validation externe.

Les modèles proposés (MLR et SVM) sont stables, robustes et prédictifs.La meilleure approche QSPR a été basée sur la méthode SVM. Le domaine d'applicabilité chimique du modèle MLR étudié et la fiabilité des prédictions ont été vérifiés par l'approche des leviers.

Les descripteurs moléculaires sélectionnés ont un sens mécanistique clair: ils sont liés à la taille des molécules et à leurs caractéristiques électroniques, ainsi qu'à la capacité de la substance chimique à former des liaisons hydrogènes avec l'eau.

IV. Modélisation de la pression de vapeur

IV. 1. Introduction

L'utilisation de quantités de plus en plus importantes de pesticides, en particulier dans l'agriculture, a conduit de nombreux chercheurs et de nombreux gestionnaires de la qualité de l'environnement à se poser des questions sur l'impact que ces produits pourraient avoir sur la qualité des eaux de surface (Khan, M A Q, 1977).

En outre, la connaissance de la pression de vapeur est nécessaire pour la prédiction del'évolution des polluants dans l'environnement.

Les méthodes QSPR sont souvent utilisées pour estimer les propriétés physico-chimiques des composés organiques et prédire leur comportement dans l'environnement. Des méthodes chimiométriques peuvent être utilisées pour décrire la manière dont les propriétés physicochimiques varient en fonction des caractéristiques de la structure moléculaire exprimées en termes de descripteurs moléculaires appropriés. Les modèles QSPR peuvent également donner un aperçu général de la structure moléculaire qui influe sur ces propriétés. Cette technique statistique est souvent utilisée pour remplacer les tests biologiques coûteux ou des expériences d'une propriété physico-chimique donnée avec des descripteurs calculés, et peut également être utilisée pour prédire les réponses d'intérêt pour de nouveaux composés (Kubinyi H, 1993). Les résultats de ces modèles révèlent des différences substantielles dans les domaines d'application et dans les capacités de prédiction (Mackay D *et al.,* 2000, Dearden J C & Schüürmann G, 2003, Estrada E *et al.,* 2004).

IV. 2. Résultats et discussion :

IV. 2. 1. Régression linéaire multiple

La régression linéaire multiple (MLR) est la méthode de modélisation la plus utilisée dans les études QSPR.La sélection par algorithme génétique conduit à un bon modèle MLR à six descripteurs qui décrit au mieux la pression de vapeur. Le modèle retenu a pour équation :

$$\begin{aligned}\textbf{log Pv} = &- 7{,}71(\pm 0{,}718) + 0{,}0637\ (\pm 0{,}0106)\ \textbf{TIE} - 1{,}73(\pm 0{,}4378)\ \textbf{EEig10d} - 1{,}60\ (\pm 0{,}4182)\\ &\textbf{EEig12d} - 0{,}347\ (\pm 0{,}093)\ \textbf{RDF060v} - 7{,}42\ (\pm 3{,}042)\ \textbf{HATS6v} +\\ &0{,}128\ (\pm 0{,}020)\ \textbf{ALOGP2} \qquad (66)\end{aligned}$$

R^2= 81,64 Q^2_{LOO} = 75,10 Q^2_{LMO} (20%) = 73,14 Q^2_{BOOT} = 65,12

EQMP = 1,313 EQMC= 1,127 K_{xx}= 47,99 K_{xy} = 50,50

n= 53 S= 1,21 F= 34,08

n_{ext}= 22 Q^2_{ext} = 79,33 $EQMP_{ext}$ = 1,221

La matrice de corrélation est reproduite ci- après:

Tableau XIV : Matrice de corrélation

	log Pv	TIE	EEig10d	EEig12d	RDF060v	HATS6v
TIE	-0,368 0,007					
EEig10d	-0,670 0,000	0,819 0,000				
EEig12d	-0,677 0,000	0,775 0,000	0,873 0,000			
RDF060v	-0635 0,000	0,660 0,000	0,734 0,000	0,651 0,000		
HATS6v	0,059 0,674	-0,135 0,337	-0,219 0,116	-0,283 0,040	-0,081 0,807	
ALOGP2	0,232 0,094	0,220 0,114	-0,219 0,115	0,135 0,336	0,151 0,280	0,002 0,986

Laisse apparaître de fortes corrélations entres plusieurs descripteurs d'ailleurs confirmées par une faible différence ΔK (Kxy – Kxx), inférieur à la valeur limite ΔK= 5.

La qualité de l'ajustement à été vérifiée par la présentation des valeurs calculées du logarithme de la pression de vapeur pour l'ensemble de calibrage, et celles prédites pour l'ensemble de validation en fonction de celles expérimentales. La figure 24 représentelog $Pv_{cal-pred}$en fonction de log Pv_{exp}.

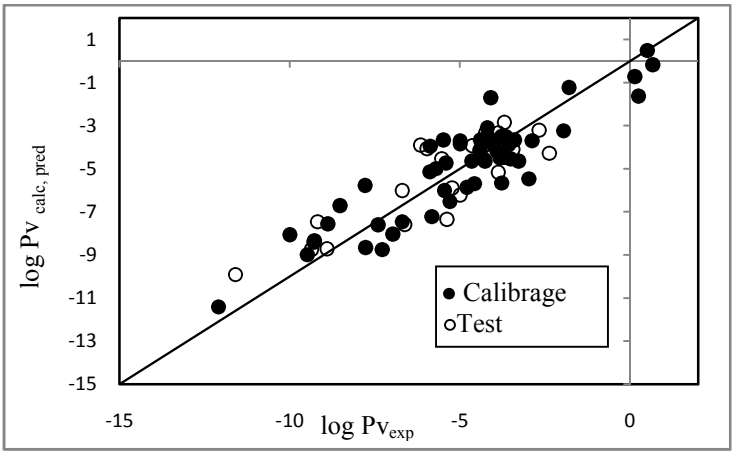


Figure 24 : Graphe des valeurs calculées, prédites en fonction des valeurs expérimentales

Le diagramme de Williams représenté dans la figure 25 permet d'afficher les valeurs des résidus de prédiction standardisés en fonction de leviers (h_{ii}), pour les deux ensembles (calibrage et validation).

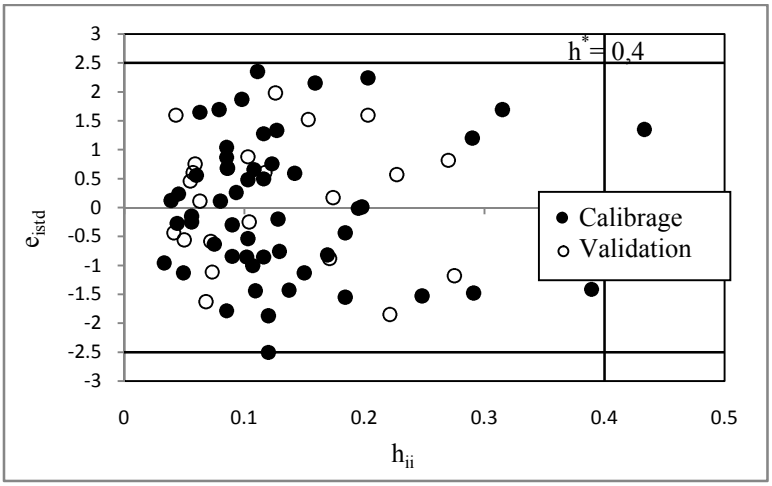


Figure 25: Diagramme de Williams

Tous les résidus standardisés de prédiction sont compris entre les limites ± 2,5 à l'exception d'un seul composé de l'ensemble de calibrage, un autre composé de l'ensemble de calibrage présente un bras de levier important ($h_i > h^* = 0,4$).

Les valeurs de la constante de Henry expérimentales, calculées, et prédites pour l'ensemble de validation, ainsi que les valeurs des leviers et des erreurs standardisées sont regroupées dans le tableau XV.

Tableau XV : Valeurs de log Pv expérimentales, calculée, prédites, h_{ii}, et e_{istd}

N°	Composé	log Pv $_{Calc, Pred}$	log Pv $_{Exp}$	h_{ii}	e_{istd}
1	Amidosulfuron	-4,65	-4,66	0,198	0,01
2	Atrazine	-4,14	-4,41	0,093	0,26
3	Bromacil	-3,67	-4,39	0,086	0,68
4	Butylate	-1,63	0,24	0,12	-1,87
5	Chorimuronethyl	-8,35	-9,27	0,085	0,87
6	Chlorpropham	-5,46	-2,97	0,12	-2,5
7	Chlorsulfuron	-6,71	-8,52	0,079	1,69
8	Clopyralide	-3,69	-2,88	0,389	-1,41
9	2,4-D	-3,23	-1,96	0,049	-1,13
10	2,4-D dimethylammonium	-8,04	-6,97	0,291	-1,48
11	2,3,6-TBA	0,5	0,51	0,195	-0,01
12	Desmedipham	-7,6	-7,4	0,128	-0,2
13	Dichlorprop	-3,71	-5	0,116	1,28
14	DichlorpropP	-3,71	-4,21	0,116	0,5
15	Dipropetryn	-3,89	-4,01	0,08	0,11
16	EPTC	-3,84	-5	0,315	1,69
17	Ethametsulfuronmethyl	-11,4	-12,1	0,433	1,35
18	Fluazifopbutyl	-4,65	-4,26	0,184	-0,44
19	FluazifopP butyl	-4,65	-3,27	0,184	-1,55
20	fluometuron	-4,2	-3,9	0,044	-0,27
21	Fluoxypyrmeptyl	-3,94	-5,87	0,203	2,24
22	Haloxyfop	-5,13	-5,88	0,123	0,76
23	Haloxyfopethoxyethyl	-5,78	-7,79	0,159	2,15
24	Hexazinone	-5,68	-4,57	0,033	-0,96
25	Isoproturon	-3,67	-5,48	0,063	1,65
26	Lenacil	-7,45	-6,7	0,129	-0,76
27	Mecoprop	-3,67	-3,51	0,056	-0,15
28	MecopropP	-3,67	-3,4	0,056	-0,25
29	Metamitron	-4,98	-5,7	0,086	0,68
30	Metsulfuronmethyl	-8,98	-9,48	0,103	0,48
31	Napropemide	-8,41	-9,28	0,29	1,2
32	Pebulate	-0,18	0,67	0,116	-0,85
33	Phenmedipham	-7,56	-8,88	0,127	1,33
34	Picloram	-1,71	-4,09	0,111	2,35
35	Primisulfuron methyl	-6,51	-5,3	0,248	-1,53
36	Prometon	-4,54	-3,51	0,107	-1
37	Prometryn	-3,5	-3,77	0,045	0,24
38	Propazine	-4,74	-5,41	0,108	0,66
39	Prosulocarb	-3,54	-4,16	0,06	0,56
40	Prosulfuron	-6,01	-5,46	0,103	-0,54
41	Rimsulfuron	-7,21	-5,82	0,137	-1,43

N°	Composé	log Pv $_{Calc, Pred}$	log Pv $_{Exp}$	h_{ii}	e_{istd}
42	Tebuthiuron	-3,88	-3,57	0,09	-0,3
43	Terbacil	-3,1	-4,2	0,085	1,04
44	Terbutylazine	-4,5	-3,82	0,075	-0,63
45	Terbutryn	-3,52	-3,65	0,039	0,12
46	Thifensulfuronmethyl	-8,66	-7,77	0,09	-0,84
47	Triallate	-1,23	-1,8	0,142	0,59
48	Triasulfuron	-8,06	-10	0,098	1,87
49	Tribenuronmethyl	-8,74	-7,28	0,109	-1,44
50	Triclopyr	-4,46	-3,7	0,169	-0,82
51	Triclopyrbutotyl	-5,65	-3,77	0,085	-1,78
52	Vernolate	-0,73	0,14	0,102	-0,85
53	Vinclozolin	-5,87	-4,8	0,15	-1,13
54	Ametryn*	-4,1	-3,44	0,05	-0,56
55	Bensulfuronmethyl*	-9,9	-11,6	0,153	1,52
56	Chloroxuron*	-7,59	-6,62	0,171	-0,88
57	Cyanazine*	-6,01	-6,7	0,117	0,61
58	Cycloate*	-3,2	-2,67	0,041	-0,44
59	2,4,5-T*	-3,91	-6,15	0,126	1,98
60	Desmetryn*	-3,34	-3,88	0,055	0,46
61	Difenxuron*	-8,72	-8,91	0,174	0,17
62	Diuron*	-4,07	-5,96	0,043	1,6
63	Fenoxapropethyl*	-7,35	-5,38	0,221	-1,85
64	Linuron*	-4,57	-4,29	0,104	-0,25
65	MCPA*	-3,93	-4,64	0,057	0,6
66	MCPA isoctyl*	-2,85	-3,7	0,27	0,82
67	MCPB*	-3,88	-4,01	0,063	0,11
68	Methabenzthiazuron*	-5,9	-5,23	0,072	-0,58
69	Methazol*	-5,17	-3,88	0,073	-1,11
70	Metoxuron*	-4,27	-2,37	0,068	-1,63
71	Metribuzin*	-3,36	-4,24	0,059	0,75
72	quizalofop ethyl*	-7,45	-9,18	0,203	1,6
73	Simazine*	-4,52	-5,53	0,103	0,88
74	Triflusulfuronmethyl*	-6,22	-5	0,275	-1,18
75	Propaquizafop*	-8,75	-9,36	0,227	0,57

* Composés de validation,

La validité du modèle a été éprouvée par le test de randomisation de log Pv (Figure 26).

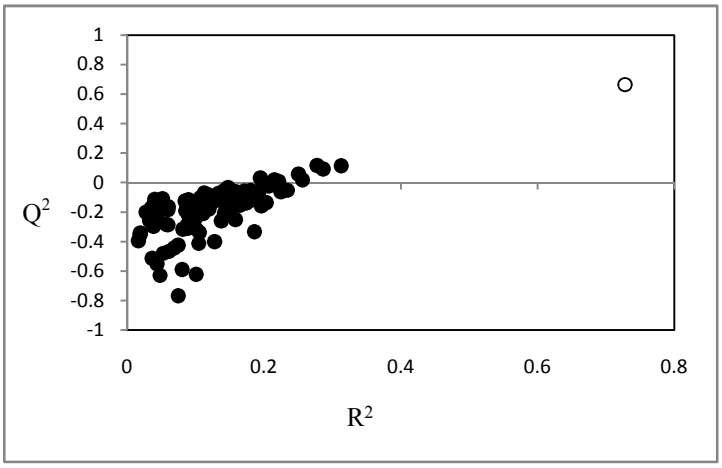


Figure 26 : Test de randomisation

Les 100 modèles pour lesquels nous avons randomisé les valeurs des logarithmes de la pression de vapeur ont des valeurs de Q^2 ou faibles ou négatives, et des valeurs R^2 petites. Seul le cercle videqui est isolé dans le graphe a des valeurs élevées et proches pour ces deux statistiques, il représente notre modèle qui, par conséquent, n'est pas dû au hasard.

IV. 2. 2.Moindres carrées partiels

Dans le but d'éliminer les corrélations entre les descripteurs on a appliqué la méthode des moindres carrées partiels (PLS).

Le tableau XVI montre que les trois premières composantes sont significatives, elles seront utilisées comme régresseurs.

Tableau XVI: Signification des composantes :

A	R^2X	R^2X(cum)	Valeurs propres	R^2Y	R^2Y(cum)	Q^2	Q^2(cum)	Significativité
1	0,546	0,546	3,28	0,51	0,51	0,443	0,443	R1
2	0,161	0,707	0,967	0,201	0,71	0,318	0,62	R1
3	0,0543	0,761	0,326	0,103	0,814	0,213	0,701	R1
4	0,128	0,89	0,771	0,00416	0,818	-0,0428	0,688	N4
5	0,0882	0,978	0,529	0,00166	0,819	-0,0104	0,684	N4
6	0,0221	1	0,132	0,0006	0,82	-0,0236	0,677	N3

L'équation de régression a été obtenue par une RLM appliquée sur les trois composantes:

$$\mathbf{log\ Pv} = 1{,}10(\pm 0{,}096)\mathbf{t_1} + 1{,}33\ (\pm 0{,}185)\mathbf{t_2} + 1{,}48\ (\pm 0{,}295)\mathbf{t_3} \qquad (67)$$

R^2= 80,93 Q^2_{LOO} = 77,50 Q^2_{LMO} (20%) = 75,94 Q^2_{BOOT} = 75,12

EQMP = 0,485 EQMC= 0,446 K_{xx}= 0,00 K_{xy} = 29,99

n= 53S= 0,464 F= 69,33

n_{ext}= 22 Q^2_{ext} = 79,62 $EQMP_{ext}$ = 0,438

D'après les statistiques on remarque qu'il n'y pas une grande amélioration en ce qui concerne les capacités prédictives du modèle obtenu par MLR, mais on a éliminé le problème de corrélation entre les descripteurs (K_{xx}=0).

