وزارة التعليم العالى و البحث العلمى

Université BADJI Mokhtar – Annaba
BADJI Mokhtar – Annaba University

جـامـعـة بـاجـي مـخـتـار – عـنـابـة

Faculté des Sciences
Département de Chimie

MÉMOIRE

Présenté pour l'obtention du diplôme Magistère en chimie analytique

Par M^me DAOUDI Souheila

Option : Chimie de l'environnement

THÈME

Application de la méthodologie QSAR à un ensemble d'alcools aliphatiques et d'amines

Devant le jury :

Président :	**M. A. DJELLAL**	**M. C**	**U. B. M. Annaba**
Rapporteur :	**M^me. H. LARKEM**	**M. C**	**U. B. M. Annaba**
Examinateurs :			
	M. D. MESSADI	**Professeur**	**U. B. M. Annaba**
	M^me C. BIDJOU-HAIOUR	**M. C**	**U. B. M. Annaba**

Année 2009

Dédicace

A Mes très chers parents qui m'ont donné un magnifique

modèle de labeur et de persévérance,

A Mon mari TATHA

A Ma belle mère,

A Mes sœurs et frères,

A Mes belles sœurs et beaux frères,

A Toute ma famille,

Je dédie ce mémoire.

REMERCIEMENTS

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

Je tiens également à remercier : Madame. H .Larkem, pour la direction de ce travail ;

Monsieur M. A.DJELLAL, pour avoir accepté la présidence de ce jury ;

Madame C. BIDJOU -HAIOUR, pour avoir accepté de participer à ce jury ;

Enfin, je ne saurais ignorer mes collègues de laboratoire et également tous ceux qui par leur présence ou par leur aide m'ont permis de mener à bien ce travail, spécialement Mohamed Lotfi, Imen et Khadidja

Résumé:

La toxicité de l'ensemble d'alcools aliphatiques et d'amines, caractérisée par la concentration d'inhibition a 50 % de la croissance (CIC_{50}) d'une population de Tetrahymena pyriformis, ont été traitées par la méthodologie QSAR, Ainsi le logarithme de l'inverse de CIC_{50} (variable dépendante) a été corrélé avec deux régresseurs significatifs: le coefficient de partage Octanol/eau (logp) et Vee.

La valeur du coefficient de détermination (R^2 = 96.75%) et les coefficients de prédiction (Q^2=95.73%) renseignent, respectivement, sur l'adéquation des modèles et ses bonnes capacités prédictives.

L'analyse des résidus permet d'identifier une observation aberrante (tridecanol).

Mots clés:

Les alcools , les amines aliphatiques - Toxicité-Milieu biologique-Méthodologie QSAR.

ملخص:

التسمم ل 21 كحول خطي, و 9 أمينات خطية, يعرف بتركيز معيقات التفاعل بنسبة 50 ٪ لفصيلة tetrahymena pyriformis ,التي تمت معالجتها بطريقة العلاقات الكمية, بنية /نشاط QSAR, وكذلك اللوغريتم العكسي ل CIC50 الذي هو عبارة على(المتغير المرتبط) .الذي تم إرتباطه مع الصفتان: عدد ذرات الهيدروجين (nH) و مأشر زاغراب الاول(ZM1V) .

ولقد تم الحصول على القيم التالية : قيمة معامل التحديد (R^2=96,14) , اما قيمة معامل التنبؤ هي (Q2=93,14) هذه النتائج تدل على ملائمة النموذج وحسن امكانيته على التنبئى .

اما معالجة البواقي تسمح لنا بتعرف على وجود ملاحظة واحدة خارجة على الحالة العادية وهي المركب الكحولي (tridecanol) .

مفاتيح الكلمات:

كحول خطي, أمينات خطية, التسمم, الوسط البيئي, طريقة العلاقات الكمية بنية /نشاط QSAR.

Résumé:

La toxicité de l'ensemble d'alcools aliphatiques et d'amines, caractérisée par la concentration d'inhibition a 50 % de la croissance (CIC_{50}) d'une population de Tetrahymena pyriformis, ont été traitées par la méthodologie QSAR, Ainsi le logarithme de l'inverse de CIC_{50} (variable dépendante) a été corrélé avec deux régresseurs significatifs: le nombre d'atomes d'hydrogène (nH) et ZM1V.

La valeur du coefficient de détermination (R^2= 98,49 %) et les coefficients de prédiction (Q^2=97,47%) renseignent, respectivement, sur l'adéquation des modèles et ses bonnes capacités prédictives.

L'analyse des résidus permet d'identifier une observation aberrant (tridecanol).

Mots clés:

Les alcools aliphatiques , les amines aliphatiques - Toxicité-Milieu biologique-Méthodologie QSAR.

Abstract:

The toxicity of series of aliphatic alcohols and aliphatic amines, characterized by the concentration of a 50% inhibition of growth (CIC_{50}) a population of Tetrahymena pyriformis, was treated by the QSAR methodology. Thus the inverse logarithm of CIC_{50} (dependent variable) was correlated with two significant regressors: the number of hydrogen atoms (nH) and first Zagreb index (ZM1V) .

The values of the coefficient of determination (R^2 = 98.49%) and coefficients of prediction (Q^2= %97.47) information, respectively, on the adequacy of models and predictive capabilities .

The traitment of the residue, permitus to analysis can identify aberrant observation (tridecanol)

key words:

aliphatic Alcohols, aliphatic amines – Toxicity - Biological Environment- QSAR Methodology

Abstract:

The toxicity of series of aliphatic alcohols and aliphatic amines, characterized by the concentration of a 50% inhibition of growth (CIC_{50}) a population of Tetrahymena pyriformis, was treated by the QSAR methodology. Thus the inverse logarithm of CIC50 (dependent variable) was correlated with two significant regressors: the coefficient of partage Octanol/eau (logp) and Vee.

The values of the coefficient of determination (R^2 = 96.75%) and coefficients of prediction (Q^2=95.73) information, respectively, on the adequacy of models and predictive capabilities .

The analysis of the residue, permit to can identify aberrant observation of (tridecanol)

key words:

aliphatic Alcohols, aliphatic amines – Toxicity - Biological Environment- QSAR Methodology

ملخص:

التسمم ل21 كحول خطي, و 9 أمينات خطية, يعرف بتركيز معيقات التفاعل بنسبة 50 ٪ لفصيلة tetrahymena pyriformis ,التي تمت معالجتها بطريقة العلاقات الكمية, بنية /نشاط QSAR, وكذلك اللوغريتم العكسي لCIC_{50} الذي هو عبارة على(المتغير المرتبط).الذي تم إرتباطه مع الصفتان: (logp) وVee.

ولقد تم الحصول على القيم التالية: قيمة معامل التحديد (R^2=96,75) , اما قيمة معامل التنبؤ هي (Q^2= 95,73) هذه النتائج تدل على ملائمة النموذج وحسن امكانيتة على التنبؤ .

اما معالجة البواقي تسمح لنا بتعرف على وجود ملاحظة واحدة خارجة على الحالة العادية وهي المركب الكحولي (tridecanol) .

مفاتيح الكلمات:

كحولات خطية, أمينات خطية, التسمم, الوسط البيئي, طريقة العلاقات الكمية بنية /نشاط QSAR.

SOMMAIRE

CHAPITRE II

CHAPITRE III

ANNEXE

AM1 : Austin Model 1.

CIC_{50} : Concentration d'inhibition de la croissance à 50 %.

DFFITS : Statistique permettant de mesurer l'influence d'une observation i sur la valeur ajustée.

Di : Distance de COOK., renseigne sur l'influence d'une observation i sur Les coefficients de régression estimés par les moindres carrés.

d : Statistique de Durbin-Watson.

di : Résidu : standardisé.

EQMC: erreur quadratique moyenne sur l'ensemble de calibrage.

EQMP*:* erreur quadratique moyenne sur l'ensemble de prédiction.

$EQMP_{ext.}$: erreur quadratique moyenne sur l'ensemble de prédiction externe.

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

F : Statistique de Fisher.

$\underset{\sim}{\mathbf{H}}$: Matrice de projection, ou matrice châpeau.

hii : Eléments diagonaux de la matrice chapeau.

LOO: Cross-validation by leave-one-out: Validation croisée par omission

logP ou ($\log k_{ow}$) : Coefficient de partage octanol /eau.

n: Dimension de la population .

n-p : Nombre de degrés de liberté de la somme des carrés des résidus.

p : Nombre de descripteurs en comptant la constante.

PRESS : Somme des carrés des erreurs de prédiction.

$pCIC_{50}$: log 1/(CIC50).

Q^2 : Coefficient de prédiction.

QSAR : Quantitative Structure/ Activity Relationships.

R^2 : Coefficient de détermination.

r_i : Résidu studentisé interne.

S : Erreur standard.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

t : t de Student.

t_i : Résidu studentisé externe.

ri Résidu studentisé interne.

$\underset{\sim}{\mathbf{X}}$:	Matrice des valeurs observées des variables explicatives.
$\underset{\sim}{\mathbf{X}}'$:	Matrice transposée de $\underset{\sim}{X}$.
x_j :	$j^{ième}$ valeur .
$\underset{-}{x_j}$:	Variable explicative.
x_{norm} :	Valeur normalisée.
$x_{\max}$:	Valeur maximale.
$x_{\min}$:	Valeur minimale.
$\underset{-}{y}$:	Vecteur de dimension n.
yi:	Valeur observée.
$\hat{y}_i$:	Valeur estimée.

INTRODUCTION :

Les alcools et les amines sont des composés organiques qui sont largement utilisés dans le milieu industriel et dans la vie quotidienne. Que ce soit professionnellement ou à la maison, ces deux familles de produits sont incontournables. Les alcools sont trouvés à l'état pur ou en mélange dans des préparations spécifiques. Ils sont utilisés comme diluants des encres d'imprimerie, des résines, des vernis, des peintures et des colles à moquette. Ils sont aussi largement utilisés comme excipients pour les produits pharmaceutiques ou cosmétiques ou comme milieu réactionnel dans l'industrie chimique [1]. Les amines sont utilisées comme intermédiaires chimiques pour la synthèse de produits pharmaceutiques, cosmétiques, détergents…. . Elles sont aussi utilisées comme solvants et inhibiteurs de corrosion [2].

Les alcools ont des effets néfastes bien connus, entrainant notamment des incoordinations motrices ou une excitation intellectuelle. Les alcools liquides et leurs vapeurs sont irritants pour la peau, les yeux et les muqueuses en cas de contact prolongé ou répété. L'inhalation accidentelle d'une grande quantité de vapeurs d'alcool peut conduire à des syndromes ébrieux ou narcotiques avec nausées, malaises, vomissements et maux de tête [1]. Quant aux effets indésirables des amines, les problèmes de santé pouvant se développer chez les travailleurs surexposés sont divers, allant de l'irritation cutanée au cancer [3].