La qualité de l'ajustement peut être vérifiée en procédant àla représentation des valeurs calculées de l'ensemble de calibrage et les valeurs prédites de l'ensemble de validation en fonction desvaleurs observées pour le logarithme de la pression de vapeur. La figure 27 reproduit la distribution autour de la première bissectrice des log $Pv_{calc,\ pred}$ en fonction du log Pvexpérimentale.

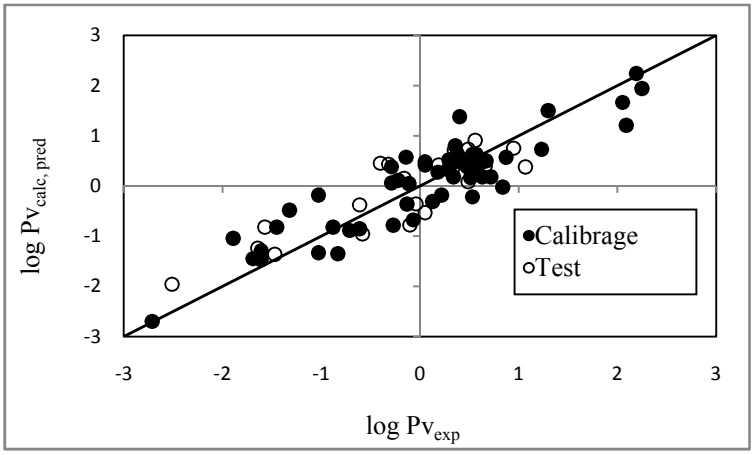


Figure 27 : Graphe des valeurs log $Pv_{calc,\ pred}$ en fonction des valeurs log Pv_{exp}

La faible dispersion des points autour de la première bissectrice montre que les valeurs prédites (pour l'ensemble de validation) et calculées (pour l'ensemble de calibrage) sont en adéquation avec les valeurs expérimentales.

Le diagramme de Williams représenté dans la figure 28 permet d'afficher les valeurs des résidus de prédiction standardisés en fonction de leviers (h_{ii}), pour les deux ensembles (calibrage et validation).

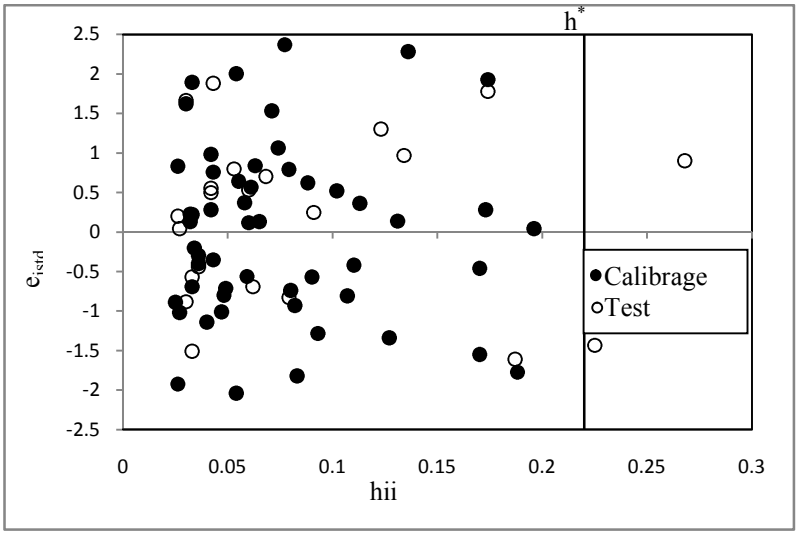


Figure 28 : Diagramme de Williams

Tous les résidus de prédiction standardisés sont compris entre les limites ±2,5σ. Le graphe fait ressortir deux points influents ($h_{ii} > h^* = 0,22$) de l'ensemble de validation.

IV. 3. Conclusion :

Les valeurs expérimentales de la pression de vapeur de 75 pesticides ont été séparées aléatoirement en deux sous-ensembles d'éléments l'un de 53 pour la sélection, par algorithme génétique, des descripteurs moléculaires théoriques calculés à partir de la structure des molécules puis la construction du modèle, et l'autr de 22 pour la validation externe.

Deux modèles linéaires ont été développés le premier par l'approche hybride GA/ MLR, et le deuxième par PLS. La comparaison de la qualité des modèles RLM et PLS pour la pression de vapeurmontre que les deux modèles RLM et PLS sont acceptables tant par la qualité de l'ajustement, la robustesse ou la capacité prédictive.

Le modèle PLS nous a permis d'éliminer le problème de colinéarité des descripteurs.

V. Modélisation du coefficient de partage octanol/carbone organique

V. 1. Introduction

Le coefficient de partage octanol / carbone organique, Koc (en cm^3/g), rend compte de la mobilité du pesticide dans le sol.

$$Koc = K / (\%OC) \quad (68)$$

où K est la constante de sorption, elle représente la quantité de produit adsorbé en équilibre.

La rétention d'une substance dans le sol dépend des propriétés de la substance et de la composition du sol (généralement du contenu en matière organique), K est le rapport des concentrations de produit dans le sol, en g/g de sol et dans l'eau, en g/cm^3 de solution. L'unité de K (et de Koc par conséquent) est le cm^3/g, Koc est donc une forme de la constante de sorption rapportée au carbone organique du sol. C'est une grandeur relativement indépendante du sol et qui permet de comparer les produits entre eux. C'est en effet une bonne estimation de la mobilité moyenne quand la sorption s'effectue sur la matière organique du sol. Dans le cas où la constante de sorption de la substance n'est pas corrélée à la matière organique du sol, le Koc n'est plus représentatif de la mobilité, c'est le cas pour les produits phytosanitaires de nature ionique qui ont une forte affinité pour les colloïdes minéraux du sol (argiles, oxydes de fer, oxydes de manganèse,…).

Le coefficient de partage K est une constante thermodynamique qui dépend de la nature de la substance à extraire, et du système de solvants utilisés. Partant du principe « qui se ressemble s'assemble », un solvant dissout bien un composé qui lui ressemble, les solvants polaires et dissociant tels que l'eau dissolvent les composés ioniques et / ou polaires hydrophiles. Les molécules apolaires présentent souvent un $K > 10$ entre un solvant hydrophobe et l'eau, les solvants apolaires et peu dissociant dissolvent les molécules et les composés hydrophobes (solvants chlorés ou hydrocarbures) (http://www,unige,ch/cabe/chimie-anal/extraction-liquide, pdf).

Le solvant intervient aussi par son caractère protique ou aprotique, les solvants protiques tels que l'eau, ammoniac, alcools, phénols, acides et amides non substituées influent sur l'extraction à cause de l'existence d'un hydrogène labile.

D'autre part, la nature du soluté intervient aussi dans le phénomène de partage ; en effet, la structure d'une molécule organique joue un rôle important sur son coefficient de partage, l'accroissement de la chaîne augmente la valeur de K d'environ 4 unités pour chaque groupement méthylénique (CH_2) incorporé dans la molécule. Un composé ramifié présente un K inférieur à celui de son isomère linéaire, il en va de même pour un composé non saturé par rapport au composé saturé correspondant. La présence d'hétéroatome (O, N) diminue parfois considérablement la valeur de K à cause des liaisons hydrogène avec l'eau. L'introduction d'un halogène au contraire favorise le passage en phase organique donc l'augmentation de K (http://www,unige,ch/cabe/chimie-anal/extraction-liquide, pdf).

L'inclusion d'une molécule simple dans un complexe chargé diminuera fortement le coefficient de partage, en effet la forme ionique ainsi obtenue est du fait de sa charge, beaucoup plus fortement retenue par l'eau, solvant polaire. Une molécule possédant un moment dipolaire (μ) non nul, définissant en quelque sorte l'énergie de liaisons van der Waals, la constante diélectrique (ε) elle, définit la polarisabilité moléculaire (ou polarisation moléculaire), elle permet de se rendre compte de l'affinité des solvants entre eux même (http://www,unige,ch/cabe/chimie-anal/extraction-liquide,pdf).

En résumé, pour une substance donnée, K augmente avec:

- la longueur de chaîne,
- la linéarité de la chaîne,
- la saturation de la chaîne,
- le contenu en chlorure,
- une diminution du contenu en oxygène ou azote,

V. 2. Résultats et discussion :

V. 2. 1. Régression linéaire multiple

La sélection par algorithme génétique conduit à un bon modèle à six descripteurs qui décrit le mieux le coefficient de partage octanol / carbone organique. Pour la modélisation par MLR, le modèle retenu est le suivant :

log Koc = - 0,065(±0,4316) + 0,0420(±0,007) **ALOGP2** - 6,31 (±1,500) **HATS5v** + 3,04 (±0,7237) **E2e** + 0,564 (±0,0861) **ATS6m**+ 0,181 (±0,0322) **SEigv** + 0,806 (±0,4807) **R5m** (72)

Les diagnostiques statistiques réunis ci- après permettent de faire des comparaisons et de tirer plusieurs conclusions.

R^2= 75,84 Q^2_{LOO} = 68,02 Q^2_{LMO} (20%) = 65,10 Q^2_{BOOT} = 64,72

EQMP = 0,411 EQMC= 0,358 K_{xx}= 35,26 K_{xy} = 40,45

n= 60 S= 0,380 F= 27,72

n_{ext}= 18 Q^2_{ext} = 61,71 $EQMP_{ext}$ = 0,452

Le modèle MLR obtenue est acceptable Q^2_{LOO}> 0,5, et significatif car il est caractérisé par une valeur de F de Fisher assez grande.

Tableau XVII : Matrice de corrélation

	log Koc	ALOGP2	HATS5v	E2e	ATS6m	SEigv
ALOGP2	0,618 0,000					
HATS5v	-0,468 0,000	-0,156 0,234				
E2e	0,245 0,059	0,190 0,146	-0,065 0,619			
ATS6m	0,399 0,002	0,266 0,040	-0,122 0,354	-0,062 0,640		
SEigv	0,024 0,856	-0,141 0,283	0,018 0,894	-0,195 0,136	-0,652 0,000	
R5m	-0,322 0,012	-0,091 0,491	0,653 0,000	-0,043 0,745	0,266 0,040	-0,524 0,000

La qualité de l'ajustement est représentée par le graphe des valeurs calculées du log Koc de l'ensemble de calibrage et les valeurs prédites de l'ensemble de validation en fonction de celles de log Koc mesurées.

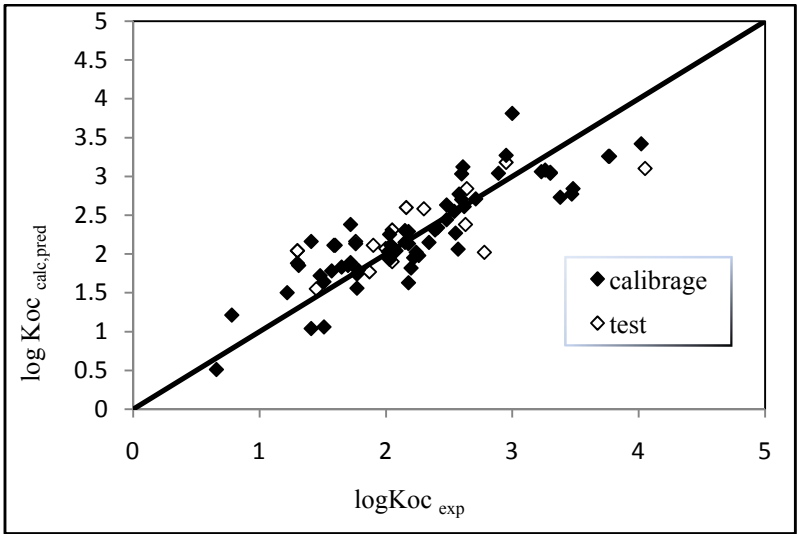


Figure 29 : Graphe des valeurs calculées, prédites en fonction des valeurs expérimentales

La figure ci- dessus (figure 29) montre que les points sont assez dispersés de la première bissectrice, ceci confirmé par la valeur de Q^2_{LOO} = 68,02%.

Le diagramme de Williams nous a permis de discuter le domaine d'application, le graphe des valeurs des erreurs standardisées en fonction des valeurs de hi est présenté dans la figure suivante :

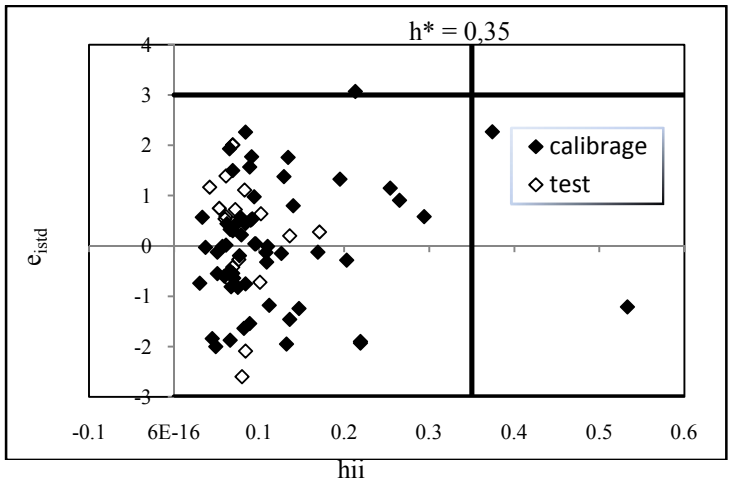


Figure 30: Diagramme de Williams

Le diagramme fait ressortir deux points influents de l'ensemble de calibrage hi > h* = 0,35, et un point aberrant de l'ensemble de calibrage (e_{istd}>3).

Les valeurs du coefficient de partage octanol/ carbone organique expérimentales, calculées, et prédites pour l'ensemble de validation, ainsi que les valeurs de leviers et des erreurs standardisées sont regroupées dans le tableau XVIII,

Tableau XVIII : Valeurs de log Koc expérimentales, calculée, prédites, h_{ii}, et e_{istd}

N°	Composé	log Koc_{Exp},	log $Koc_{Calc, Pred}$	h_{ii}	e_{istd},
1	Ametryn	2,48	2,63	0,062	0,44
2	Bensulfuron methyl	2,57	2,06	0,089	-1,54
3	Bromacil	1,51	1,64	0,294	0,58
4	Butylate	2,6	2,7	0,069	0,3
5	Carbetamide	1,77	1,56	0,059	-0,6
6	Chorimuron ethyl	2,04	2,11	0,079	0,22
7	Chloroxuron	3,26	3,08	0,069	-0,54
8	Clopyralide	0,66	0,51	0,533	-1,21
9	Clopyralid olamine	0,78	1,21	0,374	2,27
10	Cyanazine	2,26	1,98	0,075	-0,82
11	Cycloate	2,34	2,15	0,051	-0,55
12	2,4-D	1,22	1,5	0,254	1,15
13	2,4-D dimethylammonium	2,02	2,03	0,061	0,02
14	2,4,5-T-trolamine	1,41	1,04	0,147	-1,24
15	Desmedipham	3,47	2,77	0,049	-2
16	Desmetryn	2,18	2,29	0,066	0,33
17	Dichlorprop	1,41	2,16	0,084	2,26
18	Dipropetryn	2,95	3,27	0,094	0,98
19	Diuron	2,6	3,03	0,129	1,38
20	EPTC	2,24	2,02	0,07	-0,64
21	Ethametsulfuron methyl	2,03	1,93	0,109	-0,32
22	Fluazifop	1,31	1,85	0,134	1,76
23	Fluazifop butyl	3,77	3,26	0,219	-1,93
24	Fluazifop P butyl	3,76	3,26	0,219	-1,9
25	fluoxypyr eptyl	4,02	3,42	0,132	-1,95
26	Glyphosate	2,2	1,82	0,112	-1,18
27	Haloxyfop ethoxyethyl	2,03	2,25	0,265	0,91
28	Hexazinone	1,57	1,78	0,033	0,57
29	Isoproturon	1,72	2,38	0,065	1,93
30	Lenacil	2,22	1,95	0,03	-0,74
31	Linuron	2,54	2,55	0,095	0,04
32	MCPA isoctyl	3	3,81	0,213	3,07
33	MCPB	1,3	1,88	0,091	1,77
34	Methabenzthiazuron	2,15	2,15	0,057	-0,01
35	Methazol	3,48	2,84	0,066	-1,87
36	Metribuzin	1,59	2,11	0,089	1,57
37	Metsulfuron methyl	1,65	1,83	0,092	0,54
38	Napropemide	2,48	2,44	0,051	-0,12
39	Phenmedipham	3,38	2,73	0,045	-1,84
40	Prometon	2,18	2,13	0,126	-0,15

N°	Composé	log $Koc_{Exp,}$	log $Koc_{Calc, Pred}$	h_{ii}	$e_{istd,}$
41	Prometryn	2,58	2,77	0,078	0,57
42	Propazine	2,08	2,04	0,108	-0,13
43	Prosulocarb	3,23	3,06	0,066	-0,5
44	Prosulfuron	1,48	1,72	0,14	0,8
45	Quizalofop ethyl	2,71	2,71	0,11	-0,01
46	Rimsulfuron	1,78	1,79	0,096	0,04
47	Siduron	2,62	2,61	0,037	-0,03
48	Simazine	2,18	1,63	0,082	-1,64
49	Terbacil	1,51	1,06	0,136	-1,46
50	Terbutylazine	2,55	2,27	0,067	-0,81
51	Terbutryn	3,3	3,05	0,084	-0,75
52	Thifensulfuron methyl	1,7	1,85	0,084	0,45
53	Triasulfuron	2,15	2,3	0,074	0,45
54	Tribenuron methyl	1,72	1,89	0,09	0,51
55	Triclopyr	1,77	1,74	0,169	-0,12
56	Triclopyr butotyl	2,89	3,04	0,084	0,46
57	Triflusulfuron methyl	1,76	2,13	0,195	1,33
58	Vernolate	2,41	2,34	0,077	-0,19
59	Vinclozolin	2,39	2,31	0,203	-0,28
60	Propaquizafop	2,61	3,12	0,069	1,5
61	Atrazine*	2,05	1,9	0,069	-0,42
62	Chlorpropham*	2,64	2,84	0,06	0,54
63	Chlorsulfuron*	1,6	2,11	0,061	1,39
64	2,4-D-methyl*	2	2,07	0,136	0,2
65	Dichlorprop P*	1,76	2,16	0,083	1,11
66	Difenxuron*	2,3	2,58	0,053	0,75
67	Fenoxaprop*	2,16	2,6	0,042	1,17
68	Fenoxapropethyl*	4,05	3,1	0,08	-2,6
69	Fluometuron*	1,87	1,77	0,076	-0,27
70	MCPA*	1,45	1,55	0,171	0,28
71	Mecoprop*	1,3	2,04	0,069	2,01
72	MecopropP*	1,3	2,04	0,069	2,01
73	Metamitron*	2,78	2,02	0,084	-2,09
74	Metoxuron*	2,05	2,31	0,072	0,72
75	Pebulate*	2,63	2,38	0,065	-0,68
76	Tebuthiuron*	1,9	2,11	0,061	0,58
77	Thiobencarb*	2,95	3,18	0,102	0,64
78	Triallate*	3,3	3,04	0,101	-0,72

* Composés de validation

Le modèle a été également vérifié par Y-scrambling. La figure 31 fait ressortir des statistiques faibles (Q^2<0,2 ; R^2< 0,3) pour les vecteurs modifiés, alors que le point représentatif du modèle réel, qui est isolé dans le graphe, présente d bons paramètres statistiques, ce qui garantit l'existence d'une relation (multi), linéaire entre log Koc et les descripteurs sélectionnés.

Les statistiques des vecteurs modifiés du logarithme du coefficient de partage octanol/ carbone organique sont plus petites que celles du modèle réel.

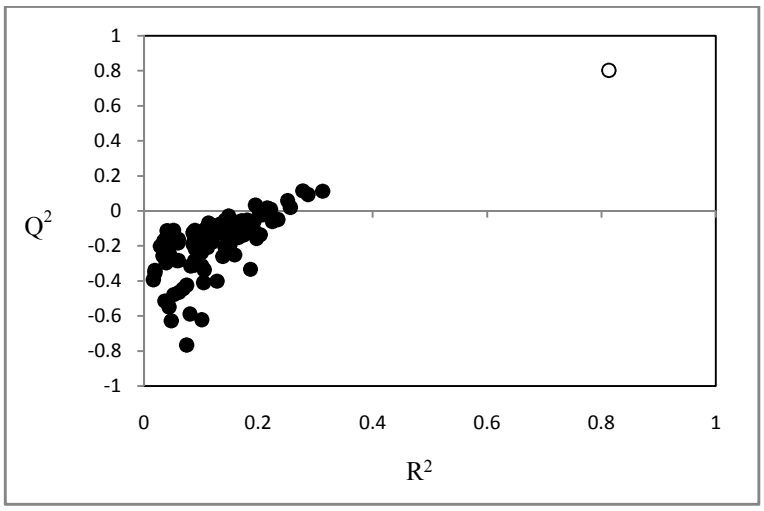


Figure 31 : test de randomisation

V. 2. 2. Les réseaux de neurones artificiels

Le choix du nombre de neurones de la couche cachée est fixé à 5 et le nombre d'itérations à 600. Le graphe suivant explicite ce choix.

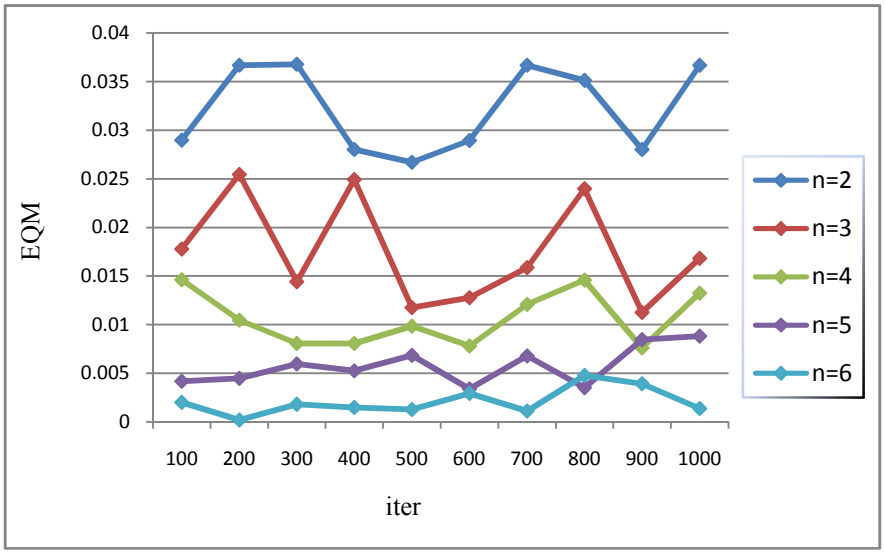


Figure 32 : Choix du nombre de neurones de la couche cachée

La structure optimale adoptée est reproduite dans le tableau XIX:

Tableau XIX : Structure optimale adopté pour le réseau de neurones

Entrées	06 (les descripteurs)
Sortie	01 (log Koc)
Couche cachée	Une couche cachée
Nombre de neurones dans la couche cachée	05
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonction d'apprentissage	Tangente hyperbolique (couche cachée) Linéaire (couche de sortie)

R^2= 81,30 Q^2_{LOO} = 80,22 Q^2_{LMO} (20%) = 79,32 Q^2_{BOOT} = 80,72

EQMP = 0,211 EQMC= 0,249 K_{xx}= 35,26 K_{xy} = 56,15

n= 60 S= 0,298 F= 47,72

n_{ext}= 18 Q^2_{ext} = 81,46 $EQMP_{ext}$ = 0,220

Les valeurs de R^2montre la qualité de l'ajustement, alors que la petite différence entre R^2 et Q^2_{LOO} renseigne sur la robustesse du modèle qui est, en outre, hautement significatif (grande valeur de la statistique F de Fisher), Les valeurs assez proches de EQMC et EQMP signifient que la capacité de prédiction interne du modèle n'est pas trop dissemblable de son pouvoir d'ajustement,

Le faible écart entre Q^2_{LOO} et $Q^2_{LMO/20}$ démontre la bonne stabilité dans la validation interne, et la validation par bootstrap (Q^2_{boot}) confirme tout à la fois la bonne capacité de prédiction interne et la stabilité du modèle,

La validation statistique externe (Q^2_{ext}) atteste de la bonne capacité prédictive des composés n'ayant pas participé au calcul du modèle,

La figure 45 qui représente le diagramme des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (cercles noirs) au modèle de départ (cercle vide), Il est clair que les statistiques obtenues pour les vecteurs modifiés de logKoc sont plus petites que celles des modèles réels ; les Q^2 sont inférieurs à 0,10, et pour la

majeure partie on obtient même un $Q^2 < 0$, Ceci permet d'assurer que le modèle établi à une base réelle, et n'est pas dû au hasard,

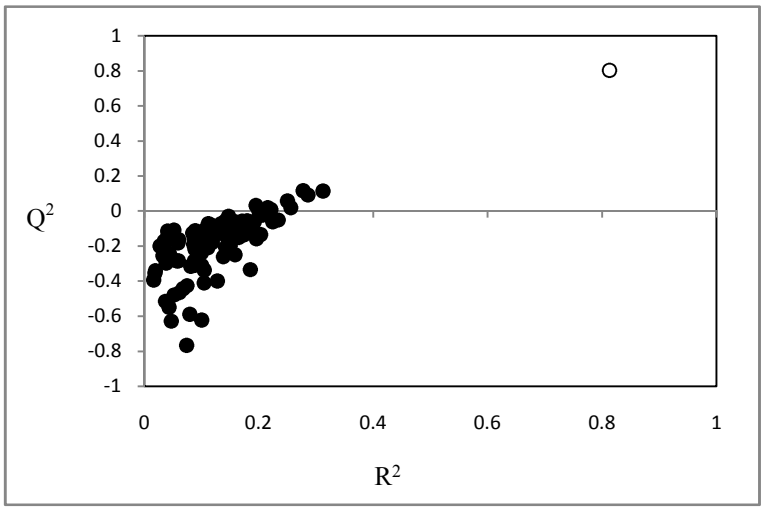


Figure 33: Test de randomisation

V. 3. Conclusion :

Les valeurs expérimentales de log Koc de 78 pesticides ont été séparées, aléatoirement, en deux sous-ensembles disjoints :

- de 60 éléments réservés au calcul des modèles à partir de descripteurs théoriques reflétant la structure moléculaire et en adoptant une approche hybride soit AG/ MLR, soit AG/ RNA ;
- de 18élements, exclusivement réservés à la validation externe

L'approche par réseaux de neurones conduit au meilleur modèle à tous les points de vue : capacités prédictives interne et externe, qualité de l'ajustement…, ce qui prouve dans ce cas, que les corrélations variable dépendante/ variablesexplicatives sont fondamentalement non linéaire.

Conclusion générale

L'emploi de quantités de plus en plus importantes de produits phytosanitaires, tout particulièrement en agriculture, a amené de nombreux chercheurs et de nombreux responsables de la gestion de la qualité de l'environnement à se poser des questions sur l'impact que ces produits pouvaient avoir sur la qualité des eaux et spécialement des eaux de surface.

La solubilité est un facteur important de l'impact des pesticides sur le milieu aquatique car elle limite obligatoirement la concentration.

On distingue :

- Les pesticides très peu solubles, en général les plus stables, dont la solubilité est de l'ordre de quelques mg par litre ;

- Les pesticides très solubles, à raison de plusieurs grammes par litre, correspondant à des composés très volatils et également très instables,

Si un pesticide est relativement soluble, son transfert de la zone d'utilisation au système aquatique, puis sa répartition dans celui- ci en est d'autant plus rapide. Au contraire, les pesticides insolubles mettent plus de temps pour atteindre le milieu aquatique et leur diffusion dans l'eau est très vite limitée par fixation sur les matières en suspension ou sur les sédiments. Ils affecteront alors la biologie du milieu uniquement, si le produit est très toxique, par accumulation dans les tissus ou par fixation sur la matière organique servant d'aliment aux différents organismes.

Abstraction faite de la stabilité chimique et de la biodégradation, la diminution de la concentration en pesticides reste essentiellement liée à leurs caractéristiques physico-chimiques comme la pression de vapeur ou la constante de Henry.

Nous avons utilisé la méthodologie QSPR pour relier cinq propriétés (constante de Henry, solubilité dans l'eau, pression de vapeur, et coefficient de partage octanol/ eau et octanol/ carbone organique) d'un mélange hétérogène de pesticides ayant des propriétés chimiques et des origines diverses, à des descripteurs moléculaires théoriques caractéristiques

de la molécule entière ou de ses fragments, calculés à l'aide de logiciels spécialisés du commerce.

Nous avons recherché des corrélations linéaires entre variables dépendantes et variables explicatives sélectionnées par algorithme génétique, en utilisant tour à tour la régression linéaire multiple et les moindres carrés partiels. Enfin, nous avons recherché des corrélations non linéaires en utilisant 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- Marquardt), et les Machines à vecteurs supports.

Pour la constante de Henry, la solubilité aqueuse, la pression de vapeur et le coefficient de partage ontanol/ eau (Kow) les modèles obtenus montrent que la relation entre ces propriétés et les descripteurs moléculaires est linéaire.

Pour le coefficient de partage ontanol/ carbone organique (Koc)l'approche par réseaux de neurones conduit au meilleur modèle à tous les points de vue : qualité de l'ajustement, robustesse interne et externe, capacité prédictive…, ce qui prouve que les corrélations variable dépendante- variables indépendantes sont fondamentalement non linéaires.

La comparaison de la qualité des modèles RLM et PLS montre qu'il n y a pas une différence tant par la qualité de l'ajustement, la robustesse ou la capacité prédictive.

En ce qui concerne la modélisation du coefficient de partage octanol/ eau l'approche SVM s'est a vérée la meilleure.

Références bibliographiques

Abraham M H, Andonian-Haftvan J, Whiting G S, Leo A, Taft S, (1994) "Hydrogen bonding 34: the factors that influence the solubility of gases and vapors in water at 298 K, and a new method for its determination", *Journal of the Chemical Society, Perkin Transactions 1*, Vol, 2, pp, 1777–1791.

Allen D M, (1974) "The relationship between variable selection and data augmentation and a method for prediction", *Technometrics*, Vol, 16, pp, 125- 127.

Allinger N L, (1976) "Calculation of molecular structure and energy by force field methods" *Adv, Phys, Org, Chem*, Vol, 13 (1), pp, 82- 221.

Allinger N L, (1977) "Conformational analysis. 130. MM2. a hydrocarbon force field utilizing V_1 and V_2 torsional terms", *J American Chemical Society*, Vol, 99, pp, 8127- 8134.

Allinger N L, Yuh Y H, Li J H, (1989) "Molecular mechanics. The MM3 force field for hydrocarbons.1",*J American Chemical Society*, Vol, 111, pp, 8551- 8565.

Allinger N L, Chem K, Lii J H, (1996) "n improved force field (MM4) for saturated hydrocarbons",*Computational Chemistry*, Vol, 17, pp, 642- 668.

Andrea T A, (1991) "Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors", *J, Med, Chem*, Vol, 34, pp, 2824 – 2836.

Anker L S, Jurs P C, (1992) "Prediction of carbon-13 nuclear magnetic resonance chemical shifts by artificial neural networks", *Anal, Chem*, Vol, 64, pp, 1157- 1164.