Le transport, la distribution, l'accumulation et l'absorption des xénobiotiques (médicaments ou toxiques), pour certains leur fixation sur les protéines plasmatiques, leur passage à travers les membranes, leur entrée dans les cellules, les interactions enzyme-inhibiteur ou ligand-récepteur de nature hydrophobe, l'activité pharmacologique et pharmacocinétique des médicaments, la toxicité des médicaments ou des contaminants,…, les propriétés liées à la formulation telles que solubilité, sont autant d'exemples dans lesquels la lipophilie des molécules constitue un descripteur physico-chimique de la première importance. La lipophilie d'une molécule est mesurée classiquement par son aptitude à se distribuer dans un système biphasique soit liquide-liquide soit solide-liquide. Le coefficient de partage dans le système *n*-octanol/eau est connu depuis longtemps comme étant un des paramètres physico-chimiques quantitatifs qui est le mieux corrélé à l'activité des molécules organiques [4].

L’élaboration de relations structure-activité quantitatives (QSAR) étant devenue cruciale dans le développement de substances chimiques organiques, que ce soit dans le domaine de la pharmacie ou dans celui de l’agrochimie, les logarithmes des coefficients de partage (log *P*) ont trouvé une utilisation extensive dans la conception de médicaments ou de pesticides. Ils entrent ainsi fréquemment dans l’élaboration des QSAR. De plus, le coefficient de partage *n*-octanol/eau joue un rôle important dans les stades précoces de l’évaluation des risques engendrés par un produit chimique sur l’environnement [4].

Dans toutes les techniques QSAR, les applications pratiques nécessitent au départ un ensemble de composés d'activités connues : les données d'estimation. Cet ensemble, dit de calibration, est utilisé dans une première étape pour le calcul d'un modèle prédictif. Ce qui conduit à des équations mathématiques simples associant, aussi bien que possible, les descripteurs (ou régresseurs) et les propriétés mesurées (observations ou encore variables dépendantes). Si l'ensemble de calibration constitue un échantillon représentatif de la population, on admet alors que l'introduction de nouveaux éléments dont la valeur de la variable dépendante est inconnue (qu'on désigne encore par données de prédiction), n'affectera pas la stabilité du modèle, et que l'on peut ainsi estimer ces valeurs manquantes avec suffisamment de confiance. Les études QSAR courantes en toxicologie mettent en jeu des descripteurs 2D comme le coefficient de partage octanol / eau (log *P*) ou d'autres paramètres, qui simulent les différentes interactions moléculaires. Il n’est donc pas surprenant de construire des modèles QSAR satisfaisants en faisant intervenir le coefficient de partage octanol / eau (log *P*). Pour améliorer les capacités prédictives de ces modèles, on y incorpore généralement d’autres descripteurs 2D [5].

Dans ce travail, nous nous sommes intéressés à la toxicité des alcools aliphatiques et des amines caractérisée par une concentration d’inhibition à 50 % de la croissance (CIC_{50}) en utilisant le calcul d'un modèle de régression linéaire et la validation externe.

1 LES ALCOOLS ALIPHATIQUES

Les alcools sont des solvants couramment utilisés aussi bien dans l'industrie que dans les foyers domestiques. Ils peuvent être à l'origine d'intoxications aiguës accidentelles ou volontaires parfois gravissimes. A terme, l'exposition répétée par voie pulmonaire ou contact cutané peut être à l'origine d'effets toxiques systémiques variés [6].

1-1 Un peu de chimie :

Les alcools sont préparés industriellement à partir d'hydrocarbures pétroliers dans des usines pétrochimiques. L'un d'eux (l'éthanol) peut être produit à partir de la fermentation naturelle de jus sucrés. Ils sont caractérisés chimiquement par la présence, sur une chaîne hydrocarbonée, d'un ou plusieurs groupements « Alcool » composés d'un atome d'oxygène et d'un atome d'hydrogène (groupement hydroxyle -OH). Les mono-alcools possèdent un groupement -OH. Ils sont normalement désignés par le nom de la chaîne hydrocarbonée auquel on ajoute la terminaison –ol ou par une dénomination alcool, par exemple :

- Méthanol ou alcool méthylique (CH_3OH),
- Éthanol ou alcool éthylique (CH_3-CH_2-OH),
- Isopropanol ou alcool isopropylique (CH_3-CHOH-CH_3) [1].

1-2 Propriétés physico-chimiques :

La grande majorité des alcools utilisés industriellement sont liquides à température ambiante. Ils sont incolores et ont une odeur qui peut être agréable (éthanol), sucrée (cas des diols), acre ou amère (propanol ou alcool furfurylique), ou encore piquante (alcool isoamylique).

Les alcools communément utilisés sont miscibles dans l'eau, totalement pour les molécules les plus courtes (méthanol, éthanol...), partiellement pour les autres. Les alcools sont inflammables ou facilement inflammables. Le point d'éclair des plus utilisés se situe entre 12 et 40 °C. Leurs vapeurs peuvent former des mélanges explosifs avec l'air.

Les alcools sont très volatils, leur diffusion dans le milieu ambiant ou dans l'atmosphère sera très importante. Ils dissolvent les graisses et certaines matières plastiques. Tous les alcools sont des liquides déshydratants [1].

1-3 Utilisation :

Les alcools sont très utilisés comme diluants des encres d'imprimerie, des résines, des vernis, des peintures et des colles à moquette. Ce sont d'excellents agents déshydratants possédant une bonne action dégraissante, ils sont donc utilisés comme agents de séchage en mécanique ou en optique et pour les nettoyages difficiles (encres, silicones…). Ils sont aussi largement utilisés comme excipients pour les produits pharmaceutiques ou cosmétiques ou comme milieu réactionnel dans l'industrie chimique [1].

1-4 Dangers et risques :

1-4-1 Toxicité :

De tous les alcools, le plus toxique est le méthanol dans la mesure où il exerce une action sélective au niveau du nerf optique, pouvant provoquer la cécité ou la mort. Les effets néfastes de l'absorption d'éthanol sont aussi bien connus, l'alcoolémie entraînant notamment des incoordinations motrices ou une excitation intellectuelle. De manière générale, les manifestations d'une intoxication modérée se traduiront par des maux de tête, des troubles digestifs et un syndrome ébrieux.

Les alcools liquides et leurs vapeurs sont irritants pour la peau, les yeux et les muqueuses en cas de contact prolongé ou répété.

L'alcool furfurylique, plus agressif que les autres alcools, peut provoquer des larmoiements à de très faibles expositions (15 ppm) et des irritations respiratoires.

L'inhalation accidentelle d'une grande quantité de vapeurs d'alcool peut conduire à des syndromes ébrieux ou narcotiques avec nausées, malaises, vomissements et maux de tête.

1-4-2 Absorption et métabolisme:

L'alcool est un toxique systématique dont la toxicité dépend de son métabolisme.il peut être absorbé par voie orale, cutanée et par inhalation. Ainsi, il se distribue uniformément dans les tissus et organes.

De tous les alcools, le méthanol diffuse rapidement et complètement dans l'eau totale de l'organisme. Après absorption, le méthanol est oxydé initialement en formaldéhyde soit par l'alcool déshydrogénase chez l'homme et les primates, soit par le système catalase-éroxydase chez les autres mammifères. La seconde étape est la transformation du formaldéhyde en acide formique / formate. Enfin, la voie métabolique des composés à un atome de carbone conduit à la production de CO_2 qui est l'étape limitant de la biotransformation du méthanol [6].

1-4-3 Risque incendie et explosion :

Les alcools couramment utilisés sont tous facilement inflammables. À température ambiante, en présence d'une flamme nue, d'une étincelle ou d'une source de chaleur, ils s'enflammeront instantanément.

De même, la présence de vapeurs alcooliques dans l'air (entre 3 et 30 % en volume) créera une atmosphère explosive extrêmement dangereuse.

C'est l'un des problèmes majeurs lors de leur utilisation en tant que solvant ou réactif de synthèse [1].

1-4-4 Réactivité :

Dans des conditions normales de stockage, ce sont des produits relativement stables. Mais ils peuvent réagir violemment notamment avec les oxydants puissants comme les mélanges sulfo-chromiques ou nitro-chromiques, les peroxydes, l'acide nitrique.

L'action du chlore sur un alcool peut produire un composé qui se décompose avec explosion lorsqu'il est exposé à la lumière ou à la chaleur. Les alcools peuvent aussi réagir avec les métaux alcalins (sodium, potassium…) avec dégagement d'hydrogène, gaz extrêmement inflammable [1].

1-4-5 Risque pour l'environnement :

Tous les alcools font partie des COV (Composés Organiques Volatils). Leur émission dans l'atmosphère contribue à augmenter la production d'ozone dans la troposphère par réaction photochimique, augmentant ainsi les risques pour les personnes asthmatiques ou souffrant d'insuffisance respiratoire.

Les alcools sont solubles dans l'eau et rapidement biodégradables. Leur rejet massif à l'égout peut cependant contribuer sensiblement à la détérioration de la faune et la flore peuplant les fleuves et les rivières [1].

5 -Protection :

1-5-1 Protection individuelle :

Toute manipulation manuelle d'alcools ou de préparations en contenant doit s'assortir des précautions suivantes :

- Éviter l'inhalation des vapeurs. Pour des travaux exceptionnels de courte durée dans des atmosphères polluées par des vapeurs d'alcools ou en cas d'urgence, il est nécessaire de porter des appareils de protection respiratoire.
- En cas d'utilisation de masque à cartouche, le type de filtre à utiliser est désigné par le marquage A1, A2 ou A3 (le chiffre représentant la capacité de piégeage) accompagné d'une bande de couleur marron.
- Éviter le contact cutané. Dès lors qu'il y a probabilité de contact avec la main, il s'avère indispensable de porter des gants de protection appropriés à la tâche effectuée et au produit manipulé [1].

1-5-2 Protection de l'environnement :

Les rejets atmosphériques de vapeurs d'alcools sont fortement limités et réglementés dans le cadre d'une directive appelée directive COV (1999/13/CE).

Les alcools seront donc préférentiellement utilisés en circuit fermé afin d'éviter toute vaporisation dans l'atmosphère et respecter ainsi les valeurs d'émission établies par la directive.

De nombreux alcools « usés » peuvent être régénérés par distillation et réutilisés.

La destruction des alcools est effectuée par incinération par des sociétés spécialisées dans le traitement des déchets industriels [1].

2-LES AMINES ALIPHATIQUES

2 -1-Histoire:

Découvertes par Wurtz en 1849, les amines furent initialement appelées alcaloïdes artificiels. Ce sont des composés organiques azotés, dérivant de l'ammoniac par remplacement d'un, de deux ou de trois atomes d'hydrogène par autant de groupes hydrocarbonés (alkyl, aryl) et désignés respectivement par amine primaire, secondaire ou tertiaire. Les arylamines, dont l'aniline est le représentant le plus simple, ont une importance pratique considérable pour la synthèse de médicaments, de colorants et d'autres dérivés aromatiques ; les alkylamines n'ont que des débouchés limités dans l'industrie pharmaceutique. Par contre, plusieurs diamines aliphatiques sont à la base de fibres textiles de haute qualité (nylons). La classe de l'amine est relative au degré de substitution de l'azote et non (comme dans le cas des alcools) à celui du carbone qui porte la fonction. Si la fonction amine est fréquente dans de nombreuses espèces d'importance biologique fondamentale, il s'agit presque toujours de composés à fonction mixte, les plus importants étant les acides aminés et les alcaloïdes. En revanche, les amines naturelles, à fonction simple, sont exceptionnelles ; on peut citer la présence de triméthylamine $(CH_3)_3N$ parmi les produits de putréfaction de la chair des poissons, celle de l'ion tétraméthylammonium $(CH_3)_4N^+$ dans les vinasses de betteraves. (...)[7].

2-2 Nomenclature :

Si le groupement amine est prioritaire, la molécule comprend le suffixe -amine. Sinon, elle possède le préfixe amino- [3].

2-3Synthèses :

Typiquement, les amines sont obtenues par alkylation d'amines de rang inférieur. En alkylant l'ammoniac, on obtient des amines primaires, qui peuvent être alkylées en amines secondaires puis amines tertiaires. L'alkylation de ces dernières permet d'obtenir des sels d'ammonium quaternaire.

D’autres méthodes existent :

- Les amines primaires peuvent être obtenues par réduction d'un groupement azoture.

- Les amines peuvent aussi être obtenues par la réduction d'un amide, à l'aide d'un hydrure.
- L'amination réductrice permet l'obtention d'amines substituées à partir de composés carbonylés (aldéhydes ou cétones) [3].

2-4 Propriétés :

La présence de l'atome d'azote est la cause des propriétés des amines. Cet atome présente un doublet non liant, ce qui donne aux amines un caractère basique et nucléophile. Dans le cas d'amine primaire et secondaire, la liaison N-H peut se rompre, ce qui leur donne un (faible) caractère acide. Les amines sont volatiles, ont une odeur forte et sont hydrosolubles [3].

2-5 Réactivité :

En général les amines tertiaires réagissent aisément avec la plupart des dérivés aromatiques. Par exemple en introduisant du naphtalène dans une solution saturée d'orthophénantroline on obtient le tétrahydrocannabinol plus connu sous l'abréviation THC. Cette réaction est catalysée par du paladium en milieu acide. Cette réaction est connue sous le nom de la réaction de Roucoux-Crévisy [3].

2-6 Utilisation:

Les amines sont utilisées en grande quantité dans des procédés industriels variés, comme un intermédiaire réactionnel, dans des mélanges ou comme solvant. On les emploie surtout dans l'industrie chimique, dans celle des polymères et du caoutchouc, en agriculture comme pesticides, de même que dans la composition de peintures, d'adhésifs et de textiles, ainsi que dans l'industrie pharmaceutique [2].

2-7 Toxicité:

Les amines sont un produit irritant et corrosif pour la peau, les yeux, les voies respiratoires et digestives. La gravité des symptômes peut varier selon les conditions d'exposition (durée de contact, concentration du produit, etc.).

L'exposition aux brouillards peut causer une irritation des yeux, de la peau et des voies respiratoires. L'exposition à de fortes concentrations peut provoquer l'œdème pulmonaire. Les symptômes de l'œdème pulmonaire (principalement toux et difficultés respiratoires) se manifestent souvent après un délai pouvant aller jusqu'à 48 heures. L'effort physique peut aggraver ces symptômes. Le repos et la surveillance médicale sont par conséquents essentiels [8].

2-8 Métabolisme :

Ces substances sont généralement bien absorbées par toutes les voies. Elles peuvent être en partie oxydées in vivo par les enzymes monoamine oxydase et diamine oxydase. Les amines tertiaires sont en partie oxydées en dérivé N-oxydes par des mono-oxygénases dépendant de la flavine ou déalkylées en amines secondaires (Bickel). Dans le cas de la diméthyléthylamine le principal métabolite urinaire (outre la substance inchangée) sera le diméthylamine-N-oxyde.

Certaines amines (par exemple diéthylamine, diméthylamine) sont surtout excrétées inchangées par voie urinaire, d'autres après biotransformation partielle. Certaines amines aliphatiques comme par exemple la diéthanolamine s'accumulent notamment dans le foie, le rein, la rate et le cerveau. En ce qui concerne la diéthanolamine on a aussi constaté que la substance mère et ses métabolites la N-méthyl-et la N-diméthyléthanolamine pouvaient être incorporés dans les phospholipides .Leur toxicité chronique pourrait donc en partie résulter de la formation de phospholipides anormaux [2].

Il est généralement admis que la toxicité de nombreuses substances, particulièrement les produits chimiques organiques industriels, est la conséquence de leur solubilité dans les lipides, alors que leurs caractéristiques moléculaires spécifiques ont peu ou pas d'influence. Leur mode d'action consisterait en la destruction des processus physiologiques associés aux membranes cellulaires.

Les protozoaires sont souvent utilisés pour l'évaluation de la toxicité. Les méthodes mises en œuvre sont basées sur des critères morphologiques, ultra-structuraux, éthologiques et métaboliques [9].

L'inhibition de la croissance d'une population est un indicateur très en vogue, parce qu'il peut être déterminé directement ou indirectement à l'aide d'un équipement électronique. Ce qui permet l'acquisition rapide des observations nécessaires pour les analyses de régression. Nous considérerons la concentration d'inhibition à 50% de la croissance (CIC_{50}), dont le logarithme de l'inverse, soit $pCIC_{50} = \log(CIC_{50})^{-1}$, servira d'indicateur de toxicité.

II – 1 – Collecte des données

Les tests de toxicité ont été réalisés en examinant la croissance d'une population de Tetrahymena pyriformis. Les essais ont été menés dans des erlenmeyers de 250 ml, contenant 50 ml d'un milieu dont la composition est précisée ci après :

Eau distillée	1000 mL
Protéose peptone	20 g
D-glucose	5 g
extrait de levure	1 g
FeEDTA	1 mL d'une solution à3 % (masse/v)
pH	7,35

La température a été fixée à 27 ± 1°C.

Ce milieu est inoculé avec 0,25 ml d'une culture contenant approximativement 36 000 cellules par ml. La croissance des ciliés est suivie par spectrophotométrie, en mesurant la densité optique (absorbance) à 540 mm après 48 heures d'incubation.

Plusieurs critères ont guidé au choix des composés toxiques examinés. Tous sont disponibles dans le commerce avec une pureté suffisante (95 % et plus), ce qui ne nécessite pas une re-purification préalablement au test. Des précautions ont été observées afin d'assurer une diversité concernant, à la fois, les propriétés physico-chimiques et la position des substituants.

Les solutions stocks des divers composés toxiques, ont été préparées dans le diméthylsulfoxide (DMSO) à des concentrations de 5, 10, 25 et 50 grammes par litre. Dans chaque cas, le volume de solution stock ajouté à chaque fiole est limité par la concentration finale de DMSO qui ne doit pas excéder 0,75 % (350 ml par fiole), quantité qui n'altère pas la reproduction de Tétrahymena.

II – 2 – Méthode de calculs :

Les relations quantitatives structure / activité (QSAR) ont été déterminées en prenant pour variable dépendante le logarithme de l'inverse de CIC_{50} (en mmol / litre)**,** et deux variables explicatives le coefficient de partage Octanol/eau et descripteur quantique électronique(Vee).

II-2-1 Optimisation de la géométrie moléculaire :

Les structures des molécules ont été obtenues à l'aide du logiciel de modélisation moléculaire Hyperchem 7.5 [10], et les géométries finales à l'aide de la méthode semi empirique AM1 du même logiciel. Tous les calculs ont été menés dans le cadre du formalisme RHF sans interaction de configuration. Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak-Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,001 $kcal.mol^{-1}$.

II – 2 – 2 Calcul des descripteurs moléculaires :

Les géométries obtenues ont été transférées dans les logiciels informatiques [10-11-12] à savoir :

- L'HyperChem [10] utilisé pour le calcul des descripteurs électroniques.
- Dragon [11] utilisé pour le calcul de plus de 1600 descripteurs appartenant à 20 classes différentes.
- Le logiciel Ecalc [12] est muni d'une interface graphique qui permet à l'utilisateur d'introduire les molécules, puis de calculer les indices électrotopologiques.

II – 2 – 3 A**nalyse de régression :**

Les données ont été modelées en utilisant la régression linéaire multiple (MLR) de Minitab [13] pour calculer les résidus, vérifier les observations aberrantes ainsi que celles influentes et vérifier la qualité de l'ajustement des modèles.

Les valeurs des données utilisées dans la suite de ce travail sont condensées dans le tableau I.

Tableau I –Toxicité relative, hydrophobicité et valeurs de descripteur Vee pour les alcools aliphatiques et des amines.

N°	Composés	Numéro de CAS	$pCIC_{50}$	log P	‹ Vee ›
1	methanol	67-56-1	-2,77	-0.77	81.07
2	éthanol	64-17-5	-2,41	-0.31	135.21
3	Propan-1-ol	71-23-8	-1,84	0.25	196.33
4	Butan-1-ol	71-36-3	-1,52	0.88	263.38
5	pentan1-ol	71-41-0	-1,12	1.56	335.28
6	Hexan-1-ol	111-27-3	-0,47	2.03	411.32
7	Heptan-1-ol	111-70-6	0,02	2.57	490.92
8	Octan-1-ol	111-87-5	0,5	3.15	573.71
9	Nonan-1-ol	143-08-8	0,77	3.69	659.30
10	Decan-1-ol	112-30-1	1,1	4.23	747.48
11	Undecan-1-ol	112-42-5	1,87	4.77	837.97
12	Dodecan-1-ol	112-53-8	2,07	5.13	930.62
13	Tridecan-1-ol	112-70-9	2,28	5.67	1121.73
14	Propan-2-ol	67-63-0	-1,99	0.05	200.66
15	Pentan-2-ol	6032-29-7	-1,25	1.21	347.32
16	Pentan-3-ol	584-02-1	-1,33	1.21	350.49
17	2-methyl-1-butanol	137-32-6	-1,13	1.42	350.05
18	3-methyl-1-butanol	123-51-6	-1,13	1.42	346.51
19	3-methyl-2-butanol	598-75-4	-1,08	1.28	356.16
20	(tert) pentanol	75-85-4	-1,27	1.21	359.82
21	(neo) pentanol	75-84-3	-0,96	1.57	359.80
22	1-propylamine	107-10-8	-0,85	0.48	187.94
23	1-butylamine	109-73-9	-0,7	0.97	254.77
24	1-amylamine	110-58-7	-0,61	1.49	328.44
25	1-hexylamine	111-26-2	-0,34	2.06	405.22
26	1-heptylamine	111-68-2	0,1	2.57	485.70
27	1-octylamine	111-86-4	0,51	3.04	569.45
28	1-nonylamine	112-20-9	1,59	3.57	656.12
29	1-decylamine	2016-57-0	1,95	4.10	745.42
30	1-undecylamine	7307-55-3	2,26	4.63	837.10

III -1 Calcul du modèle:

Avant de procéder au développement effectif des équations de régression, nous avons vérifié la qualité statistique des variables dépendantes et explicatives.