Aoyama T, Suzuki Y, Ichikawa H, (1990) "Neural networks applied to quantitative structure-activity relationship analysis*", J, Med, Chem*, Vol, 33, pp, 2583- 2590.

Atkinson A C, (1985), "Plots, Transformations and Regression", Clarendon Press, Oxford.

Benhamouche Z, (2010) "Plus de 200 cas d'intoxications/an par ingestion de pesticides à Oran".

Blinder SM, (1965) "Basic Concepts of Self-Consistent-Field Theory", *American Journal of Physics*, Vol, 33, pp, 431-520.

Bodor N, Huang M J, (1992) "An Extended Version of a Novel Method for the Estimation of Partition Coefficients", *Journal of Pharmaceutical Sciences*, Vol, 81, pp, 272–281.

Boivin A, (2003) "Disponibilité spatio-temporelle et transfert des pesticides dans le sol", thèse de doctorat : Sciences agronomiques, France, pp, 228.

Bordjiba O, Ketif A, (2009) "Effet de Trois Pesticides (Hexaconazole, Bromuconazole et Fluazifop-p-butyl) sur quelques Métabolites Physio Biochimiques du Blé dur : Triticum durum,Desf", *European Journal of Scientific Research*, Vol,36 (2), pp 260-268.

Bosque R, Sales J, Bosch E, Rosès M, Garcia- Alvarez- Coque M C, Torres- Lapasio J R, (2003) "A QSPR study of the p solute polarity parameter to estimate retention in HPLC", *J, Chem, Inf, Comput, Sci*, Vol, 43, pp, 1240- 1247.

Boulesteix A L, Strimmer K, (2006) "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data", *Brief, Bioinform*, Vol, 8, pp, 32-44.

Bounechada M, FenniM , Mekhlouf A, Tedjar L, (2011) "Conséquences des pesticides dans les eaux et leur impact sur la santé des population dans les hautes plainesSétifiennes (nord est de l'Algérie)".

Boust C, Institut national de recherche et sécurité, 1ere édition avril 2004, page 1- 6. www, INRS,fr accès le 1juin 2016.

Bouveresse D J R,Maalouly J,Jaillais B, (2004), "Sélection d'échantillons représentatifs par des méthodes chimiométriques : Application à des modèles d'étalonnage", Vol, 33 (237), pp 23-27.

Bouziani M, (2007), "L'usage immodéré des pesticides, De graves conséquences sanitaires", Le guide de la médecine et de la santé, Santé maghreb. http://www.santemaghreb.com/algerie/pointvue.htm accès le 29 mai 2016.

Bradbury SP, Russom CL, Ankley GT, Schultz TW, Walker JD, (2003) "Overview of data and conceptual approaches for derivation of Quantitative Structure- Activity Relationships for ecotoxicological effects of organic chemicals", *Environmental Toxicology* and *Chemistry*, Vol, 22 (8), pp, 1789-1798.

Burkert U N L, Allinger D C, (1982), "Molecular Mechanics", American Chemical Society, Washington.

Burns J, A, Whiteside G M, (1993) "Feed- forward neural networks in chemistry: mathematical systems for classification and pattern recognition", *Chem, Rev*, Vol, 93(8), pp, 2583- 2601.

Carson R, (1962) "Silent Spring", Boston, New York.

Chatfield C, Collins A J, (1980) "Introduction to multivariate analysis", Chapman & Hall, London.

Chen N, Lu W, Yang J, Li G, (2004) "Support Vector Machine in Chemistry", Singapore: World Scientific Publishing Co, Pte, Ltd.

Chouquet C, (2010) "Modèles Linéaires", Laboratoire de Statistique et Probabilités-Université Paul Sabatier-Toulouse.

Christopher M, Bishop, (1995) "Neural Networks for Pattern Recognition", Oxford University Press.

CIRAD-CA GEC AMATROP, (2000), Les herbicides,agroecologie,cirad,fr /2007/docs/1015714804,pdf, P1-7.

Clare B W, (1994)"Frontier orbital energies in quantitative structure- activity relationships: a comparison of quantum chemical methods"*Theor, Chim, Acta*, Vol, 87, pp, 415 – 430.

Clark RD, (1997) "OptiSim: an extended dissimilarity selection method or findingdiverserepresentativesubsets",*J, Chem, Inf, Comput, Sci*, Vol, 37, pp, 1181–1188.

Colas A, (1971) "La Technique de l'eau", Vol, 290, pp, 21-36.

Connolly M L, (1985) "Computation of molecular volume",*J, Am, Chem, Soc*, Vol, 107, pp, 1118- 1124.

Consonni V, Todeschini R, Pavan M, Gramatica P, (2002)"Structure/ responsecorrelations and similarity/ diversity analysis by GETAWAY descriptors, Part II, Application of the novel 3D molecular descriptors in QSAR/ QSPR studies", *J, Chem, Inf, Comput, Sci*, Vol, 42, pp, 693 – 705.

Cremlyn R, (1978) "Pesticides, Preparationand Mode of Action ",New York: Wiley.

Daszykowski M, WalczakB, MassartDL, (2002) "Representative subset selection", *Analytica Chimica Acta*, Vol, 468, pp, 91-103.

Dearden JC, Schüürmann G, (2003)"Quantitative Structure- Property Relationships for Predicting Henry's law constant from molecular structure", *Environ, Toxicol, Chem*, Vol, 22 (8), pp, 1755- 1770,

deBodt E, Cottrell M, Verleysen M,(2002) "Statistical tools to assess their liability of self-organizing maps", *Neural Networks*, Vol, 15, pp, 967 – 978.

Deeb, O, Goodarzi M, (2010) "predicting the solubility of pesticide compounds in water using QSPR methods", *Molecular physics*, Vol,108, pp,181-192.

Despagne F, MassartDL,(1998) "Neural networks in multi variate calibrage", *Analyst*, Vol, 123, pp, 157-178.

Dewar M J S, Thiel W J, (1977) "Results of MNDO calculations on molecules with H, C, N, O", *Am, Chem, Soc*, Vol, 99, pp, 4899- 4907.

Dewar M J S, Thiel W J, (1977) "A semienpirical model for the two-center repulsion integrals in the NDDO formalism", *J Theoretica chimica acta*, Vol, 46, pp, 89- 104.

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.

Directive européenne 91/414/CE du 15 juillet (1991).

Document d'aide technique pour les normes directives et objectif associés à la qualité de l'eau potable en Ontario, (2003).

Draper N R, Smith H, (1998)"Applied Regression Analysis", Third Edition, WileySeries in Probability and Statistics, New York.

Edelahid MC, (2004) "Contribution à l'étude de dégradation un situ des pesticides par procédés d'oxydation avancés faisant intervenir le fer, Application aux herbicides phénylurées", Thèse de doctorat (Université de Marne la Vallée), Chapitre 1, p 22-25.

Effets chroniques des pesticides sur la santé : état actuel des connaissances Janvier 2001 Observatoire Régional de la Santé de Bretagne.

Efron B, Tibshirani R J, (1993) "An Introduction to the Bootstrap", Chapman & Hall.

Erikson L, JohannsonE, Kettaneh- WoldN, (2001) "Multi and megavariate data analysis- principles and applications", Umetrics Academy, Umeå.

EriksonL, JaworskaJ, WorthA, CroninM,McDowell RM,Gramatica P, (2003)"Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs", *Environmental Health Perspectives*, Vol, 111 (10), pp, 1361-1375.

Escofier B, Pages J, (1998),"Analyses factorielles simples et multiples: Objectifs, méthodes et interprétation", 3ème ed,Dunod, Paris.

Estrada E, Delgado E J, Alderate J B, Jana G A, (2004)"Quantum- connectivity descriptors in modeling solubility of environmentally important organic compounds", *J, Comput, Chem*, Vol, 25(14), pp, 1787- 1796.

European Commission (2001) White Paper on a strategy for a future Community Policy for Chemicals, available at: http://europa,Eu,Int/comm,/enterprise/reach/ ,accés le 29 mai 2016

EXTOXNET, (1996) "Pesticide Information Profiles : Oxyfluorfene.

Fayet G, (2010) "Développement de modèles QSPR pour la prédiction des propriétés d'explosibilité des composes nitroaromatiques", Thèse de doctorat de l'université pierre et marie curie.

Fdil F, (2004)"Etude de la dégradation des herbicides chlorophénoxyalcanoïques par des procédés photochimique et électrochimique, Applications environnementales", Thèse (Docteur de l'Université de Marne-La-Vallée), Chapitre 1 pp, 8-25.

FEPS les pesticides et la pollution de l'eau. https://www.safewater.org/PDFS/resourcesknowthefacts/pesticides+pollution+eau.pdf. accés le 29 mai 2016

Fleischer G,Andoli V, Coulibaly M, RandolphT, (1998) "Analyse socio-économique de lafilière des pesticidesen Côte d'Ivoire", Série de Publications du Projet de Politique des PesticidesN° 06/F.

Flogeac K,(2004) "Etude de la capacité de rétention de produits phytosanitaires par deux solides modèles des sols, Influence de la présence des cations métalliques", Thèse de doctorat (université de Reims Champagne-Ardenne), Chapitre 1, p10-20.

Fortier J C, Messier C, (2005) Revue en science de l'environnement Vertigo, Vol 6 n° 2 Canada.

Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V, (1996)"Chemical information in 3D space", *J, Chem, Inf, Comput, Sci*, Vol, 36, pp, 1030 – 1037.

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.

Gelada P, Kowalski B, R, (1986) " Partial least- squares regression: tutorial", *Anal,Chim,Acta*, Vol, 185, pp, 1- 17.

Ghose A K, Crippen G M, (1986)"Atomic physico- chemical parameters for three-dimensional- structure- directed quantitative structure- activity relationships, I, Partition coefficients as a measure of hydrophobicity", *J, Comput, Chem*, Vol, 7, pp, 565 – 577.

Golbraikh A, Tropsha A, (2002),"Beware of q^2!",*J, Mol, Graph, Model*, Vol, 20, pp, 269-276,

Graybill F A,(1976),"Theory and Application of the Linear Model", Duxbury, North Scituate, Mass., pp, 231 – 236.

Gunn S R, (1998) "Support Vector Machines for Classification and Regression", Technical Report, University of Southampton, available at:http://users,ecs,soton,ac,uk/srg/publications/pdf/SVM,pdfaccés le 26 mai 2016.

Halgren T A, (1996), "Merck Molecular Force Field I. Basis, Form, Scope, Parameterization and Performance of MMFF94",*J Computational Chemistry*, Vol, 17, pp, 490- 519.

Halgren T A, (1996), "A General Program for Modeling Molecules and their Interactions",*J Computational Chemistry*, Vol, 17, pp, 520- 552, 553- 586, 616- 641.

Halgren T A, Nackbar R B, (1996), ""Merck Molecular Force Field IV. Conformational Energies and Geometries for MMFF94",*J Computational Chemistry*, Vol, 17, pp, 587- 615.

Hall G G, (1951) "The molecular orbital theory of chemical valency.VIII A method of calculating ionization potentials", *J Proceedings of the Royal Society of London A*, Vol, 205, pp, 541- 552.

Hansch C, Fujita T, (1964) "π –σ-π Analysis, A method for the correlation of biological activity and chemical structure",*J, Am, Chem, Soc*, Vol, 86, pp, 1616- 1626,

Hansch C, Lien E J, (1971) "Structure- activity relationships in antifungial agents, A survey",*J, Med, Chem*, Vol, 14, pp, 653- 670.

Hansen O C, (2004)"Quantitative Structure-Activity Relationships (QSAR) and Pesticides", (Pesticides Research No, 94, The Danish Environmental Protection Agency,), http://www2,mst,dk/udgiv/publications/2004/87-7614-434-8/pdf/87-7614-435-6,pdf, accés le 26-mai-2014)

Hecht-Nielson R, (1990) "Neurocomputing", Addison-Wesly Publishing Company, New York, pp, 433.

Hinton G, (1992) "Apprentissage et réseaux de neurones", *Pour la science*, Vol, 181, pp, 124-132.

Hohenberg P, Kohn W, (1964) "Inhomogeneous Electron Gas",*Phys, Rev*, Vol, 136, pp, 864-871.

Hopfield J J, (1982)"Neural Networks andphysical systems with emergent collective computational abilities",*Proceedings of the National Academy of sciences*, USA, Vol,79, pp,2554-2588.

Höskuldsson A, (1988) "PLS regression methods",*J, Chemometrics*, Vol, 2, pp, 211- 228.

http://agroecologie-cirad,fr mars 2000

Huang Q G, Kong I, Wang L S, (1996)"Applications of Frontier molecular orbital energies in QSAR studies", *Bull, Environ, Contam, Toxicol*, Vol,56, pp, 758 – 765.

Hyperchem ™ Release 6,03 for windows, Molecular Modeling System (2000).

Jager G, (1983) "Herbicides", in Buchel KH (ed): Chemistry of Pesticides, New York: Wiley, pp, 322- 392.

Jin C, Lei B, Li J, Li S, Shenb Y, Yao X, (2008) "Accurate and Validated Quantitative Structure–Activity Relationship Model of Caspase-mediated Apoptosis-inducing Activity of Phenolic Compounds Using Density Functional Theory Calculation and Genetic Algorithm–Multiple Linear Regression", *QSAR Comb, Sci*, Vol, 27, pp, 1318–1325.

Jollcliffe I T, (1986) "Principal Component Analysis", Springer- Verlag, Berlin.

Jorgensen W L, Duffy E M, (2002)"Prediction of drug of drug solubility from structure", *Adv, Drug Deliver, Rev*, Vol, 54, pp, 355 – 366.

Jurs P C, (1996) "Computer Software Applications in Chemistry", Second Edition, J, Wiely.

Kamlet M, Doherty RM,Abraham MH,Marcus Y, TaftRW, (1988)"Linear solvation energy relationships, 46, an improved equation for correlation and prediction of octanol/waterpartition coefficients of organic nonelectrolytes, (Including Strong Hydrogen Bond DonorSolutes)",*Journal of Physical Chemistry*, Vol, 92, pp, 5244–5255.

Karelson M, (2000) "Molecular descriptors in QSAR/ QSPR", Wiley- Interscience, pp, 385.

Katritzky A R, KuanarM, SlavovS, Hall CD,(2010)"Quantitative correlation of physical and chemical properties with chemicalstructure: utility for prediction",*Chemical Reviews*, Vol, 110, pp,5714–5789.

Kennard R W, Stone L A, (1969) " Computer Aided Design of Experiments", *technometrics*, Vol, 11 (1), pp 137-148.

Khan M A Q, (1977) "Pesticides in Aquatic Environements", Plenum Press, New York.

Kirby C, (1980) "The Hormone Weed killers a short history of their discovery and development", Croydon, UK, pp, 55.

KirkwoodJ,(1934)"Theory of solutions of molecules containing widely separated charges, withspecial application to zwitterions",*Journal of Chemical Physics*, Vol, pp,351–361,

Kleter G A, Bhula R, Bodnaruk K, Carazo E, Felsot A S, Harris C A, Katayama A, Kuiper H A, Racke K D, Rubin B, Shevah Y, Stephenson G R, Tanaka K, Unsworth J,

Wauchope R D, Wong S S, (2007) "Altered pesticide use on transgenic crops and the associated general impact from an environmental perspective", *Pest Management Science*, Vol, 63 (11), p 1107–1115.

KnuthDE, (1997) "The art of computer programming",Vol,2 (3rded,),Boston: Addison Wesley.

Kohonen T, (1995) Springer series in information sciences: Vol,30, Self-organizingmaps, Berlin: Springer-Verlag.

Kolos W, Wolniewiez L, (1964) "Accurate Adiabatic Treatment of the Ground State of the Hydrogen Molecule", *J, Chem, Phys*, Vol, 43, pp, 3663- 3673.

Koopmans T A, (1933), *Physica*, Vol, 1, pp, 104.

Kowalski B, GerlachR, WoldH, (1982) "Systems under indirect observation", (K, Jörenskoget H, Wold, eds,), North Holland, Amsterdam, pp, 191-206.

Kubinyi H, (1994)"Variable selection in QSAR studies: I, An evolutionary algorithm", *Quant, Struct,- Act, Relat*, Vol, 13, pp, 285- 294.