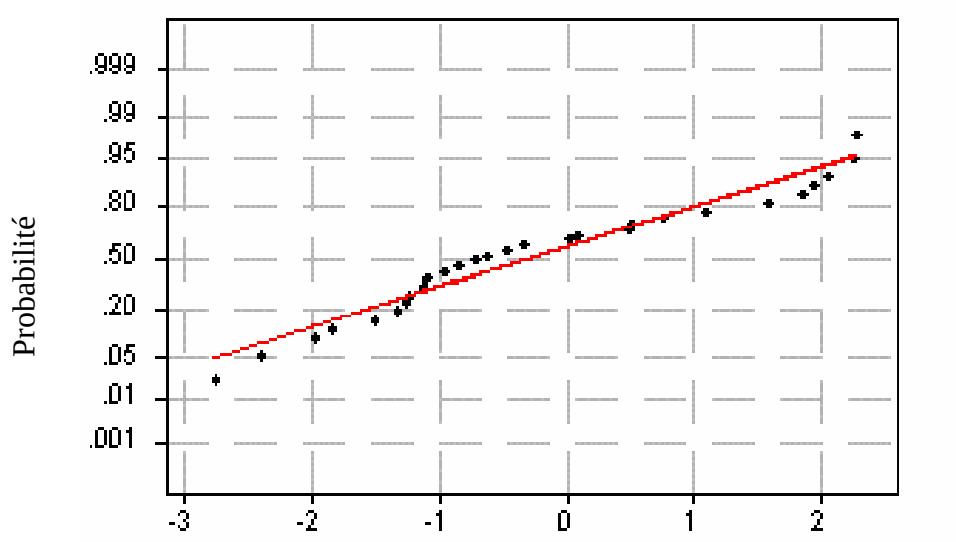
Quelques résultats partiels seront présentés à titre d'illustration, en exploitant les données qui apparaissent dans le tableau I.

Les diagrammes de probabilité obtenus à partir de ces données (figures 2, 3 et 4) montrent que les variables considérées ne se distribuent pas toujours selon la loi normale, puisque les R^2 obtenus ne sont pas systématiquement supérieurs aux R^2 critiques (R^2c) donnés par les tables (Tableau II pour n = 30 et 20 composés).

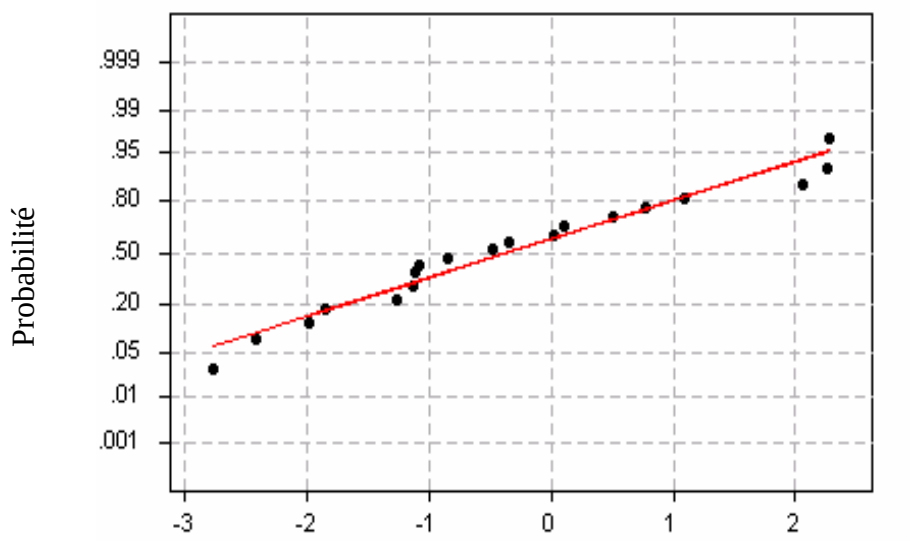
Tableau II - les valeurs de R^2 (%) pour test de normalité

		$pCIC_{50}$	logP	Vee		α =0.01	α =0.05
pour n=30	**R^2 (%)**	97.43	98.25	96,70	**R^2 c (%)**	94.9	96.39
pour n=20	**R^2 (%)**	98.05	98.69	96,42	**R^2 c (%)**	92.9	95.03

Test de normalité a été vérifiée pour les niveaux α = 0,01et α = 0,05 après élimination de 10 composes : octan-1-ol ; butan-1-ol ; undecan-1-ol ; pentan-2-ol ; pentan-3-ol; (neo) pentanol; 1-butylamine ;amylamine ; nonylamine ; decylamine

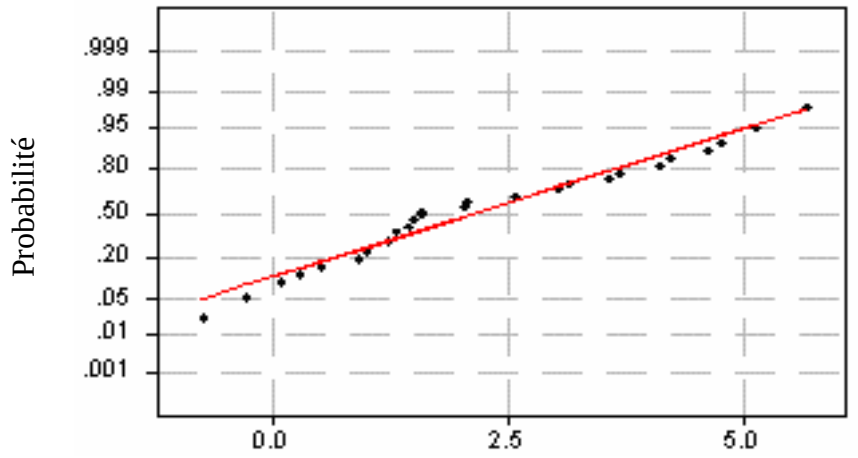


$pCIC_{50}$ (n=30)

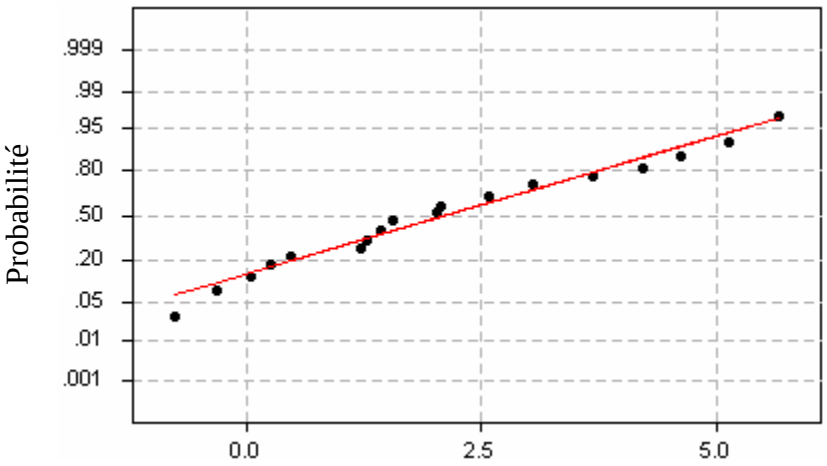


$pCIC_{50}$ (n=20)

Figure 1 Diagramme de test de normalité pour $pCIC_{50}$

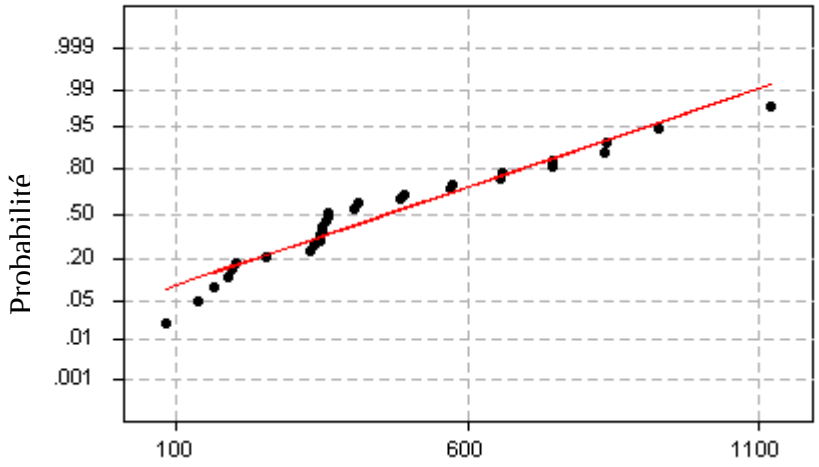


Log*P* (n=30)

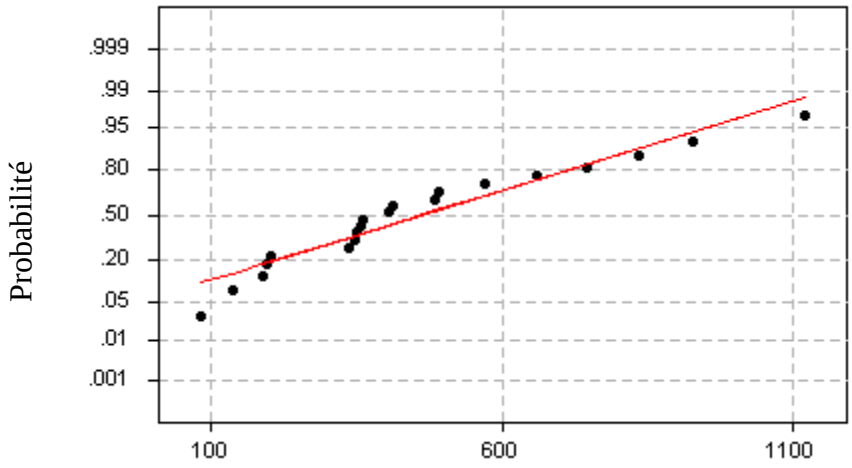


log*P* (n=20)

Figure 2 Diagramme de test de normalité pour log*P*



Vee (n=30)



Vee (n=20)

Figure 3 Diagramme de test de normalité pour Vee

Parmi les modèles optimaux générés, celui qui fournit la valeur maximale pour les paramètres statistiques Q^2, R^2 et Q^2_{ext} .comporte les 2 descripteurs calculés par le logiciel HYPERCHEM, dont les symboles, la classe et la signification sont réunis dans le tableau III.

Tableau III - Descripteurs moléculaires intervenant dans la modélisation de la $pCIC_{50}$

N°	Descripteur	Classe	Signification
1	log*P*	Proprietés moléculaire	le coefficient de portage Octanol/eau
2	Vee	Proprietés électroniques	descripteur quantique électronique

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

$$pCIC_{50} = - 1.87 + 1.03 \log P - 0.00139 \langle Vee \rangle \qquad (1)$$

$$n = 20 \;\; ; \;\; \sigma_N = 0.352 \; ; \;\; R^2 = 96.75\% \;\; ; \;\; Q^2 = 95.73\% \;\; ; \;\; F = 253.23$$

III – 2 Analyse de régression:

Les valeurs des paramètres statistiques montrent que les deux descripteurs (logp, Vee) (tableau -II) permettent de corréler la $pcic_{50}$ des 20 composés.

En effet, la valeur du cœfficient de détermination (R^2) signifie que 96.75%de la variabilité de $pCIC_{50}$ peut être expliquée par ces deux descripteurs, alors que la racine de l'erreur quadratique moyenne de prédiction est de l'ordre de 2 ($\sigma_N = 0.352$) ; en outre ce modèle est très hautement significatif (grande valeur du paramètre de Fischer : F=253.23).

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

L'analyse des résidus (colonne 1 tableau IV). Tous les résidus ordinaires e_i sont inférieur à 2 fois l'erreur standard ($|e_i| < 2S$), soit 2 x 0.327 = 0.654.

Tous les résidus standardisés d_i de la colonne (2) sont compris entre les limites ± 2, à l'exception de point 16.

La colonne (3) rassemble les résidus studentisés internes r_i . On a ici p = 3 et n = 20, et on constate que tous les r_i exceptés r_{16} sont inférieurs en valeur absolue à $t_{(0,025;n-p)}$ [= 2,11] qui est le 0,975 quantile d'une loi de Student avec (n-p) degrés de liberté.

La colonne (4) donne les valeurs de h_{ii}, $i^{ème}$ terme diagonal de la matrice de projection : $\underset{\sim}{H} = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\ \underset{\sim}{X}'$ où $\underset{\sim}{X}$ est la matrice des valeurs observées des variables explicatives et $\underset{\sim}{X}'$ sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques. La valeur critique pour déterminer les points leviers correspond à h* = $\frac{3 \times 3}{20} = \frac{3p}{n}$ = 0,45. On constate que tous les hii sont inférieures à cette valeur critique 0,45.

Les résidus studentisés externes t_i, rassemblés dans la colonne (5).les valeurs t_i sont inférieurs en valeur absolue à $t_{(0,025;n-p-1)}$ [2,101], à l'exception des points 16.

III – 3 Autres diagnostics d'influence:

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

On remarque que les distances de Cook sont toutes inférieures 1 à l'exception de point16, et que les DFITS une observation 16 est supérieure à la valeur critique $2\sqrt{p/n} = 2\sqrt{3/20} = 0{,}7745$. Cette observation est donc inhabituelle.

III – 4 Vérification de la qualité de l'ajustement:

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par

« Leave –one –out ». La figure (7), qui reproduit les valeurs prédites $\hat{p}CIC_{50}$ en fonction de celles observées, fait ressortir une faible dispersion caractéristique d'un bon ajustement, d'ailleurs confirmé par la grande valeur de Q^2 95,73%).

$$\hat{p}CIC_{50} = -0,10449 - 0,894\ pCIC_{50} \qquad (2)$$

S = 0, 45 R-carré = 91, 1% R-carré (adjust) = 90, 6%

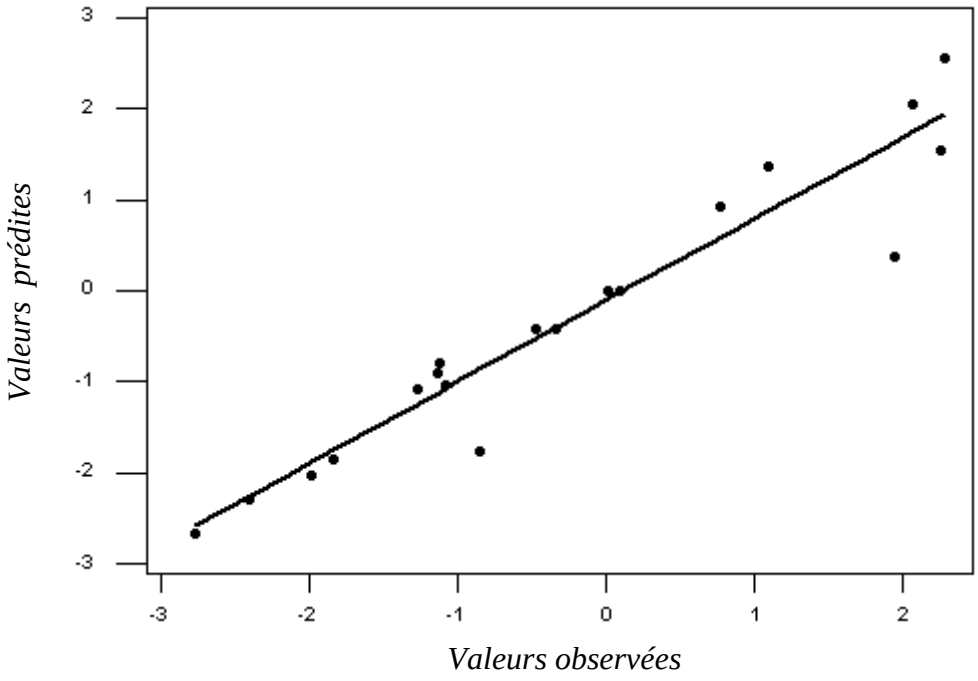


Figure 4– *Graphe des valeurs prédites* $\hat{p}CIC_{50}$ *en fonction des valeurs observées*

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

Ainsi 100 nouveaux vecteurs de $pcic_{50}$ ont été générés par permutation des positions des composantes du vecteur réel:

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

Et utilisés comme sources d'observations pour des modèles QSAR dans les conditions optimales établies (2 paramètres).

La figure 5 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (points noircis) au modèle réel de départ (astérisque).

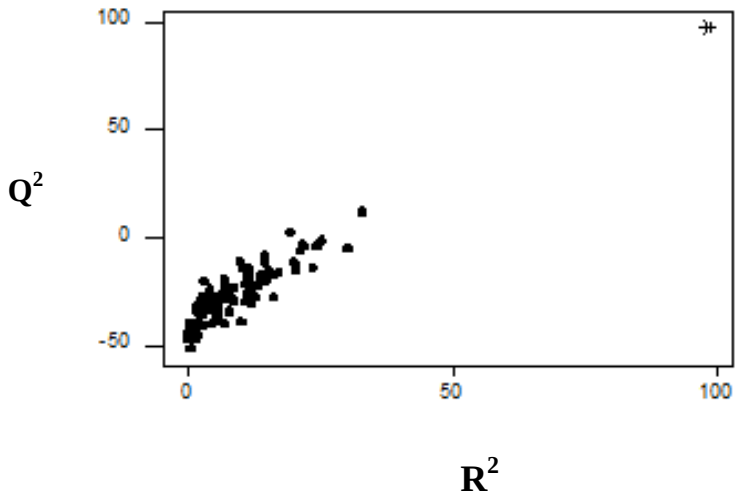


Figure 5 – *Test de randomisation associé au modèle QSAR. Les points noircis représentent* $pCIC_{50}$ *ordonnées de façon aléatoire, et l'astérisque correspond au modèle réel.*

Il est clair que les statistiques obtenues pour les vecteurs modifiés de $pcic_{50}$ sont plus petites que celles du modèle QSAR réel, et pour la majeure partie on obtient un $Q^2 < 0$. Ceci permet d'assurer qu'une relation structure/ activité réelle a été établie.

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

Tableau IV - Résidus caractéristiques, diagnostics d'influence et valeurs estimées de $pCIC_{50}$.

Observation i	Colonne 1	Colonne 2	Colonne 3	Colonne 4	Colonne 5	Colonne 6	Colonne 7	Colonne 8
	e_i	di	r_i	h_{ii}	t_i	pCIC50 estimée	D_i	DFITS
1	-0,022642	-0,11615101	-0,12914	0,186126	-0,12534	-2,74736	0,00127	-0,05994
2	-0,089849	-0,46091564	-0,49894	0,141417	-0,48763	-2,32015	0,01367	-0,1979
3	0,052943	0,27159186	0,28818	0,106426	0,28026	-1,89294	0,0033	0,09672
4	-0,081473	-0,41794767	-0,43368	0,065596	-0,42308	-1,03853	0,0044	-0,1121
5	0,14132	0,7249563	0,7499	0,059757	0,73985	-0,61132	0,01191	0,18652
6	0,204112	1,04707247	1,08535	0,063636	1,09144	-0,18411	0,02669	0,28453
7	0,099696	0,51142969	0,5409	0,100547	0,52932	0,6703	0,0109	0,17698
8	0,002489	0,0127683	0,01376	0,133578	0,01335	1,09751	0,00001	0,00524
9	0,118073	0,60570171	0,69181	0,228794	0,68081	1,95193	0,04733	0,37082
10	-0,099135	-0,50855182	-0,60579	0,290978	-0,59415	2,37913	0,0502	-0,38062
11	-0,061783	-0,3169401	-0,33786	0,114627	-0,32887	-1,92822	0,00493	-0,11833
12	-0,056199	-0,28829479	-0,30102	0,077202	-0,29282	-1,0738	0,00253	-0,08469
13	-0,056199	-0,28829479	-0,30102	0,077202	-0,29282	-1,0738	0,00253	-0,08469
14	0,029075	0,1491516	0,15724	0,09478	0,15266	-1,10907	0,00086	0,0494
15	-0,125652	-0,64458116	-0,68855	0,118328	-0,67751	-1,14435	0,02121	-0,2482
16	0,511876	2,62586847	3,1036	0,279815	4,57363	-1,36188	1,24749	2,85085
17	-0,259748	-1,33247912	-1,49591	0,201748	-1,55736	-0,08025	0,18852	-0,78293
18	-0,246955	-1,26685242	-1,4164	0,195161	-1,46316	0,34696	0,16216	-0,7205
19	-0,264163	-1,3551276	-1,51805	0,198292	-1,584	0,77416	0,19	-0,78777
20	0,204214	1,04759572	1,22647	0,265989	1,24627	2,05579	0,1817	0,75022

III – 5 Validation externe:

Pour s'assurer de la bonne capacité prédictive effective du modèle, nous avons opéré par validation externe sur l'ensemble de 10 composés choisis aléatoirement et qui ne font pas partie de l'ensemble d'essai.

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

$$\hat{p}CIC_{50} = -0,008265 + 0,9565\ pCIC_{50} \qquad (3)$$

S = 0,399 R-carré = 90, 7% R-carré (adjust) = 89, 5%

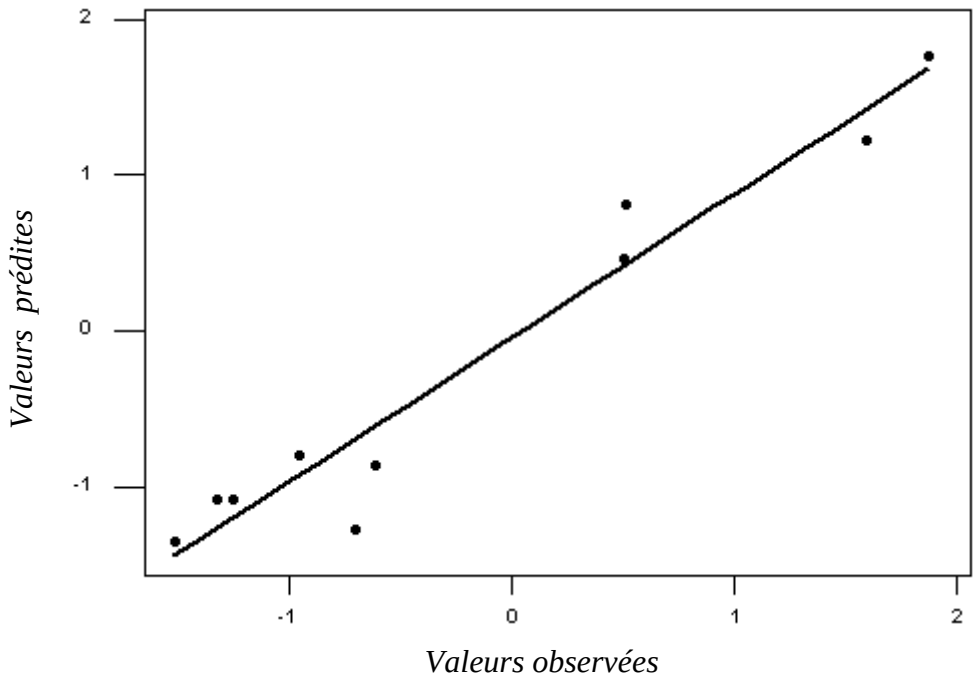


Figure 6 – *Graphe des* $\hat{p}CIC_{50}$ *prédites en fonction des* $pCIC_{50}$ *observées pour validation*