Leahy D, MorrisJ J, TaylorP J,(1992)"Model solvent systems for QSAR,Part 3, An LSERanalysis of the "Critical Quartet:" New Light on Hydrogen Bond Strength and Directionality",*Journal of the Chemical Society, Perkin Transactions 1*, Vol,2,pp, 705–731.

Leardi R, Boggia R, Terrile M, (1992)"Genetic algorithms as a strategy for feature selection", *J, Chemom*, Vol, 6, pp, 267- 281.

Lebart L, Morineau A, Piron M, (2004) "Statistique exploratoire multidimensionnelle", 3èmeed,Dunod, Paris.

Lejeune M, (2004) "Statistiques : la théorie et ses applications", Springer-Verlag, Paris.

Leo JA,(1981) "Hydrophobicity, the underlying property in most biochemical events, in environmental health chemistry", *J, McKinney*, Ed, Chapitre, 16, pp,323–336.

Leo JA, Hansch C, ElkinsD,(1971)"Partition coefficients and their uses",*Chemical Reviews*, Vol, 71, pp,525–616.

Levine I N, (2000)"Quantum Chemistry", 5th ed, New Jersey: Prentice Hall.

Lipinski C A, Lombardo F, Doming D W, Feeney P J, (2001)"Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings", *Adv, Drug Deliver, Rev*, Vol, 64, pp, 3- 25.

Lomas J S, (1986) "La mécanique moléculaire, un nom quantique pour le calcul de la structure et de l'énergie d'entités moléculaires", *l'actualité chimique*, pp, 7- 20.

Lowdin P O J, (1950) "On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals", *Chem, Phy*,, 18, pp, 365.

Mackay D, (2000) "In handbook of property estimation methods for chemicals: environmental and health sciences", Ed, By R, S, Boethling and D, Mackay (CRC Press LLC, Boca Raton) pp, 205.

Mackay D, Shiu W S, Ma K C, (2000) "Handbook of property estimation methods for chemicals: environmental and health sciences", Boethling R,S,, Mackay D,, eds, Lewis, Boca Raton, FL, USA.

Malinowsky E R, Howery D G, (1980) "Factor analysis in chemistry", Wiley Interscience, New York.

Marliere F ,(2000) "Mesure des pesticides dans l'atmosphère, INERIS DRC-00–23449–AIRE –569a-CDu-FMr.

Marquis J K, (1982) "Contemporary issues in pesticide toxicology and pharmacology", Basel: Karger, pp, 87- 95.

MATLAB, Version 7,0,0,19920 (Release 14), The Language of Technical Computing, The Math Works, Inc, May 06 (2004).

Maya R J, Maier H R, Dandy G C, (2010) "Data splitting for artificial neural networks using SOM- based stratified sampling", *Neural Networks* Vol, 23, pp, 283-294.

McCulloh WS, Pitts W, (1943) "A Logicial calculus of the ideas imminent in nervous activity", *Bull, Math,Biophys*, Vol, 5, pp, 115-133.

Mckerell A D, Jr Bashford D, Bellott M, Dunbrack R L, Jr, Evanseck J D, Field M J, Fischer S, Gao J, Guo H, Ha S, (1998), "All-atom empirical potential for molecular modeling and dynamics studies of proteins", *J chemical physics*, vol, 102, pp, 3586- 3616.

McEween FL, Stephson GR, (1979) "The use and significance of pesticides in the environment", New York: Wiley, pp, 91- 154.

Meloum M, Militky M, Forina M, (1992) "Chemometrics in Analytical Chemistry", Ellis Horwood, New York.

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.

Michael J L, Smith M C, Knisel W G, Fowler W P, Turton D J, (1996) "Using a hydrological model to determine environmentally safer windows for herbicide application",*New Zealand journal of forestry science*, Vol, 26, pp, 288-297,

Michael J L, Webber E C, Bayne D R, Fischer J B, Gibbs H L, Seesock W C, (1999) "Hexazinone dissipation in forest ecosystems and impacts on aquatic communities", *Canadian Journal of Forest Research*, Vol, 29 (7), pp, 1170-1181,

Michael J L, "Impact des herbicides sur les écosystèmes forestiers et aquatiques et la faune sauvage : l'expérience américaine", Rev. For. Fr. LIV - 6-2002.

MINITAB, Release 13,31, Statistical software, (2000),

Molegro, Data Modeller (MDM), v,2,1,0, Copyright Molegro(2009),

Moussaoui K M, Boussahe R, Tchoulak Y, Haouchine O, Benmami M, Dalachi A N, (2001) "Utilisation, évaluation et impacts des pesticides en Algérie, Ecole Nationale Polytechnique.www.recy.net/actualites/colloques/adep/20010605-pesticides.pptaccess le 30 mai 2016.

Nello C, Shawe-Taylor J, (2000) "An Introduction to Support Vector Machines and other kernel-based learning methods", Cambridge University Press.

Nello C, Joh S T, (2005) "An introduction to support vector machines and other kernel- based learning methods", Beijing: Publishing House of Electronics Industry.

Niketic S R, RasmussenK,(1977) "The Consistent Force Field: A Documentation", Springer, Berlin.

Orlando S, Sironi M, Bianchi G, Drummond AH, Boraschi D, Yabes D, Mantovani A, (1997) "Role of metalloproteases in the release of the IL-1 type II decoy receptor", *J boil chem*, Vol, 272(50), pp, 31764-31769.

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.

Perrin R, Scharff J P,(1997) Chimie industrielle 2^{ième}édition, Paris, Chapitre 7, pp 873-897.

Price N R, Watkins R W, (2003)"Quantitative structure-activity relationships (QSAR) in predicting the environmental safety of pesticides", *Pestic, Outlook*, Vol, 14, pp, 127- 129.

Ramachandran K I, Deepa G, Namboori K, (2008) " Computational chemistry and molecular modeling: principles and applications", DOI 10. 1007/978- 3- 540- 77304- 7.

Ran Y, He Y, Yang G, Johnson J L H, Alkowsky S HY, (2002)"Estimation of aqueous solubility of organic compounds by using the general solubility equation", *Chemosphere*, Vol, 48, pp, 487- 509.

Ren Y,Y,, H,X,Liu, X,J,Yao, M,C,Liu (2007) " Prediction of ozone tropospheric degradation rate constants by projection pursuit regression", *Anal, Chim,Acta*, Vol, 589, pp, 150–158.

Riahi S, Pourbasheer E, Ganjali M R, Norouzi P, (2009) "Investigation of different linear and non linear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine", *Journal of Hazardous materials*, Vol, 166, pp,853–859.

Roothan C C J, (1951) "New developments in molecular orbital", *J Theory, Reviews of Modern Physics,* Vol, 23, pp, 69- 89.

Ruelle P, Kesselring U W, (1997),"Aqueous solubility prediction of environmentally important chemicals from the mobile order thermodynamics", *Chemosphere*, Vol, 34, pp, 275- 298.

Rumelbart D E, Mcclelland J, L, et al ,(1988) Parallel Distributed processing, Vol, 1, Massachusetts: MIT press, pp, 547.

SCAN- Software for Chemometric Analysis- 1995, version 1,1- for Windows, Minitab USA

Scheyer A, (2000)"Développement d'une méthode d'analyse par CPG/MS/MS de 27 pesticides identifiés dans les phases gazeuses, particulaires et liquides de l'atmosphère, Application à l'étude des variations spatio-temporelles des concentrations dans l'air et dans les eaux de pluie", Chapitre 1, pp, 8-11 ;pp, 22-27 et chapitre 2,pp, 30-36.

Schuur J, Selzer P, 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", *J, Chem, Inf, Comput, Sci*, Vol, 36, pp, 334 – 344.

Shen M, Béguin C, Golbraikh A, Stables J P, Kohn H, Tropsha A, (2004)"Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds",*J, Med, Chem*, Vol, 47, pp, 2356 – 2364.

Shi LM, FangH, TongW, WuJ, PerkinsR, BlairR M, BranhamWS, DialSL, MolandCL,SheehanDM,(2001) "QSAR Models using a large diverse set of estrogens", *Journal of Chemical Information and Computer Science*, Vol, 41, pp, 186- 195.

Si H, YuanS, Zhang K, FuA, DuanYB, HuZ,(2008)"Quantitative structure activity relationship study on EC50 of anti-HIV drugs", *Chemometrics and Intelligent Laboratory Systems*, Vol, 90, pp, 15–24.

Skyner R E, McDonagh J L, Groom C R, van Mourik T, Mitchell J B O, (2015)"A review of methods for the calculation of solution free energies and the modeling of systems in solution", *Phys, Chem, Chem, Phys*, Vol, 17, pp, 6174- 6191.

Smola Alex J, Bernhard Schölkop, A,(2004), "Tutorial on support vector regression", *Statistics and Computing*, Vol, 14(3), pp, 199-222.

Snee R D, (1977)"Validation of Regression Models: Methods and Examples", *Technometrics*, Vol, 19, pp, 415- 428.

SprevakD, Azuaje F, Wang H, (2004) "Anon-random data sampling method For classification model assessment", In 17th international conference on pattern recognition,Vol,3,pp, 406-409.

Stevens J T, Breckenridge C B, (2001) "Agricultural chemicals: regulation, risk assessment, and risk management", in Regulatory Toxicology, Ed, By S, C, Gad (Taylor & Francis Ltd,, London,),pp, 215.

Strouf,O,(1986), Chemical Pattern Recognition,Wiley, New York.

Suteliffe B T, (1997) "The nuclear motion problem in molecular physics",*Adv, Quantum, Chem*, Vol, 28, pp, 65- 80.

Tenenhaus M, (1998), la régression PLS, théorie et pratique Paris : Technip.

Tham S Y, Agatonovic- Kustrin S,(2002) "Application of the artificial neural network in quantitative structure- gradient elution retention relationship of phenylthiocarbamyl amino acids derivaties", *J, Pharm, Biomed, Anal*, Vol, 28, pp, 581- 590.

The Pesticide Manual, (2006) 14th edition, pp, 1349.

Todeschini R, (1997)"Data Correlation, Number of Significant Principal Components and Shape of Molecules, The K correlation Index", *Anal,Chim, Acta*, Vol, 348, pp, 419- 430.

Todeschini R, Ballabio D, Consonni V, Mauri A, Pavan M, (2009) MOBY DIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, Release 1,1 for windows, Milano.

Todeschini R, Consonni V, Maiocchi A, (1999)"The K correlation index: theory development and its application in chemometrics", *Chemom, Intell, Lab, Syst*, Vol, 46, pp,13-29.

Todeschini R, ConsonniV, PavanM, (2005),DRAGON, Software for the Calculation of Molecular Descriptors, Release 5.3 for windows, Milano.

Todeschini R, Consonni V, Pavan, M (2006) DRAGON Software for the Calculation of Molecular Descriptors, Release 5.3 for windows, Milano.

Todeschini R, BallabioD, ConsonniV, MauriA, PavanM, (2004)MOBY DIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm,Release 1.0 for Windows, Milano.

Todeschini R, Gramatica P, Marengo E, Provenzani R, (1995)"Weighted holistic invariant molecular descriptors, Part, 2, Theory development and applications on modeling physico-chemical properties of polyaromatic hydrocarbons (PAH)", *Chemom, Intell, Lab, Syst*, Vol, 27, pp, 221 – 229.

Todeschini R, ConsonniV, (2009) "Molecular Descriptors for Chemoinformatics Volumes I & II", WILEY-VCH Verlag GmbH & Co, KGaA, Weinheim.

Todeshini R, Ballabio D, Consonni V, Mauri A, Pavan M, (2009)MOBY DIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, Release 1,1 for windows, Milano.

Tomassone R, Lesquoy E, Miller C, (1983) "La régression: nouveaux regards sur une ancienne méthode statistique", Masson, INRA.

Tourassi GA,FrederickED,MarkeyMK,FloydJrCE,(2001) "Application of the mutual information criterion for feature selection in computer-aided diagnosis", *Medical Physics*, Vol, 28(12), pp, 2394-2402.

Toussaint MW, SheddT R, Van der SchalieW H, LeatherG R,(1995) "A comparison of standard acute toxicity tests with rapid- screening toxicity tests",*Environmental Toxicology and Chemistry*, Vol, 14 (5), pp, 907- 915.

Tremaine LM, Diamond GL, QuebbemanAJ, (1984)"In vivo quantification of renal glucuronide and sulfate conjugation of 1- naphtol and p- nitrophenol in the rat", *BiochemPharmacol*,Vol, 33, pp, 419-427.

Tropsha A, Gramatica P, Gombar V K, (2003)"The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models", *QSAR Comb, Sci*, Vol, 22, pp, 69- 77.

User's Guide to SIMCA-P, SIMCA-P+, Version 11,0, Umetrics AB, 2005.

Vancolen, S, (2004), la régression PLS, groupe de statistique, université de Neuchâtel, Suisse.

Vapnik,V,(1995), " The nature of Statistical Learning Theory ", *Springer,* New York.

Viswanadhan V N, Reddy M R, Bacquet R J, Erion M D, (1993)"Assessment of methods used for predicting lipophilicity : application to nucleopides and nucleoside bases", *J, Comput, Chem*, Vol, 14, pp, 1019 – 1026.

Viswanadhan V N, Ghose A K, Revankar G R, Robins R K, (1989) "Atomic physicochemical parameters for three dimensional structure directed quantitative structure- activity relationships, 4, Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics", *J, Chem, Inf, Comput*, Vol, 19, pp, 163- 172.

Vladimir N, Vapnik V, (1995) "The nature of statistical learning theory", Springer-Verlag, New York, NY, USA.

Walker JD, (2003) "Applications of QSARs in toxicology: a US Government perspective", *Journal Of Molecular Structure-Theochem*, Vol, 622, pp, 167-184.

WangWJ,XuZB,LuWZ,ZhangXY,(2003)"DeterminationofthespreadparameterintheGaussiankernelforclassificationand regression",*Neurocomputing*, Vol,55, pp, 643–663.

Wauchope R D, Butler T M, Hornsby A G, Augustijn-Beckers P W M, Burt J P, (1992) "The SCS/ ARS/ CES pesticide properties database for environmental decision- making", *Reviews of environmental toxicology and chemistry*, Vol, 123, pp, 1-164.

Wehrens R H, Putter L M C, (2000) " The bootstrap: a tutorial", *Buydens, Chemom, Int, Lab, Syst*, Vol, 54, pp, 35- 52.

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

Willy J, Peumans EIS Van DammeJ M, (1995) "Lectins as piant defense proteins", *Plant Physiol*, Vol, 109, pp 347-352.

Wold S, (1984),"Chemometrics: Mathematics and Statistics in Chemistry, Reidel, Dordrecht, The Netherlands.

Wold S, Eriksson L, (1995)"Statistical validation of QSAR results", In: H, Van de Waterbeemd ed, Chemometrics methods in molecular design, VCH, New York, Vol, 2, pp, 309- 318.

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.

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.

Wu W, Walczak B; Massart D L; Heuerding S; Erni F; Last I R; Prebble K A, (1996)"Artificial neural networks in classification of NIR spectral data: Design of the training set", *Chemometrics and Intelligent Laboratory Systems*, Vol, 33, pp, 35-46.

Xu J, Zhang H, Wang Lei, Liang G, Wang L, Shen X, Xu W, (2010)"QSPR study of absorption maxima of organic dye- sensitized solar cells based on 3D descriptors", *Spectrochimica Acta Part A*, Vol, 76, pp, 239- 247.

Xu HY, ZhangJ Y, ZouJ W, ChenX S,(2008) "QSPR models for the physicochemical properties of halogenated methyl-phenyl ethers", *Journal of Molecular Graphics and Modelling*, Vol, 26, pp, 1076–1081.

Yao XJ, LiuMC, ZhangXY, Hu ZD, FanBT,(2002) "Radial basis function network-based quantitative structure–property relationship for the prediction of Henry's law constant", *AnalyticaChimicaActa*, Vol, 462, pp, 101–117.