Tableau V – Valeurs observées, prédites et les résidus ordinaires de $pCIC_{50}$ par MLR pour l'ensemble de validation externe.

composés	$pCIC_{50}$	$\hat{p}CIC_{50}$	ei
Butan-1-ol	-1,52	-1.35	-0.264251
Octan-1-ol	0,5	0.47	-0.179768
Undecan-1-ol	1,87	1.77	-0.221634
Pentan-2-ol	-1,25	-1.09	-0.142651
Pentan-3-ol	-1,33	-1.09	-0.21926
(neo) pentanol	-0,96	-0.8	-0.215875
1-butylamine	-0,7	-1.28	0.559382
1-amylamine	-0,61	-0.86	0.184257
1-nonylamine	1,59	0.81	-0.059261
1-decylamine	1,95	1.23	0.559062

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

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

EQMC	= 0,263	(20 objets)	R^2	= 96,75
EQMP	= 0,301	(20 objets)	Q^2	= 95,73
EQMP (ext)	= 0,392	(10 objets)	Q^2 (ext)	= 91.71

Suggèrent, tout à la fois, une bonne capacité prédictive (faibles valeurs des EQM) et une possibilité d'extension suffisante (valeurs proches ou similaires) du modèle.

CONCLUSION GENERALE

Notre travail a consiste en l'application de la méthodologie QSAR à un ensemble de composés, constitué de vingt et un (21) alcools aliphatiques et neuf (09) amines aliphatiques, pour relier la toxicité de 30 composés précités ,caractérisés par la concentration d'inhibition à 50 % de la croissance (CIC_{50}) , avec 02 descripteurs : le coefficient de partage Octanol/eau logp et un descripteur électronique quantique Vee.

Le modèle obtenu très hautement significatif, permet de retrouver les valeurs prédites avec une erreur standard inferieure à 0,32 .

Les statistiques calculées réunis dans le tableau suivant:

n	σ_N	R^2	Q^2	F
20	0.352	96.75	95.73	253.23

L'analyse des résidus ne permet pas toujours vérifier les hypothèses d'un modèle linéaire statistique à effets fixes; les diagnostics d'influence ne permettent pas de mettre en évidence de point influant ou aberrant.

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par "leave-one-out". le grande valeur de Q^2 obtenus, indique que est un excellent résultat.

Notons enfin que le modèle obtenu contient une observation abérrante .

REFERENCES

[1] C. Boust, Institut national de recherche et sécurité .1ere édition avril 2004, page 1-6, www.inrs.fr**.**

[2] R. R. Lauwerys, V. Haufroid, P. Huet, D. Lison Toxicologie industrielle et intoxications professionnelles, 2007, page 643-644.

[3] les amines aliphatiques, Encyclopædia Universalis France S.A 2005 et 2009, http://www. Encyclopædia .com.

[4] A.Carpy, Importance de la lipophilie en modélisation moléculaire, analusis magazine, 1999, 1, 18 et 21.

[5] H.Hamada, Relations Quantitatives Structure/Activité d'une série de phénols, promotion 2007,3

[6] J.-P. ANGER, Effets à long terme et surveillance biologique des expositions professionnelles aux alcools et aux glycols, 2008, page6.

[7] Les amines, Wikipedia, 2008, http:// www. Wikipedia.com

[8] National Institute for Occupational Safety and Health, *RTECS (Registry of toxic effects of chemical substances)*, Hamilton, Ont.: Canadian Centre for Occupational Health and Safety (CD-ROM) http://ccinfoweb.ccohs.ca/rtecs/search.html **.**

[9] G. Persoone, D. Dive, Ecotox. Environ. Saf, 1978, 2, 105-144.

[10] Hyperchem™ Release 7.5 for windows, Molecular Modelling system, 2000.

[11] R. Todeschini, V. Consonni, M. Pavan, DRAGON, Software for the calculation of Molecular Descriptors. Release5.3 for Windows, Milano, 2005.

[12] E-calc computes all the E-stat values and displays them in a convent form of the screen. The computational parts of this program have been taken from: -Molconn-Z™ software. Lowell H,Hall Associates Consulting, 2 Devis street, Quincy, MA02170 for DOS version only. -Sci. QSAR™ 2D. Sci. Vision, Inc, 200 Wheeler Rd, Burlington, MA, 01803 for PC versions

[13] MINITAB, Release 14.1, Statistical Software 2003.

[14] P. Dagnélie, Statistique Théorique et Appliquée, Tome 1, De Boeck université Paris, Bruxelles, 1998,pp. 508.

[15] J. Durbin, G.S Watson, Testing for serial correlation in least squares regression. I, Biometrika, 1950, 37, 409-438.

[16] J. Durbin, G.S. Watson, Testing for serial correlation in least squares regression. II, Biometrika, 1951, 38, 159-178.

[17] J. Durbin, G.S. Watson, Testing for serial correlation in least squares regression. I, Biometrika, 1971, 58, 1-19.

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

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

[20] D. Cook, Detection of Influential Observations in linear Regression. Technometrics. 1977, 19, 15-18.

[21] Matlab Version 7.0.0.19920 (Release 14) The Language of Technical Computing The MathWorks, Inc. May 06, 2004.

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

[23] S. Weisberg, Applied linear Regression. J. Wiley, Inc., New York, 1980.

INTRODUCTION

Dans toute la suite, on désignera par **y** le vecteur de R^n dont les composantes y_1 , y_2 , .. , y_n représentent des réalisations particulières d'une certaine variable aléatoire Y. Typiquement **y** représente un vecteur d'observations.

Nous nous intéresserons dans ce qui suit, à 1'approximation du vecteur **y** de R^n par un vecteur **Y**, appelé vecteur ajusté, à travers une relation fonctionnelle de la forme :

$$\mathbf{Y} = \mathbf{X}\, b \qquad (4)$$

où **X** est une matrice à n lignes et p colonnes et b un vecteur de R^k dont les composantes peuvent être évaluées.

I – REGRESSION LINEAIRE SIMPLE

Un problème de régression consiste à étudier les changements de la valeur moyenne d'une variable (aléatoire) quand une autre variable ou plusieurs autres variables prennent différentes valeurs fixes. La première variable est appelée variable dépendante ou variable expliquée, les autres variables sont appelées "variables indépendantes" ou variables explicatives.

Lorsqu'il y a une seule variable explicative on dit qu'il y a une régression simple.

Lorsqu'il y a au moins deux variables explicatives on dit qu'il y a une régression multiple.

Pour écrire formellement un modèle de régression simple, on désigne par Y la variable dépendante, par **X** la variable explicative et par E (Y / X = x) la valeur moyenne de Y quand la variable X prend la valeur fixe x. Pour exprimer les changements susceptibles d'intervenir au niveau de la variable Y quand la variable X prend différentes valeurs, on posera :

$$E(Y / X = x) = f(x,\beta) \qquad (5)$$

où f est une fonction quelconque, dépendant du paramètre inconnu β, appelée la fonction de régression. Lorsque f est une fonction linéaire du paramètre β, on dira qu'on a une régression linéaire.

Le cas le plus élémentaire du modèle de régression est le cas où la fonction de régression $f(x,\beta)$ s'écrit $f(x,\beta) = \beta_0 + \beta_1 x$ où $\beta_0$et β_1 sont des paramètres inconnus. Dans ce cas, 1'expression :

$$E\,(Y/X=x) = \beta_0 + \beta_1 x \qquad (6)$$

est appelée modèle linéaire simple.

L'écriture (6) n'est pas très usitée puisque dans la pratique nous n'observons pas E (Y / X = x) mais une réalisation particulière, notée y, d'une variable Y quand X = x. Nous interprétons alors cette valeur y comme étant une valeur "bruitée" de la moyenne conditionnelle inconnue E (Y / X = x) que nous cherchons à estimer. Nous exprimerons tout ceci dans la relation :

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (7)$$

où ε représente une erreur aléatoire inobservable. Nous supposons généralement que 1' erreur a une moyenne nulle, E(ε) = 0 et une variance ε inconnue, Var (ε) = σ^2; naturellement β_0, β_1 et σ^2 sont des paramètres inconnus. IIs doivent être estimés à partir d'un échantillon d'observations de taille n, (yi, xi), i = 1 , ... , n par exemple. Il est donc naturel de considérer n relations du type (7) ; c'est-à-dire :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \qquad i=1 , ... n \qquad (8)$$

Les relations (8) peuvent être résumées sous une forme concise dans 1écriture matricielle suivante:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (9)$$

où y, x et ε sont des vecteurs de R^n et 1 un vecteur de R^n dont toutes les composantes sont égales à l.

L'analyse de la régression consiste à estimer les paramètres inconnus β_0, β_1 et éventuellement à "approcher" le modèle (9) par la relation linéaire :

$$y = b_0 + b_1 x = \begin{bmatrix} 1 & X_1 \\ . & . \\ . & . \\ 1 & X_n \end{bmatrix} \begin{pmatrix} b0 \\ \\ b1 \end{pmatrix} = Xb \qquad (10)$$

où les composantes b_o et b_l du vecteur b sont maintenant des quantités observables. Les quantités b_o et b_1 sont des estimations de β_0 et β_1. Le modèle (10), appelé droite de régression, est le plus simple cas particulier du modèle général (9).

II – REGRESSION LINEAIRE MULTIPLE

Dans le cas où nous considérons plusieurs variables explicatives simultanément nous obtenons une généralisation du modèle (9), examiné précédemment. Cela donne lieu au modèle :

$$E\ (Y / X_1 = x_1\ ,\ .....,\ X_k = x_k\ ) = \beta_0 + \beta_1 x_1 + ...+ \beta_1 x_1 + \beta_k x_k \qquad (11)$$

où x_l , ... x_k sont des valeurs fixes des variables "indépendantes" X_I , ... X_k.

Pareil modèle est appelé modèle de régression linéaire multiple.

II – 1 – Estimation des coefficients de régression par les moindres carrés

On utilise la méthode des moindres carrés pour estimer les coefficients de régression dans:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...+ \beta_n x_n + \varepsilon \qquad (12)$$

Supposons que n (> k) observations soient disponibles, et notons y_i la i^{eme} réponse observée et x_{ij} la i^{eme} observation (ou niveau) du régresseur x_j**.** Les données peuvent être présentées comme dans le tableau VI.

Tableau VI – ***Données d'une régression linéaire multiple***.