Zeeman M, AuerC M, ClementsR G, NabholzJ V, BoethlingRS,(1995) "U,S, EPA Regulatory perspectives on the use of QSAR for new and existing chemical evaluations",*SAR and QSAR in EnvironmentalResearch***,** Vol, 3, pp, 179-201.

Annexe

Tableau I : Liste des composés étudiés

N°	Objet	Formule	N° CAS	N°	Objet	Formule	N° CAS
1	2,3,6-TBA	$C_7H_3Cl_3O_2$	50-31-7	*20*	chlorimuron-ethyl	$C_{15}H_{15}ClN_4O_6S$	90982-32-4
2	2,4,5-T	$C_8H_5Cl_3O_3$	93-76-5	*21*	clopyralid olamine	$C_8H_{10}Cl_2N_2O_3$	57754-85-5
3	2,4,5-T-trolamine	$C_{14}H_{20}Cl_3NO_6$	3813-14-7	*22*	clopyralid	$C_6H_3Cl_2NO_2$	1702-17-6
4	2,4-D	$C_8H_6Cl_2O_3$	94-75-7	*23*	cyanazine	$C_9H_{13}ClN_6$	21725-46-2
5	2,4-DB	$C_{10}H_{10}Cl_2O_3$	94-82-6	*24*	cycloate	$C_{11}H_{21}NOS$	1134-23-2
6	2,4-D-dimethylammonium	$C_{10}H_{13}Cl_2NO_3$	2008-39-1	*25*	desmedipham	$C_{16}H_{16}N_2O_4$	13684-56-5
7	2,4-DB	$C_{10}H_{10}Cl_2O_3$	94-82-6	*26*	desmetryn	$C_8H_{15}N_5S$	1014-69-3
8	2,4-D-methyl	$C_9H_8Cl_2O_3$	1928-38-7	*27*	dichlorprop	C9H8Cl2O3	120-36-5
9	ametryn	$C_9H_{17}N_5S$	834-12-8	*28*	dichlorprop p	C9H8Cl2O3	15165-67-0
10	amidosulfuron	$C_9H_{15}N_5O_7S_2$	120923-37-7	*29*	difenoxuron	C16H18N2O3	14214-32-5
11	atrazine	$C_8H_{14}ClN_5$	1912-24-9	*30*	diphenamid	$C_{16}H_{17}NO$	957-51-7
12	bensulfuron methyl	$C_{16}H_{18}N_4O_7S$	268-00-57	*31*	dipropetryn	$C_{11}H_{21}N_5S$	4147-51-7
13	bromacil	$C_9H_{13}BrN_2O_2$	314-40-9	*32*	diuron	$C_9H_{10}Cl_2N_2O$	330-54-1
14	butylate	$C_{11}H_{23}NOS$	2008-41-5	*33*	EPTC	$C_9H_{19}NOS$	759-94-4
15	carbetamide	$C_{12}H_{16}N_2O_3$	16118-49-3	*34*	ethametsulfuron methyl	$C_{15}H_{18}N_6O_6S$	97780-06-8
16	chloroxuron	$C_{15}H_{15}ClN_2O_2$	1982-47-4	*35*	fenoxaprop	$C_{16}H_{12}ClNO_5$	95617-09-7
17	chlorpropham	$C_{10}H_{12}ClNO_2$	101-21-3	*36*	fenoxaprop ethyl	$C_{18}H_{16}ClNO_5$	66441-23-4
18	chlorthal dimethyl	$C_{10}H_6Cl_4O_4$	1861-32-1	*37*	flamprop M isopropyl	$C_{19}H_{19}ClFNO_3$	63782-90-1
19	chlorsulfuron	$C_{12}H_{12}ClN_5O_4S$	64902-72-3	*38*	flamprop methyl	$C_{17}H_{15}ClFNO_3$	52756-25-9

N°	Objet	Formule	N° CAS	N°	Objet	Formule	N° CAS
39	fluazifop	$C_{15}H_{12}F_3NO_4$	69335-91-7	*60*	metribuzin	$C_8H_{14}N_4OS$	21087-64-9
40	fluazifop butyl	$C_{19}H_{20}F_3NO_4$	69806-50-4	*61*	metsulfuron methyl	$C_{14}H_{15}N_5O_6S$	74223-64-6
41	fluazifop P butyl	$C_{19}H_{20}F_3NO_4$	79241-46-6	*62*	napropamid	$C_{17}H_{21}NO_2$	15299-99-7
42	fluometuron	$C_{10}H_{11}F_3N_2O$	2164-17-2	*63*	pebulate	$C_{10}H_{21}NOS$	1114-71-2
43	haloxyfop	$C_{15}H_{11}ClF_3NO_4$	69806-34-4	*64*	phenmedipham	$C_{16}H_{16}N_2O_4$	13684-63-4
44	haloxyfop ethoxyethyl	$C_{19}H_{19}ClF_3NO_5$	87237-48-7	*65*	picloram	$C_6H_3Cl_3N_2O_2$	01/02/1918
45	fluroxypyr meptyl	$C_{15}H_{21}Cl_2FN_2O_3$	81406-37-3	*66*	primisulfuron methyl	$C_{15}H_{12}F_4N_4O_7S$	86209-51-0
46	hexazinone	$C_{12}H_{20}N_4O_2$	51235-04-2	*67*	prometon	$C_{10}H_{19}N_5O$	1610-18-0
47	isopropalin	$C_{15}H_{23}N_3O_4$	33820-53-0	*68*	prometryn	$C_{10}H_{19}N_5S$	7287-19-6
48	isoproturon	$C_{12}H_{18}N_2O$	34123-59-6	*69*	propaquizafop	$C_{22}H_{22}ClN3O_5$	111479-05-1
49	lenacil	$C_{13}H_{18}N_2O_2$	01/08/2164	*70*	propazine	$C_9H_{16}ClN_5$	139-40-2
50	linuron	$C_9H_{10}Cl_2N_2O_2$	330-55-2	*71*	prosulfuron	$C_{15}H_{16}F_3N_5O_4S$	94125-34-5
51	MCPA	$C_9H_9ClO_3$	94-74-6	*72*	prosulfocarb	$C_{14}H_{21}NOS$	52888-80-9
52	MCPA isoct	$C_{17}H_{25}ClO_3$	26544-20-7	*73*	quizalofop ethyl	$C_{19}H_{17}ClN_2O_4$	76578-14-8
53	MCPB	$C_{11}H_{13}ClO_3$	94-81-5	*74*	rimsulfuron	$C_{14}H_{17}N_5O_7S_2$	122931-48-0
54	mecoprop	$C_{10}H_{11}ClO_3$	7085-19-0	*75*	siduron	$C_{14}H_{20}N_2O$	1982-49-6
55	mecoprop P	$C_{10}H_{11}ClO_3$	16484-77-8	*76*	simazine	$C_7H_{12}ClN_5$	122-34-9
56	metamitron	$C_{10}H_{10}N_4O$	41394-05-2	*77*	tebuthiuron	$C_9H_{16}N_4OS$	34014-18-1
57	methabenzthiazuron	$C_{10}H_{11}N_3OS$	18691-97-9	*78*	terbacil	$C_9H_{13}ClN_2O_2$	5902-51-2
58	methazol	$C_9H_6Cl_2N_2O_3$	20354-26-1	*79*	terbutryn	$C_{10}H_{19}N_5S$	886-50-0
59	metoxuron	$C_{10}H_{13}ClN_2O_2$	19937-59-8	*80*	terbuthylazine	$C_9H_{16}ClN_5$	5915-41-3

N°	Objet	Formule	N° CAS	N°	Objet	Formule	N° CAS
81	thifensulfuron methyl	$C_{12}H_{13}N_5O_6S_2$	79277-27-3	*103*	Diazinon	$C_{12}H_{21}N_2O_3PS$	333-41-5
82	thiobencarb	$C_{12}H_{16}ClNOS$	28249-77-6	*104*	Dicrotophos	$C_8H_{16}NO_5P$	141-66-2
83	tri allate	$C_{10}H_{16}Cl_3NOS$	2303-17-5	*105*	Dichlorvos	$C_4H_7Cl_2O_4P$	62-73-7
84	triasulfuron	$C_{14}H_{16}ClN_5O_5S$	82097-50-5	*106*	Dimethoate	$C_5H_{12}NO_3PS_2$	60-51-5
85	tribenuron methyl	$C_{15}H_{17}N_5O_6S$	101200-48-0	*107*	Disulfoton	$C_8H_{19}O_2PS_3$	298-04-4
86	triclopyr	$C_7H_4Cl_3NO_3$	55335-06-3	*108*	Ethion	$C_9H_{22}O_4P_2S_4$	563-12-2
87	triclopyr butotyl	$C_{13}H_{16}Cl_3NO_4$	64700-56-7	*109*	Ethoprophos	$C_8H_{19}O_2PS_2$	13194-48-4
89	triflusulfuron methyl	$C_{17}H_{19}F_3N_6O_6S$	126535-15-7	*110*	Etrimfos	$C_{10}H_{17}N_2O_4PS$	38260-54-7
90	vernolate	$C_{10}H_{21}NOS$	1929-77-7	*111*	Oxydometon-methyl	$C_6H_{15}O_4PS_2$	301-12-2
91	vinclozolin	$C_{12}H_9Cl_2NO_3$	50471-44-8	*112*	Phorate	$C_7H_{17}O_2PS_3$	298-02-2
92	Acéphate	$C_4H_{10}NO_3PS$	30560-19-1	*113*	Phosalone	$C_{12}H_{15}ClNO_4PS_2$	2310-17-0
93	Azamethiphos	$C_9H_{10}ClN_2O_5PS$	35575-96-3	*114*	Phosmet	$C_{11}H_{12}NO_4PS_2$	732-11-6
94	Azinephos-eyhyl	$C_{12}H_{16}N_3O_3PS_2$	2642-71-9	*115*	Phosphamidon	$C_{10}H_{19}ClNO_5P$	13171-21-6
95	Azinephos-methy	$C_{10}H_{12}N_3O_3PS_2$	86-50-0	*116*	Phoxim	$C_{12}H_{15}N_2O_3PS$	14816-18-6
96	Chlorfenvinphos	$C_{12}H_{14}Cl_3O_4P$	470-90-6	*117*	Pirimiphos-ethyl	$C_{13}H_{24}N_3O_3PS$	23505-41-1
97	Chlorpyrifos	$C_9H_{11}Cl_3NO_3PS$	2921-88-2	*118*	Pirimiphos-methyl	$C_{11}H_{20}N_3O_3PS$	29232-93-7
98	Chlorpyrifos methyl	$C_7H_7Cl_3NO_3PS$	5598-13-0	*119*	Profenofos	$C_{11}H_{15}BrClO_3PS$	41198-08-7
99	Cyanophos	$C_9H_{10}NO_3PS$	2636-26-2	*120*	Thiometon	$C_6H_{15}O_2PS_3$	35400-43-9
100	Fenitrothion	$C_9H_{12}NO_5PS$	122-14-5	*121*	Trichlorfon	$C_4H_8Cl_3O_4P$	35400-43-12
101	Fenthion	$C_{10}H_{15}O_3PS_2$	55-38-9	*122*	Malathion	$C_{10}H_{19}O_6PS_2$	121-75-5
102	Fonofos	$C_{10}H_{15}OPS_2$	944-22-9	*123*	Methamidophos	$C_2H_8NO_2PS$	10265-92-6

N°	Objet	Formule	N° CAS	N°	Objet	Formule	N° CAS
124	Formothion	$C_6H_{12}NO_4PS_2$	2540-82-1	*130*	Propetamphos	$C_{10}H_{20}NO_4PS$	31218-83-4
125	Isazofos	$C_9H_{17}ClN_3O_3PS$	42509-80-8	*131*	Sulprofos	$C_{12}H_{19}O_2PS_3$	35400-43-2
126	Isofenphos	$C_{15}H_{24}NO_4PS$	25311-71-1	*132*	Temephos	$C_{16}H_{20}O_6P_2S_3$	35400-43-4
127	Mevinphos	$C_7H_{13}O_6P$	26718-65-0	*133*	Terbufos	$C_9H_{21}O_2PS_3$	35400-43-5
128	Naled	$C_4H_7Br_2Cl_2O_4P$	300-76-5	*134*	Tetrachlorvinphos	$C_{10}H_9Cl_4O_4P$	35400-43-6
129	Methidathion	$C_6H_{11}N_2O_4PS_3$	950-37-8				

QSPR study of the water solubility of a diverse set of agrochemicals: hybrid (GA/ MLR) approach

Etude QSPR de la solubilité aqueuse d'un ensemble de divers produits agrochimiques: approche hybride (AG/RLM)

Amel Bouakkadia, Hamza Haddag, Nabil Bouarra & Djelloul Messadi*

Environmental and Food Safety Laboratory, Badji Mokhtar University, Annaba , PO Box 12, 23000, Algeria.

Soumis le : 05/03/2015 *Révisé le : 21/03/2016* *Accepté le : 12/04/2016*

ملخص

أجريت علاقة بين كمية الهيكل و الخاصية للتنبؤ بلنحلالية المبيدات منتمية إلى أربعة أقسام كيميائية: أحماض، اليوريا، تريازين، وكاربامات. المجموعة المكونة من 77 مبيد قسمت إلى مجموعة بناء من 58 مبيد ومجموعة اختبار من 19 مبيد بتقنية. النموذج بستة متغيرات بمعامل ارتباط (R^2) يساوي 0.8895 و خطا معيار التقدير (s) يساوي 0.52 وحدة تم تطويره بتطبيق التراجع المتعدد الخطي باستخدام المربعات الصغرى واختيار مجموعة المتغيرات تم باستعمال الخوارزمية الجينية. قوة النموذج المقترح تأكدت باستخدام عدة تقنيات للتقييم leave- one- out ، bootstrap ، الاختبارات العشوائية، والتحقق من خلال مجموعة الاختبار.

الكلمات الدالة: *المبيدات ، الانحلالية، QSPR ، الموصفات الجزيئية، التراجع المتعدد الخطي.*

Abstract

A quantitative structure- property relationship (QSPR) was performed for the prediction of the aqueous solubility of pesticides belonging to four chemical classes: acid, urea, triazine, and carbamate. The entire set of 77 pesticides was divided into a training set of 58 pesticides and a test set of 19 pesticides according to the Snee technique. A six descriptor model, with squared correlation coefficient (R^2) of 0.8895 and standard error of estimation (s) of 0.52 log unit, was developed by applying multiple linear regression analysis using the ordinary least square regression method and genetic algorithm- variable subset selection. The reliability of the proposed model was further illustrated using various evaluation techniques: leave- one- out cross- validation, bootstrap, randomization tests, and validation through the test set.

Key Words: *pesticides- aqueous solubility- QSPR- molecular descriptors- multiple linear regression.*

Résumé

Une relation quantitative structure-propriété (QSPR) a été réalisée pour la prédiction de la solubilité aqueuse des pesticides appartenant aux quatre classes chimiques: acide, urée, triazine, et carbamate. L'ensemble des 77 pesticides a été divisé en un ensemble de calibrage de 58 pesticides et un ensemble de test de 19 pesticides selon la technique de Snee. Un modèle de six descripteurs, avec un coefficient de corrélation (R^2) de 0,8895 et une erreur standard d'estimation (s) de 0,52, a été développé en appliquant une analyse de régression linéaire multiple en utilisant la méthode de régression des moindres carrés ordinaires et les algorithme-génétiques pour la sélection des sous-ensembles de variables. La fiabilité du modèle proposé a été en outre illustrée en utilisant diverses techniques d'évaluation: validation croisée par leave- one- out, bootstrap, tests de randomisation, et la validation par l'ensemble de test.

Mots clés: *pesticides- solubilité aqueuse- QSPR- descripteurs moléculaires- régression linéaire multiple.*

* *Corresponding author : d_messadi@yahoo.fr*

1. INTRODUCTION

The massive use of agrochemicals, known generically as pesticides [1], has allowed significant reduction in the agricultural plagues, and consequently, increased the productivity. On the other hand, the massive use of these agrochemicals has an environmental cost (due to their toxicity, their persistence, or their tendency to bioaccumulation), which is necessary to evaluate to conciliate productivity and environment protection [2].