Observation	i	y	X_1	X_2	……..	X_k
	1	yi	X_{1l}	X_{12}	.....	x_{ik}
	2	y_2	x_{21}	X_{22}	.....	X_{2k}
	.	.	.	.	…….	
	.	.	.	.	…….	
	.	.	.	.	…….	
	.	.	.	.	…….	
	n	Y_n	Y_{nl}	X_{n2}	.....	X_{nk}

On suppose que le terme erreur ε dans le modèle a une moyenne nulle, E(ε)=0, une variance $V(\varepsilon)=\sigma^2$ et que les erreurs ne sont pas corrélées.

Le modèle correspondant à (17) peut être écrit comme suit:

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_I$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_I \qquad I=1,2,\ldots\ldots,n \tag{13}$$

La fonction des moindres carrés :

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{14}$$

doit être minimisée par rapport. à $\hat{\beta}_0, \hat{\beta}_1 \ldots\ldots\ldots \hat{\beta}_K$. Les estimateurs des moindres carrés pour $\beta_0, \beta_1 \ldots\ldots\ldots \beta_K$ doivent satisfaire les relations :

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1 \ldots\ldots, \hat{\beta}_K,} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \beta_j X_{ij}) = 0 \tag{15-a}$$

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1 \ldots\ldots, \hat{\beta}_K,} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \beta_j X_{ij} \right) X_{ij} = 0, \quad j=1,2,\ldots,k \tag{15-b}$$

En simplifiant (15) on obtient les équations normales des moindres carrés :

$$\begin{aligned}
&n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} X_{i2} + \ldots\ldots\ldots\ldots + \hat{\beta}_k \sum_{i=1}^{n} X_{ik} = \sum_{j=1}^{k} y_i \\
&\hat{\beta}_0 \sum_{i=1}^{n} X_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} X_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{i1} X_{i2} + \ldots\ldots + \hat{\beta}_k \sum_{i=1}^{n} X_{i1} X_{ik} = \sum_{i=1}^{n} X_i y_i \\
&\hat{\beta}_0 \sum_{i=1}^{n} X_{ik} + \beta_1 \sum_{i=1}^{n} X_{ik} X_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} X_{ik} X_{i2} + \ldots\ldots + \hat{\beta}_k \sum_{i=1}^{n} X_{ik} 2 = \sum_{i=1}^{n} X_{ik} y_i
\end{aligned} \tag{16}$$

Notons qu'il y a p = k + 1 équations normales, une pour chaque coefficient de régression inconnu. Les solutions des équations normales fourniront les estimateurs des moindres carrés $\hat{\beta}_0, \hat{\beta}_1, \ldots\ldots\ldots\ldots \hat{\beta}_k$.

Le modèle (13) peut s'écrire sous la forme matricielle suivante :

$$\underset{\sim}{y}=\underset{\sim}{x}\hat{\beta}+\underset{\sim}{\varepsilon}$$

où

$$y=\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \qquad X=\begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & & & & x_{2k} \\ . & . & & & & & \\ . & . & & & & & \\ 1 & x_{n1} & x_{2n} & . & . & . & x_{1k} \end{bmatrix}$$

$$\underset{\sim}{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix} \qquad \underset{\sim}{\varepsilon}=\begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

En général, y est un vecteur n × 1 des observations, X est une matrice n x p des valeurs des régresseurs**,** β est un vecteur p × 1 des coefficients de régression, et **ε** est un vecteur n × 1 d'erreurs aléatoires.

Nous voulons obtenir le vecteur des estimateurs des moindres carrés, β, qui minimise :

$$S(\beta)=\sum_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} =(y'-X\beta)'(y-X\beta)$$

Notons que S(β) peut encore s'écrire

$$\begin{aligned} S(\beta) &= y'y - \beta X y - y'X\beta + \beta' X'X\beta \\ &= y'y - 2\beta' X' y + \beta 'X'X\beta \end{aligned} \qquad (17)$$

puisque β'X'y est une matrice 1 × l,c'est-à-dire un scalaire, et sa transposée (β 'X'y) ' doit satisfaire à :

$$\left.\frac{\partial S}{\partial_{B0}}\right|_{\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

qui se simplifie en :

$$X'X\hat{\beta} = X'y \qquad (18)$$

Les équations **(**18**)** sont les équations normales des moindres carrés.

Pour résoudre les équations normales, multiplions à gauche les deux membres de (18)par 1'inverse de X'X. Ainsi 1'estimateur des moindres carrés de $\hat{\beta}$ est :

$$\hat{\beta} = \mathbf{(X'X)^{-1}\,X'} \qquad (19)$$

Pourvu que $\mathbf{(X'X)^{-1}}$ existe. La matrice $(X'X)^{-1}$ existera toujours si les régresseurs sont linéairement indépendants, c'est-à-dire, si aucune colonne de la matrice X n'est une combinaison linéaire des autres colonnes.

Il est facile de voir que X' X est une matrice p X p symétrique et que X' y est un vecteur colonne p × 1. Notons la structure particulière de la matrice X'X. Les éléments de la diagonale principale sont les sommes des carrés des éléments des colonnes de X, et les éléments non diagonaux sont les sommes des produits croisés des éléments des colonnes. Notons de plus que les éléments de X'y, sont les sommes des produits croisés des colonnes de X et des observations $y_{i:}$

Le modèle de régression ajusté correspondant aux niveaux des variables régresseurs $X' = [l, x_l, x_2, \ldots\ldots, x_k]$ est :

$$\hat{\mathbf{y}} = X'\hat{\beta}$$

$$= \hat{\beta}_0 + \sum_{\mathbf{j=1}}^{\kappa} \boldsymbol{\beta}_j \mathbf{X}_j$$

Le vecteur des valeurs ajustées $\hat{\mathbf{y}}$, correspondant aux valeurs observées y_i est :

$$\hat{\mathbf{y}} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy \qquad (12)$$

La matrice n × n, H = X $(X'X)^{-1}$ X', appelée matrice de projection, (ou matrice chapeau), joue un rôle central dans 1'analyse de régression.

La différence entre la valeur observée $\mathbf{y_i}$ et la valeur ajustée correspondante est appelée résidus $e_i = \mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}$. Les n résidus peuvent être mis sous la forme matricielle:

$$e = \mathbf{y} - \hat{\mathbf{y}} \qquad (20\text{-a})$$

Plusieurs autres voies permettant d'exprimer le vecteur des résidus **e** ont prouvé leur utilité, comme par exemple :

$$e = y - X\hat{\beta} \qquad (20\text{-b})$$

$$= y - Hy \qquad (20\text{-c})$$

$$= (I-H)y \qquad (20\text{-d})$$

II – 2 – Propriétés des estimateurs au sens des moindres carrés

II – 2 – 1 – Estimateur β

On peut facilement démontrer les propriétés statistiques de 1'estimateur $\hat{\beta}$ au sens des moindres carrés. Considérons le biais :

$$E(\hat{\beta}) = E[(X'X)^{-1}X'y] = E[(X'X)X'(X\beta + \varepsilon)$$

$$= E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon] = \beta$$

puisque $E(\varepsilon) = 0$ **et** $(X'X)^{-1}X'X = I$. Ainsi $\hat{\beta}$ est un estimateur sans biais de β.

La propriété de variance de $\hat{\beta}$ j s'exprime par la matrice de covariance :

$$Cov(\hat{\beta}) = E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]\}$$

qui est une matrice p × p symétrique dont le j ^{éme} élément de la diagonale principale est la variance de ßj et dont 1'élément (ij) n'appartenant pas à la diagonale principale est la covariance entre $\hat{\beta}$ i et $\hat{\beta}$ j . La matrice de covariance de $\hat{\beta}$ est :

$$Cov(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

Ainsi si 1'on pose C =$(X^{/} X)^{-1}$, la variance de $\hat{\beta}_j$ est $\sigma^2 C_{jj}$ et la covariance entre $\hat{\beta}$ i et $\hat{\beta}$ j est $\sigma^2 C_{ij}$.

L'estimateur $\hat{\beta}$ au sens des moindres carrés est le meilleur estimateur linéaire non biaisé de β (Théorème de Gauss - Markov). Si nous supposons de plus que les erreurs ε_i se distribuent normalement $\hat{\beta}$ est également 1'estimateur au sens du maximum de vraisemblance. L'estimateur au sens du maximum de vraisemblance est 1'estimateur sans biais et de variance minimale de β .

II – 2 – 2 – Estimation de σ^2

On peut développer un estimateur de σ^2 à partir de la somme des carrés des résidus.

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum e_i^2 = é\ e$$

En substituant e = y – x $\hat{\beta}$ on obtient :

$$SCE = (y - X\hat{\beta})\,(y - X\hat{\beta})$$

$$= y'y - \hat{\beta}'y + \hat{\beta}'X'X\beta$$

$$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}'$$

Comme **X' X$\hat{\beta}$ = X'y,** cette dernière équation devient :

$$SCE = y'y - \hat{\beta}'X'y \qquad (21)$$

La somme des carrés des résidus possède n - p degrés de liberté puisque p paramètres dans le modèle de régression sont estimés. Le carré moyen des écarts est :

$$CME = \frac{SCE}{n - p} \qquad (22)$$

On peut montrer que la valeur moyenne de CME est σ^2, et qu'un estimateur sans biais de σ^2 est fourni par :

$$\sigma^2 = CME \qquad (23)$$

Cet estimateur de σ^2 dépend du modèle.

III – DIAGNOSTICS DE REGRESSION ET MESURES DE L'ADEQUATION D'UN MODELE

L'évaluation de 1'adéquation d'un modèle est une part importante du problème de régression multiple. Nous présentons dans cette partie plusieurs méthodes d'évaluation de 1'adéquation d'un modèle.

III – 1 – Coefficient de corrélation multiple

Rappelons que la somme des carrés totale, SCT ou S_{yy} peut être décomposée en une somme de carrés due à la régression SC_R (ou S_R) et une somme des carrés des écarts SCE.

$$\sum_{i} (y_i - \bar{y})^2 = \sum_{i} (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \qquad (24)$$

(S_{yy} ou SCT) = (SCE) (SCR)

Le carré du coefficient de corrélation linéaire R^2 est définit par :

$$R^2 = \frac{SCR}{S_{yy}} = 1 - \frac{SCE}{S_{yy}}$$

II est de coutume de penser R^2 comme une mesure de la réduction de la variabilité de y obtenue en utilisant comme variables les régresseurs x_1, x_2, ... , x_k. Nous devons avoir $0 \leq R^2 \leq 1$.

Le carré R^2 du coefficient de corrélation, multiplié par 100, est appelé "coefficient de détermination".

Le coefficient de détermination R^2 est utilisé pour tester la qualité de 1'ajustement de y par $\hat{y}$.

Si R^2 = l, $y_i = \hat{y} \quad \forall$ i, 1'ajustement est parfait. Cependant, une grande valeur de R^2 n'implique pas nécessairement que le modèle de régression est un bon modèle. L'ajout d'un régresseurs au modèle se traduira toujours par une augmentation de R^2, indépendamment de la contribution ou pas de ce régresseur au modèle. Ainsi il est possible que des modèles à grandes valeurs de R^2 ne soient pas performants pour la prédiction ou 1'estimation.

La racine carré positive de R^2 représente le coefficient de corrélation multiple entre y et 1'ensemble des regrésseurs x_l, x_2, ... , x_k pris comme variables. C'est-à-dire que R est une mesure de 1'association linéaire entre y et x_1, x_2, ... , x_k. On peut également montrer que R^2 est le carré de la corrélation entre le vecteur des observations y et le vecteur des valeurs ajustées $\hat{y}$.

L'ajout d'une variable dans une équation de régression n'entraînant jamais une diminution de R^2**,** on préfère utiliser un R^2 ajusté :

$$R^2_a = 1 - \frac{SCE/(n-k)}{SCT/(n-1)} = 1 - \frac{n-1}{n-k}(1-R^2) \qquad (25)$$

Si les deux statistiques sont très différentes il y a de fortes chances que le modèle soit sur-spécifié, c'est- à- dire qu'i1 comporte des termes qui ne contribuent pas de façon significative à l'ajustement.

III – 2 – Analyse des résidus

III – 2 – 1 – Définitions

L'ajustement d'une régression dans le cas d'un modèle à effets fixes, par la méthode des moindres carrés, conduit à supposer que chaque valeur observée de la variable expliquée (y) peut être "correctement" reconstituée à partir des variables explicatives (x_1 x_2, ... , x_k).

Pour construire le modèle et admettre que les coefficients de la régression sont sans biais et convergents, on montre qu'il faut poser comme hypothèses :

a/ Les résidus (e) ont une espérance mathématique nulle: $E(e) = 0$**;**

b/ le modèle choisi est correct (aucune variable explicative n'a été omise) ;

c/ les résidus sont indépendants entre eux :$E(e_i e_j) = 0$ si $i \neq j$, leurs covariances sont nulles.

d/ les résidus ont tous même variance σ^2(propriété d'homoscédasticité).

Par ailleurs, l'emploi de tests statistiques pour analyser la variation expliquée par la régression conduit à admettre :

e/ les résidus suivent une distribution normale (de Laplace - Gauss).

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 1'égalité des variances résiduelles, la normalité des résidus ou 1'absence d'auto - corrélation.

Il est parfois intéressant de travailler avec des résidus standardisés :

$$d_i = \frac{e_i}{\sqrt{CME}}, \qquad i=1,2,\ldots..,n \tag{26}$$

L'équation ci-dessus standardise (réduit) les résidus en les divisant par leur écart - type moyen.

Le résidu standardisé est calculé sur 1'ensemble des données, alors que le résidu studentisé est calculé en éliminant au préalable la i^{eme} observation.

Nous avons vu que 1'on pouvait écrire comme suit le vecteur des résidus :

$$e = (I - H)\, Y \tag{27}$$

où la matrice de projection: $H = X\,(X'X)^{-1}\,X$ possède de nombreuses propriétés utiles. Elle est symétrique ($H' = H$) et idempotente ($H\,H = H$). Pareillement la matrice $I - H$ est symétrique et idempotente. En posant $Y = X\,\beta + \varepsilon$ dans (27), il vient successivement :

$$\begin{aligned} e &= (I-H)\,(X\beta + \varepsilon) \\ &= X\beta - HX\beta + (I-H)\,\varepsilon \\ &= X\beta - X\,(X'X)^{-1}X'X\beta + (I - H)\,\varepsilon \\ &= (I-H)\,\varepsilon \end{aligned}$$

Ainsi les résidus sont obtenus par les mêmes transformations linéaires des observations y et des erreurs ε.

La matrice de covariance des résidus est :

$$\begin{aligned} V(e) &= V[(I - H)\varepsilon] \\ &= (I - H)\, V(\varepsilon)\, (I - H)' \\ &= \sigma^2\, (I - H) \end{aligned} \tag{28}$$

puisque $V(\varepsilon) = \sigma^2\, I$ et que $I - H$ est symétrique et idempotente. Généralement la matrice $I - H$ n'est pas diagonale, aussi les variances des résidus sont différentes et les résidus sont corrélés.

La variance du i^{eme} résidu est :

$$V(ei) = \sigma^2 (I - hii) \quad (29)$$

où h_{ii} est le i^{eme} élément de la diagonale principale de H**.** Comme o < h_{ii}< l, 1'utilisation du carré moyen des écarts CME pour estimer la variance des résidus surestime réellement V(ei)**.** De plus comme h_{ii} est une mesure de la position du i^{eme} point dans 1' espace des **x**, la variance de e_i dépend de la région où se trouve Xi. En général les points voisins du centre de 1'espace des **x** ont de plus grandes variances (plus mauvais ajustement par les moindres carrés) que les résidus correspondants aux localisations plus éloignées. Les violations des hypothèses du modèle sont observées, le plus souvent, pour les points éloignés, et elles peuvent être difficiles à détecter à partir de 1'inspection des e_i (ou des d_i) parce que leurs résidus seront en général plus petits.

Il est suggéré de tenir compte de 1'inégalité des variances lors de la réduction des résidus, en traçant le graphe des résidus studentisés :

$$r_i = \frac{e_i}{\sqrt{CME\,(1 - h_{ii})}}, \qquad i=1, 2, \ldots, n \quad (30)$$

plutôt que celui des ei (ou d**i**). Quand la forme du modèle est correcte les résidus studentisés ont une variance constante $V(r_i) = 1$, indépendamment de la position Xi. La variance des résidus se stabilise dans beaucoup de cas, en particulier pour les grands ensembles de données. Ainsi les résidus standardisés et studentisés véhiculent souvent des informations équivalentes. Cependant, comme tout point caractirisé par un résidu important et un grand h_{ii} a de fortes chances d'influer grandement sur 1'ajustement par les moindres carrés, il est souvent recommandé d'examiner les résidus studentisés internes

La covariance entre e_i et e_j étant :

$$Cov(ei,e_j) = - \sigma^2 hij \quad (31)$$

une autre approche pour réduire les résidus consiste à transformer les n résidus dépendants en (n-p**)** fonctions orthogonales des erreurs ε**.** Ces résidus transformés sont indépendants, se distribuent normalement, et ont une variance constante σ^2.

III – 2 – 2 – Les représentations graphiques [14]

III – 2 – 2 – 1 – Diagrammes de dispersion des résidus en fonction de yi

La représentation graphique des résidus en fonction de la variable dépendante estimée fournit une série d'informations concernant 1' adéquation du modèle.

On peut également prendre en considération les résidus standardisés qui doivent se distribuer selon la loi normale réduite. En particulier, environ deux valeurs sur trois doivent être comprises entre - 1 et + l, et seulement cinq valeurs sur cent environ peuvent se situer en dehors de 1'intervalle (-2, +2).

III – 2 – 2 – 2 – Diagrammes de probabilité

Rappelons d'abord qu'on appelle quantile d'ordre α ($0 \leq \alpha \leq 1$) d'une variable aléatoire x de fonction de répartition F toute valeur X_α telle que :

$$F(x_\alpha) = \alpha \qquad \Leftrightarrow \qquad P(X \leq X_\alpha) = \alpha \tag{32}$$

Notons que si **F** est continue et strictement croissante, le quantile X_α , pour α donné, existe.

Les diagrammes de probabilité sont des diagrammes de fonctions de répartition, ou de fréquences cumulées, dans lesquels les ordonnées sont déterminées de telle sorte que les fonctions de répartition F(x) apparaissent sous la forme de droite.

Si, au contraire, on souhaite utiliser en ordonnées une échelle de quantiles de la variable normale réduite, les quantiles doivent être calculés, à partir des fréquences relatives, par la fonction inverse de la fonction de répartition $\Phi(n)$ de la distribution normale réduite

$$\mu_{i=}\ \Phi^{-1}[N'(X'_i)] \qquad \text{ou} \qquad \Phi^{-1}[(i-½)/n] \tag{33}$$

Les valeurs μ_i ainsi définies sont généralement appelées quantiles normaux ou scores normaux.

La représentation graphique d'un ensemble de fréquences cumulées sous une telle forme permet de juger, de façon visuelle, de la normalité ou de la non - normalité des données considérées. La linéarité ou la quasi - linéarité du diagramme ainsi obtenu est en effet un indice de normalité.

III – 2 – 3 – Test paramétrique : la statistique de Durbin et Watson [15-17]

La vérification de 1'indépendance des résidus peut se faire par le test de Durbin et Watson. La méthode consiste à calculer la quantité suivante :

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})2}{\sum_{i=1}^{n} e_i^e} \qquad (34)$$

Les e_i étant les résidus de la régression et n le nombre d'observations.

Cette caractéristique est comprise entre 0 et 4. Une valeur très inférieure a 2 indique 1'existence d'une corrélation positive entre les résidus successifs et une valeur très supérieure a 2 correspond a une corrélation négative entre ces résidus. Par contre, une valeur voisine de 2 ne permet pas de rejeter 1'hypothèse d'indépendance des résidus.

III – 2 – 4 les points influents:

La disposition des points dans 1'espace des **x** est importante pour la détermination des propriétés du modèle. Les observations éloignées, en particulier, ont une influence potentielle disproportionnée sur les paramètres estimés, les valeurs prédites, et les statistiques élémentaires.

III – 2 – 4 -1 La statistique représentée par le symbole DFITS :

$$\text{DFITS} = \frac{1}{p}\sqrt{\left(\frac{h_{ii}}{1-h_{ii}}\right)}\, t_i \qquad (35)$$

permet de mesurer l'influence d'une observation i sur la valeur ajustée ou prédite. Belsley, Kuh et Welsch [18] considèrent qu'une observation pour laquelle DFITS > $2\sqrt{p/n}$ (p étant le nombre de paramètres de la régression) est inhabituelle.

III – 2 – 4 -2 La distance de Cook D_i :

$$D_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}} d_i^2 \qquad (36)$$

permet d'étudier l'influence d'une observation i sur les coefficients de régression estimés par les moindres carrés. Cook [19] et Weisberg [20] suggèrent de comparer D_i au paramètre de Fisher $F_{(0,5,p,n-p)}$ et de contrôler les observations avec distances de Cook > $F_{(0,5,p,n-p)}$. Comme $F_{(0,5,p,n-p)} \approx 1$, on considère que les observations pour lesquelles $D_i > 1$ sont influentes.

.

IV –Robustesse du modèle

La stabilité du modèle a été explorée en utilisant la "validation croisée par omission d'une observation" (LOO : cross-validation by leave-one-out) [21]. Elle consiste à recalculer le modèle sur (n – 1) composés de phénol, le modèle obtenu servant alors à estimer l'indice de rétention du composé éliminé noté $\hat{y}_{(i)}$. On répète le procédé pour chacun des phénols.

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

$$PRESS = \sum_1^n \left(y_i - \hat{y}_{(i)}\right)^2 \qquad (37)$$

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

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \qquad (38)$$

Contrairement à R^2, qui augmente avec le nombre de paramètres de la régression, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier), obtenu pour un certain nombre de variables explicatives, puis décroît par la suite de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO}. Une valeur de $Q^2_{LOO} > 0{,}5$ est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [22].

V –Validation externe

En plus du test de randomisation, il est intéressant [23], pour juger de la qualité du modèle, de considérer la racine de l'écart quadratique moyen (RMSE, pour Root Mean Squared Error), calculée sur différents ensembles:

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

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

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

$$\sigma_N = EQMP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}{n}} = \sqrt{\frac{PRESS}{n}} \tag{40}$$

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}\left(y_i - \hat{y}_{(i)}\right)^2}{n_{ext}}} \tag{41}$$