Solubility in water is an important physicochemical property, having numerous applications to the modeling of the environmental effects of chemicals [3]. It is a direct measurement of hydrophobicity, that is, the tendency of water to exclude the substance from solution. Although the experimental determination of solubility is not difficult, there are some justifications to develop models that can predict it. This is especially important in environmental studies where the compounds are toxic, carcinogenic, or undesirable for some or other reason.

An extensive series of studies for the prediction of aqueous solubility has been reported in the literature [4- 10]. These methods can be categorized into three types:

1 - Correlation of solubility with experimentally data such as melting point (MP) and log P (logarithme of octanol/ water partition coefficient). However, this approach is of little use because it requires a knowledge of the compound's experimental melting point which is not available for virtual compounds. The melting point is a key index of the cohesive interactions in the solid and it is difficult to estimate.

2 - Estimation of solubility by group contribution methods. The group contribution method allows the approximate calculation of solubility by summing up fragmental values associated with substructural units of the compounds. The disadvantages of the group contribution method are that: 1/ the groups included must be defined in advance and therefore the solubility of a new compound containing new groups cannot be estimated; and 2/ the different effects of a group in different chemical environments are not considered.

3 - Correlation of solubility with descriptors derived from the molecular structure by computational methods. This third approach has been proven to be particularly successful for the prediction of solubility because it does not need experimental descriptors and can therefore also be applied to collections of virtual compounds.

The aim of the present work is to develop a robust QSPR model that could predict the aqueous solubility values for a diverse set of agrochemicals (which consists of 26 acids, 25 ureas, 13 triazines and 13 carbamates) using the general molecular descriptors computed with the help of DRAGON software [11].

2. METHODS

2.1 Experimental Data

The experimental S values (mg/l) of 77 selected, structurally heterogeneous, pesticides were taken from Hansen [12]. The water solubility values (log S) span between -1.05 and 5.90 (Table 1). The detailed structures of all studied compounds are available as Supporting Information.

2.2 Descriptor Generation

The chemical structure of each compound was sketched on a PC using the HYPERCHEM program [13] and preoptimized using MM+ molecular mechanics method (Polack- Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi- empirical PM3 method at a restricted Hartree- Fock level with no configuration interaction, applying a gradient norm limit of 0.01 kcal.$Å^{-1}$.mol^{-1} as a stopping criterion. Then the geometries were used as input for the generation of 1664 descriptors using the Dragon software (version 5.4) [11]. Quantum-chemical descriptors such as HOMO (highest occupied molecular orbital), LUMO (lowest unoccupied molecul arorbital), HOMO – LUMO gap (DHL), and ionization potential (P_{ion}), calculated by the semi empirical PM3 method using [13], were added and used for descriptor selection during model development.

Constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (when there was more than 98% pairwise correlation, one variable was deleted), and the genetic algorithm was applied for variables selection to a final set of 1230 descriptors.

2.3 Selection of the training and test sets

It is important to rationally define a training set from which the model is built and external test set on which to evaluate its prediction power. The object of this selection should be to

generate two sets with similar molecular diversity, in order to be reciprocally representative and to cover all the main structural and physicochemical characteristics of the global data set.
Several procedures can be adopted for the selection of the training and test sets, the later should contain between 15 and 40% of the compounds in the full data set.

DUPLEX algorithm adopted in this study proceeds as follows. In the first step, the two points which are furthest away from each other are selected for the training set. From the remaining points, the two- objects which are furthest away are included in the test set. In the third step, the remaining point which is furthest away from the two previously selected for the training set is included in that set. The procedure is repeated selecting a single point for the test set which is furthest from the existing points in that set. Following the procedure, points are added alternately to each set [14]. This algorithm was applied in the present study to separate data into two independent subsets: a training set of 58 compounds to build the model and a test set of the remained 19 compounds to evaluate its prediction ability.

2.4 Model Development and Validation

Multiple linear regression analysis (MLR) and variable selection were performed by the software MobyDigs [15] using the Ordinary Least Square regression (OLS) method and Genetic Algorithm-Variable Subset Selection (GA-VSS) [16].
The outcome of the application of the genetic algorithms is a population of 100 regression models, ordered according to their decreasing internal predictive performance, verified by Q^2. The models with lower Q^2 are those with fewer descriptors. First of all, models with 1-2 variables were developed by the all – subset – method procedure in order to explore all the low dimension combinations. The number of descriptors was subsequently increased one by one, and new models were formed. The best models are selected at each rank, and the final model must be chosen from among them. This has to be sufficiently correlated and, at the same time, protect against any overparameterization, which would lead to a loss of predictive power for molecules outside training set. From a statistical view point the ratio of the number of samples (n) to the number of descriptors (m) should not be too low. Usually, it is recommended that $n/m \geq 5$ [17]. The GA was stopped when increasing the model size did not increase the Q^2 value to any significant degree. Particular attention was paid to the collinearity of the selected molecular descriptors: by applying the QUIK rule (Q Under Influence of K) [18] a necessary condition for the model validity. Acceptable model is only that with a global correlation of [x + y] block (K_{xy}) greater than the global correlation of the x block (K_{xx}) variable, x being the molecular descriptors and y the response variable.
The collinearity in the original set of molecular descriptors results in many similar models that more or less yield the same predictive power (in MOBYDIGS software 100 models of different dimensionality). Therefore, when there were models of similar performance, those with higher ΔK (K_{xy}- K_{xx}) were selected and further verified.
The models were justified by the R^2, the adjusted R^2, the cross-validated values of Q^2 by leave-one-out (LOO), the F ratio values and the standard error s.
The robustness of the models and their predictivity were evaluated by both Q^2_{LOO} and bootstrap. In this last procedure K n-dimensional groups are generated by a randomly repeated selection of n- objects from the original data set.
The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 8000 times.
The proposed model was also checked for reliability and robustness by permutation testing: new models are recalculated for randomly recorded response (Y- scrambling) by using the same original independent variable matrix. After repeating this test several times (100 times in this work) it is expected to obtain new models that have significantly lower R^2 and Q^2 than the original model. If this condition is not verified the original model is not acceptable, as it was due to a chance correlation or a structural redundancy in the training set.
Obtaining a robust model does not give real information about its prediction power. This is evaluated by predicting the compounds included in the test set.The external for the test set is determined with equation (1):

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

Here n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively.

2.5 Applicability Domain Analysis

The applicability domain (AD) [19, 20] is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSPR should make reliable predictions. In this work, the structural AD was verified by the leverage (h_{ii}) approach [21].

The warning leverage h^* is, generally, fixed at $3(m + 1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation.

The presence of both the response outliers (Y outliers) and the structurally influential compounds (X outliers) was verified by the Williams plot [22], the plot of standardized residuals versus leverage values.

Table 1: Experimental and calculated logS for the studied pesticides.

No	Expt. logS	Calc. logS	Residual	No	Expt. logS	Calc. logS	Residual
1	3.89	3.86	0.03	40	1.88	2.47	-0.59
2	2.18	1.97	0.21	41	2.87	3.03	-0.16
3	2.95	3.44	-0.49	42 *	1.64	1.45	0.19
4	5.90	5.05	0.85	43	2.87	2.63	0.24
5	1.66	1.50	0.16	44	2.93	2.63	0.30
6	2.00	2.76	-0.76	45	3.23	2.72	0.51
7	2.27	1.85	0.42	46	1.77	2.56	-0.79
8	1.52	1.71	-0.19	47 *	2.83	2.10	0.73
9 *	2.08	2.86	-0.78	48 *	3.09	2.33	0.76
10 *	2.85	3.55	-0.70	49	3.98	3.53	0.45
11	1.64	1.75	-0.11	50 *	1.87	1.26	0.61
12	3.54	3.68	-0.14	51	2.00	1.78	0.22
13 *	0.40	0.85	-0.45	52	0.67	1.41	-0.74
14 *	1.95	1.88	0.07	53	2.63	3.27	-0.64
15 *	4.45	3.48	0.97	54 *	2.39	2.93	-0.54
16	3.08	2.79	0.29	55	2.86	1.94	0.92
17	5.75	5.71	0.04	56	1.52	1.49	0.03
18	5.16	4.95	0.21	57	-0.20	-0.11	-0.09
19	2.23	1.81	0.42	58	0.93	1.22	-0.29
20 *	1.98	1.80	0.18	59	3.60	2.91	0.69
21	0.95	0.27	0.68	60	1.12	0.90	0.22
22	2.76	2.10	0.66	61	-0.51	-0.14	-0.37
23	2.54	2.55	-0.01	62	3.86	3.32	0.54
24 *	2.77	2.55	0.22	63 *	0.79	1.94	-1.15
25 *	1.30	1.14	0.16	64	3.40	3.68	-0.28
26	1.20	0.93	0.27	65 *	2.85	3.48	-0.63
27	1.62	1.40	0.22	66	1.34	1.75	-0.41
28 *	2.54	2.35	0.19	67 *	0.93	0.81	0.12
29	1.70	2.81	-1.11	68	3.80	4.13	-0.33
30	-0.10	-0.66	0.56	69	1.45	1.26	0.19
31	0.00	-0.18	0.18	70	0.60	0.68	-0.08
32	0.30	-0.18	0.48	71	2.91	2.40	0.51

33	2.04	2.38	-0.34	72 *	3.18	3.44	-0.26
34	-1.05	-0.59	-0.46	73	3.91	3.17	0.74
35	4.08	4.29	-0.21	74	1.36	0.74	0.62
36	1.64	1.81	-0.17	75	2.04	2.19	-0.15
37	0.11	0.74	-0.63	76 *	2.03	2.33	-0.30
38	1.81	2.38	-0.57	77	0.53	1.47	-0.94
39	0.78	1.61	-0.83				

* Members for the test set.

3. RESULTS AND DISCUSSION

3.1 Results of the MLR Model

The dissolving process is the establishment of equilibrium between the phase of solute and its saturated aqueous solution. Aqueous solubility is almost exclusively dependent on the intermolecular forces that exist between the solute molecules and the water molecules. The solute- solute, solute- water, and water- water adhesive interactions determine the amount of compound dissolving in water. Additional solute- solute interactions are associated with the lattice energy in the crystalline state.

The solubility of a compound is thus affected by many factors: the state of solute, the relative aromatic and aliphatic degree of the molecules, the size and shape of the molecules, the polarity of the molecule, steric effects, and the ability of some groups to participate in hydrogen bonding.

In order to predict solubility accurately, all these factors correlated with solubility should be represented numerically by descriptors derived from the structure of the molecule.

A best six- parameters equation was obtained, which is as the following:

log S = - 2.80 - 1.27 $\mathbf{E_{HOMO}}$ - 0.182 **Mor02v** - 17.2 **G2e** - 9.56 **HATS7v** + 4.76 **RTu+** - 0.0821 **AlogP2** (2)

R^2= 0.8895 R^2_{adj}= 0.8765 Q^2_{LOO} = 0.8547 Q^2_{EXT} =0.8511 Q^2_{BOOT} = 0.8323 s = 0.52 log unit F = 68.42

K_{xx}= 37.68 K_{xy} = 45.67

Here, E_{HOMO} is the Highest Occupied Molecular Orbital energy [23, 24] ; Mor 02 v is the 3D-MoRSE- signal 02/ weighted by atomic van der Waals volume [25, 26] ; G2e is the second component symmetry directional WHIM index/ weighted by atomic Sanderson electronegativities [27, 28] ; HATS7v is the leverage weighted autocorrelation of lag 7/ weighted by atomic van der Waals volumes [29, 30] ; RTu+ is the R maximal index/ unweighted [29, 30] ; AlogP2 is the squared Ghose-Crippen-Viswanadhan octanol-water partition coefficient [31, 32].

More information about these descriptors can be found in [33] and the references therein.

The results for the randomized models can be compared with the real starting one only by representing in a plot the statistical coefficients R^2 and Q^2. This is depicted in figure 1. The statistics for the modified logS vectors are clearly lower than the real QSPR model. This ensures that a real structure-property relationship has been found out.

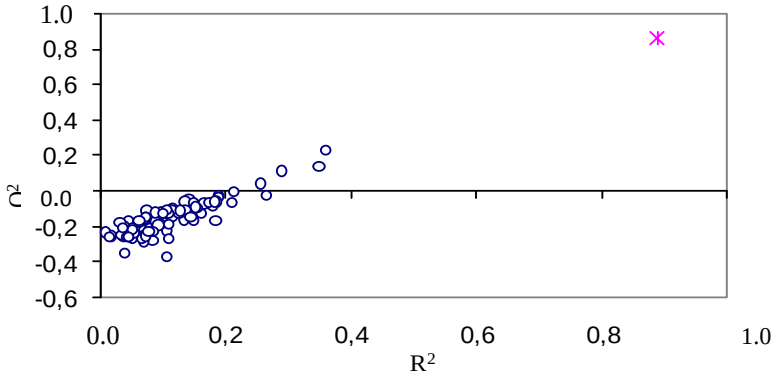


Figure1. Randomization test associated to previous QSPR model. Circles represent the randomly ordered solubilities, and star corresponds to the real solubilities.

Some important statistical parameters (as given in table 2) were used to evaluate the involved descriptors. The t-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute t-values shown in table 2 express that the regression coefficients of the descriptors involved in the MLR model are significantly larger than the standard deviation. The t-probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (i. e., descriptor's interactions). Descriptors with t-probability values below 0.05 (95% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance [34]. The smaller t-probability suggests the more significant descriptor. The t-probability values of the six descriptors are very small, indicating that all of them are highly significant descriptors. The VIF values suggest that these descriptors are weakly correlated with each others. Thus, the model can be regarded as an optimal regression equation.

The calculated log S values from equation (2) for the training and test set are showed in table 1 and figure 2. The distribution of errors for the entire data set is given in figure 3. As the errors are distributed on both sides of the zero line, one may conclude that there is no systematic error in the developed model.

Table 2. Characteristics of the selected descriptors in the best MLR model

Descriptor	Descriptor type	X	Dx	t- value	t-probability	VIF
Constant		-2.801	2.545	-1.1	0.276	
E_{HOMO}	Quantum-chemical descriptors	-1.267	0.245	-5.18	0.000	1.1
Mor02v	3D- MoRSE descriptors	-0.182	0.031	-5.78	0.000	4
G2e	WHIM Index	-17.202	4.131	-4.16	0.000	2.1
HATS7v	GETAWAY descriptors	-9.561	1.492	-6.41	0.000	1.2
RTu+	GETAWAY descriptors	4.762	1.909	2.49	0.000	3.2
AlogP2	Molecular properties	-0.082	0.014	-5.96	0.000	2.1

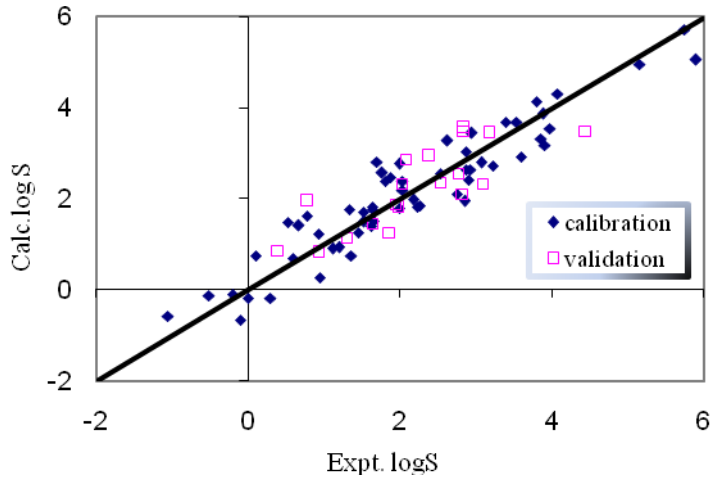


Figure 2. Plot of predicted *vs*. experimental logS for the entire data set.

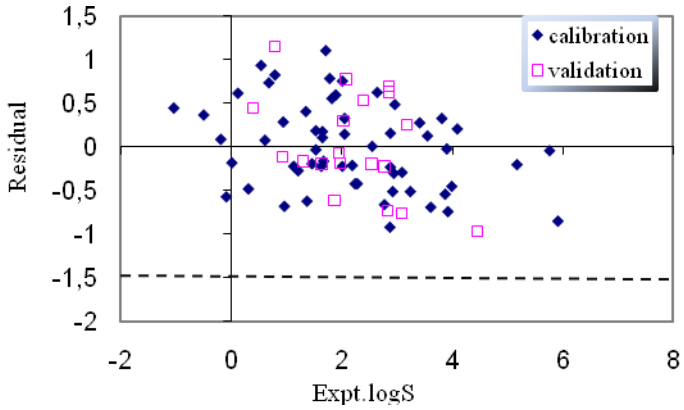


Figure 3.Plot of residual *vs*. experimental logS for the entire data set.

3.2 Descriptor Contribution Analysis and Interpretation

Based on a previously described procedure [35, 36], the relative contribution of the six descriptors to the model were determined and they decrease in the following order: HATS7v (17.91%) > Mor2v (17.67%) > HOMO (16.94%) > AlogP2 (16.80%) > G2e (15.76%) >RTu+ (14.89%). It should be noted that the difference in the descriptor contribution between any two descriptors used in the model is not significant, indicating that all of the descriptors are indispensable in generating the predictive model (Fig.4).

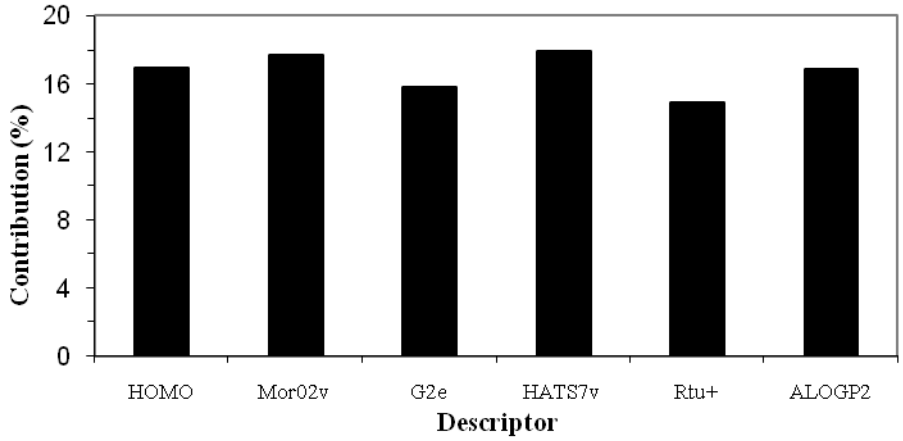


Figure 4.Relative contributions of the selected descriptors to the MLR model.

The importance of atomic van der Waals volumes on the log S values is apparent, since the descriptors weighted by atomic van der Waals explain 35.58% of the contributions (17.91% of HATS7v, and 17.67% of Mor2v). The first important descriptor is HATS7v, which has a relatively high negative correlation with the experimental log S values (R= -0.328). The negative coefficient of HATS7v indicates that the agrochemicals with larger values for this descriptor would have lower log S values.
The second important descriptor is Mor02v, a 3D- MoRSE descriptor, which has a smaller negative correlation coefficient with the experimental log S values (R= -0.787). 3D-MoRSE descriptors are the 3D molecular representations of structure based on electron diffraction descriptor [25, 26], which are calculated by summing atomic weights viewed by a different angular scattering function. The values of these descriptor functions are calculated at 32 evenly distributed values of scattering angle (s) in the range of 0- 31A° from the three dimensional atomic coordinates of a molecule. The 3D- MoRSE descriptor is calculated using following expression:

$$\text{Morsw} = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i w_j (\sin(s.r_{ij}) / s.r_{ij}) \quad (3)$$

where s is the scattering angle, nAT is the number of atoms, r_{ij} is the interatomic distance between the i^{th} and the j^{th} atoms, w is an atomic property, including atomic number, masses, van der Waals volumes, Sanderson electronegativities, and polarizabilities. The coefficient of Mor02v is negative, indicating that an increase in Mor02v would result in a decrease in log S values.

Hence, as expected, atomic volumes have a specific effect on the log S values: an increase in Mor02v (or in HATS7v) would result in a decrease in log S values.

The Squared-Ghose- Crippen-Viswanadhan octanol-water partition coefficient (AlogP2) [31, 32] is calculated from a regression equation based on the hydrophobic character of the molecule. It reflects both the interactions of the solute with the bulk of the surrounding solvent (macroscopic or non specific solvent effects) and the specific bonding between the solute and individual solvent molecules (microscopic or specific solvent effects). When this descriptor increases, the log S decreases.

Highest occupied molecular orbital energy (E_{HOMO}) is a measure of the nucleophicity of a molecule. It should explain the differences in the tendency of solutes to take part in the charge transfer interactions, i. e. the ability of electron- donating to water molecules of solute molecules. According to the Koopmans theorem [37], the energy of the HOMO is directly related to the ionization potential IP (- E_{HOMO} = IP), provided that the ionization process is adequately represented by the removal of an electron from an orbital without change in the wave functions of the other electrons. The descriptor and its coefficient in the model are negative, so the contribution of E_{HOMO} is positive.

The importance of the axial shape and symmetry of the molecule on the log S values is apparent due to the presence of G2e. In the calculations Sanderson atomic electronegativity was used for each atom because it may determine, with other atomic properties, the macroscopic properties of a compound. The positive sign of G2e means that the increase in this descriptor decreases the log S.

RTu+, as HATS7v, is a GETAWAY descriptor and correlates with the experimental log S values of 0.490. The GETAWAY descriptors [29, 30] have been proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular influence matrix. These descriptors, as based on spatial autocorrelation, encode information on molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties.

HATS7v and RTu+ are calculated by Equations. (4) and (5) respectively.

$$\text{HATSk(w)} = \sum_{i=1}^{A} \sum_{j \neq i}^{A} (w_i.h_i)(w_j.h_j)\delta(d_{ij};k)$$

for k=0,1,2,3,…D (4)

$$\text{RTu+} = \text{maxij}\left(\frac{\sqrt{h_i h_j}}{r_{ij}}.w_i.w_j.\delta(d_{ij};k)\right)$$

$i \neq j$ k= 0, 1, 2,3,… D (5)

where A is the number of atoms, w is an atomic weighting scheme, d_{ij} is the topological distance, δ (k, d_{ij}) is a Dirac- delta function (δ=1 if d_{ij}=k, zero otherwise), r_{ij} is the interatomic distance. D is the molecule topological diameter that is the maximum topological distance in the molecule.The coefficient of RTu+ is positive,

meaning that the pesticides with larger values for this descriptor have larger log S values.

3.3 Applicability Domain of the MLR Model

Before a QSPR model is put into use for screening compounds, its applicability domain must be defined and predictions for only those compounds that fall in this domain can be considered as reliable.

The AD of the MLR model was analyzed in the Williams plot (shown in Fig.5). Clearly observation 35 of the training set with leverage higher than the warning limit of 0.36 is a structurally influential compound. Deleting observation 35 could alter slightly R^2 between the experimental logS values and the selected descriptors to 0.8866 (Q^2 = 0.8485) and increase the standard error to 0.524, while utilization of a higer energy conformation geometry for this observation alter negatively the calculated model.

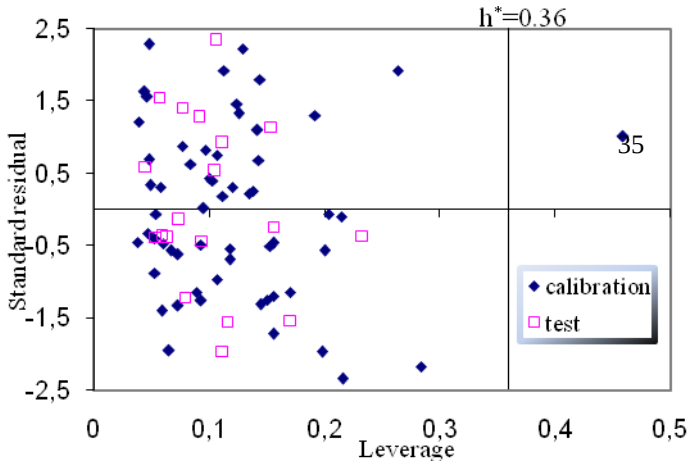


Figure 5. Williams plot of the MLR model for the entire data set.

4. CONCLUSION

In this paper, the QSPR method was applied to the prediction of the aqueous solubility of various type of pesticides. A six- parameter linear model was developed by hybrid GA/ MLR approach with R^2 of 88.95 and s of 0.52 log unit for the training set. The selected descriptors express many factors influencing aqueous solubility, to name: molecular size and shape, specific atomic properties, both macroscopic and microscopic effects and tendency of solutes to take part in the charge transfer interactions. Several validation techniques, including leave-one-out cross-validation and bootstrap, randomization tests, and validation through the test set, illustrated the reliability of the proposed model. All of the descriptors can be directly calculated from the molecular structure of the compound, thus the proposed model is predictive and could be used to estimate the solubility of pesticides. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" chemical structures.

REFERENCES

[1] Price N R., Watkins R W., 2003. Quantitative structure-activity relationships (QSAR) in predicting the environmental safety of pesticides, *Pestic. Outlook*, Vol. 14, pp. 127- 129.

[2] Stevens J T., Breckenridge C B., 2001. Agricultural chemicals: regulation, risk assessment, and risk management, in Regulatory Toxicology. Ed. By S. C. Gad (Taylor & Francis Ltd., London,). 215 p.

[3] Mackay D., 2000. In Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences. Ed. By R. S. Boethling and D. Mackay (CRC Press LLC, Boca Raton) 205 pages.

[4] Lipinski C A., Lombardo F., Doming D W., Feeney P J., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliver. Rev*, Vol. 64, pp. 3- 25

[5] Jorgensen W L., Duffy E M., 2002. Prediction of drug solubility from structure. *Adv. Drug Deliver. Rev*, Vol. 54, pp. 355 – 366

[6] Kartizky A R., Kuanar M., Slavov S., Hall C. D., 2010. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem.Rev*, Vol. 110, pp. 5714 – 5789

[7] Skyner R E., McDonagh J L., Groom C R., van Mourik T., Mitchell J B O., 2015. A review of methods for the calculation of solution free energies and the modeling of systems in solution. *Phys. Chem. Chem. Phys*, Vol. 17, pp. 6174- 6191.

[8] Ruelle P., Kesselring U W., 1997. Aqueous solubility prediction of environmentally important chemicals from the mobile order thermodynamics. *Chemosphere*, Vol. 34, pp. 275- 298.

[9] Deeb O., Goodarzi M., 2010. Predicting the solubility of pesticides compounds in water using QSPR methods. *Molecular Physics*, Vol. 108, pp. 181- 192.

[10] Ran Y., He Y., Yang G., Johnson J L H., Yalkowsky S H., 2002. Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere*, Vol. 48, pp. 487- 509.

[11] Todeschini R., Consonni V., Mauri A., Pavan M., 2005. DRAGON Software – version 5.4-TALETE srl

[12] Hansen O C., 2004. Quantitative Structure-Activity Relationships (QSAR) and Pesticides. (Pesticides Research No. 94. The Danish Environmental Protection Agency,), http://www2.mst.dk/udgiv/publications/2004/87-7614-434-8/pdf/87-7614-435-6.pdf. (accessed 26-05-2014)

[13] Hyperchem™. Release 6.02 for windows. 2000. Molecular Modeling system

[14] Snee R D., 1977. Validation of Regression Models: Methods and Examples, *Technometrics*, Vol. 19, pp. 415-428

[15] Todeschni R., Ballabio D., Consonni V., Mauri A., Pavan M., 2009. MOBYDIGS – version 1.1 – Copyright TALETE srl (2004).

[16] Leardi R., Boggia R., Tarrile M., 1992. Genetic Algorithm as a Strategy for Feature Selection, *J. Chemom*, Vol. 6, pp. 267 – 281

[17] Xu J., Zhang H., Wang Lei., Liang G., Wang Luoxin., Shen X., Xu W., 2010. QSPR study of absorption maxima of organic dye- sensitized solar cells based on 3D descriptors. *Spectrochimica Acta Part A*, Vol. 76, pp. 239-247.

[18] Todeschini R., Maiocchi A., Consonni V., 1999. The K Correlation Index: Theory Development and its Application in Chemometrics. *Chemom*, Int. Lab. Syst, Vol.46, pp. 13 – 29

[19] Tropsha A., Gramatica P., Gombar V K., 2003. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR Comb. Sc*i, Vol. 22, pp. 69 – 77

[20] Shen M., Béguin C., Golbraikh A., Stables J P., Kohn H., Tropsha A., 2004. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds, *J. Med. Chem*, Vol. 47, pp. 2356 – 2364

[21] Weisberg S., 2005. Applied Linear Regression, 3rd edn. (John Wiley and sons, Inc., New Jersey,)
[22] SCAN- Software for Chemometric Analysis- 1995. version 1.1- for Windows, Minitab USA.

[23] Clare B W., 1994. Frontier orbital energies in quantitative structure- activity relationships: a comparison of quantum chemical methods, *Theor. Chim. Acta*, Vol. 87, pp. 415 – 430

[24] Huang Q G., Kong I., Wang L S., 1996. Applications of Frontier molecular orbital energies in QSAR studies, *Bull. Environ. Contam.* Toxicol, Vol.56, pp. 758 – 765.

[25] Gasteiger J., Sadowski J., Schuur J., Selzer P., Steinhauer L., Steinhauer V., 1996. Chemical information in 3D space, *J. Chem. Inf. Comput. Sci*, Vol. 36, pp. 1030 – 1037.

[26] Schuur J., Selzer P., 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, *J. Chem. Inf. Comput. Sci*, Vol. 36, pp. 334 – 344.

[27] Todeschini R., Lasagni M., Marengo E., 1994. New Molecular descriptors for 2D- and 3D – structures, *Theory. J. Chemom*, Vol. 8, pp. 263 – 272.

[28] Todeschini R., Gramatica P., Marengo E., Provenzani R., 1995. Weighted holistic invariant molecular descriptors. Part. 2. Theory development and applications on modeling physico- chemical properties of polyaromatic hydrocarbons (PAH), *Chemom. Intell. Lab. Syst*, Vol. 27, pp. 221 – 229.

[29] Consonni V., Todeschini R., Pavan M., 2002. Structure/ response correlations and similarity/ diversity analysis by GETAWAY descriptors. Part I. Theory of the novel 3D molecular descriptors, *J. Chem. Inf. Comput. Sci*, Vol.42, pp. 682 – 692.

[30] Consonni V., Todeschini R., Pavan M., Gramatica P., 2002. Structure/ responsecorrelations and similarity/ diversity analysis by GETAWAY descriptors. Part II. Application of the novel 3D molecular descriptors in QSAR/ QSPR studies, *J. Chem. Inf. Comput. Sci*, Vol. 42, pp. 693 – 705.

[31] Ghose A K., Crippen G M., 1986. Atomic physico-chemical parameters for three- dimensional- structure-directed quantitative structure- activity relationships. I. Partition coefficients as a measure of hydrophobicity, *J. Comput. Chem*, Vol. 7, pp. 565 – 577.

[32] Viswanadhan V N., Reddy M R., Bacquet R J., Erion M D., 1993. Assessment of methods used for predicting lipophilicity : application to nucleopides and nucleoside bases, *J. Comput. Chem*., Vol. 14, pp. 1019 – 1026.

[33] Todeschini R., Consonni V. , 2009. Molecular Descriptors for Chemoinformatics Volumes I & II. (WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009).

[34] Ramsey F. L., Schafer D. W., 2002. The Statistical Sleuth: A Course in Methods of Data Analysis, 2nd edn. (Wadsworth group, USA).

[35] Zheng F., Bayram E., Sumithran S P., Ayers J T., Zhen C G., Schmitt J D., Dwoskim L P., Crooks P A., 2006. Bioorg. Med. Chem, Vol. 14, pp. 3017 – 3037.
[36] Guha R., Jurs P C., 2005. Interpreting Computational Neural Network QSAR Models: A Measure of Descriptor Importance, *J. Chem. Inf. Model*, Vol. 45, pp. 800 – 806.

[37] Koopmans T C., 1933. Ordering of wave functions and eigenenergies to the individual electrons of an atom, *Physica*, Vol. 1, pp. 104-113.