Estimation de l'écotoxicité de substances chimiques par des méthodes à noyaux

Jonathan Villain

THESE / UNIVERSITE DE BRETAGNE-SUD
sous le sceau de l'Université Bretagne Loire

pour obtenir le titre de
DOCTEUR DE L'UNIVERSITE DE BRETAGNE SUD

Mention : Mathématiques
Ecole doctorale : Santé, Information, Communication, Mathématiques, Matière

Présentée par Jonathan Villain

Préparée à l'unité mixte de recherche : UMR CNRS 6205 et UPRES EA 4258-FR CNRS 3038 INC3M

Etablissement de rattachement : Université de Bretagne Sud

Nom développé de l'unité : Laboratoire de Mathématiques de Bretagne Atlantique et Centre d'Etude et de Recherche sur le Médicament de Normandie

Estimation de l'écotoxicité de substances chimiques par des méthodes à noyaux

Thèse soutenue le 24 juin 2016
devant le jury composé de :

Avner Bar Hen
Professeur, Université Paris descartes / Rapporteur
Ronan Bureau
Professeur, Université de Caen Normandie / Directeur de thèse
Anne Claude Camproux
Professeur, Université Paris-Diderot / Rapporteur
Bertand Cuissard
Maître de conférences, Université de Caen Normandie / Examinateur
Gilles Durrieu
Professeur, Université de Bretagne Sud / Directeur de thèse
Ernest Fokoué
Professeur, Rochester Institue of Technology, New-York / Examinateur
Jean-François Petiot
Maître de conférences, Université de Bretagne Sud / Examinateur

Laboratoire de Mathématiques de Bretagne Atlantique (LMBA)
UMR CNRS 6205 et Université de Bretagne-Sud
Campus de Tohannic
56000 Vannes, France

Remerciements

Comme l'a écrit Marcel Achard : « La chance existe. Sans cela, comment expliquerait-on la réussite des autres ». Et je pense que moi-même, j'ai bénéficié de cette chance en rencontrant tout au long de mes années d'études des personnes qui se sont intéressées à moi et m'ont ainsi permis de finaliser cette thèse.

Je voudrais remercier tout d'abord la région Bretagne ainsi que l'Université de Bretagne Sud pour avoir cru en moi en me permettant, grâce à leur financement, de m'ouvrir les portes du monde de la recherche.

Je remercie tout particulièrement mes directeurs de thèse, M. Gilles Durrieu, Professeur des Universités à l'Université de Bretagne Sud et M. Ronan BUREAU, Professeur des Universités à l'université de Caen en Normandie pour avoir cru en moi et s'être battus pour m'obtenir un financement « ce qui n'était pas gagné » et sans lesquels je n'aurais pu faire ce doctorat. Je leur suis reconnaissant également pour le temps conséquent qu'ils m'ont consacré malgré leurs nombreuses responsabilités et pour tous les conseils judicieux qu'ils m'ont prodigués ainsi que pour la bienveillance dont ils ont fait preuve à mon égard.

Un très grand merci également aux Membres du Laboratoire de Mathématiques de Bretagne Atlantique pour leur convivialité et plus particulièrement à Madame Véronique Vellet pour tout ce temps passé à nous aider avec les formalités administratives. J'associe également à ces remerciements toute l'équipe du Centre d'Etude et de Recherche sur le Médicament de Normandie pour m'avoir chaleureusement accueilli pour ma première année de thèse.

Je voudrais également remercier les rapporteurs de cette thèse Mr Avner Bar hen, Professeur des universités de Paris Descartes et MmeAnne-Claude Camproux, Professeur des universités de Paris-Diderot ainsi que les membres du jury Mr Jean-François Petiot, Maître de conférence de l'université de Bretagne Sud, Mr Ernest Fokoué Professor à Rochester Institute of technology et Mr Bertrand Cuissart, Maître de conférence de l'université de Caen Basse-Normandie qui ont accepté de prendre de leur temps pour examiner mon travail.

Un grand merci aussi à toute l'équipe de l'IUT de Vannes département STID pour tout le soutien qu'ils ont pu m'apporter sur cette fin de thèse. Je remercie également mes parents pour m'avoir soutenu tout au long de mes études et surtout pour la confiance qu'ils avaient en moi, même lorsque j'ai rencontré des échecs.

Résumé :

Dans le domaine de la chimie et plus particulièrement en chémoinformatique, les modèles QSAR (pour Quantitative Structure Activity Relationship) sont de plus en plus étudiés. Ils permettent d'avoir une estimation *in silico* des propriétés des composés chimiques notamment des propriétés écotoxicologiques. Ces modèles ne sont théoriquement valables que pour une classe de composés (domaine de validité) et sont sensibles à la présence de valeurs atypiques.

La thèse s'est focalisée sur l'établissement de modèles globaux robustes (intégrant un maximum de composés) permettant de prédire l'écotoxicité des composés chimiques sur une algue *P. Subcapitata* et de déterminer un domaine de validité dans le but de déduire la capacité de prédiction d'un modèle pour une molécule. Ces modèles statistiques robustes sont basés sur une approche quantile en régression linéaire et en régression Support Vector Machine.

Mots clés : chémoinformatique, QSAR, robustesse, quantile, Support Vector Machine, machine learning, écotoxicité.

Abstract :

In chemistry and more particularly in chemoinformatics, QSAR models (Quantitative Structure Activity Relationship) are increasingly studied. They provide an *in silico* estimation of the properties of chemical compounds including ecotoxicological properties. These models are theoretically valid only for a class of compounds (validity domain) and are sensitive to the presence of outliers.

This PhD thesis is focused on the construction of robust global models (including a maximum of compounds) to predict ecotoxicity of chemical compounds on algae *P. subcapitata* and to determine a validity domain in order to deduce the capacity of a model to predict the toxicity of a compound. These robust statistical models are based on quantile approach in linear regression and regression Support Vector Machine.

Keywords : chemoinformatics, QSAR, robustness, quantile, Support Vector Machine, machine learning, ecotoxicity.

Table des matières

Introduction générale

On estime à plusieurs millions le nombre de composés chimiques actuellement synthétisés. Les préoccupations en terme de santé publique et d'environnement concernent dans un premier temps, les composés produits au niveau industriel (production supérieure à 1 tonne), composés que l'on retrouve sous forme pure ou associés à d'autres composés pour former des mélanges complexes. Ces dérivés chimiques ont fait l'objet du programme européen REACH (Registration, Evaluation, Authorization and registration of CHemicals [1]) qui oblige les industriels (production ou importation) à fournir un certain nombre d'informations (endpoints) concernant les propriétés physico-chimiques, toxicologiques et écotoxicologiques de ces substances. En considérant les données actuelles, autour de 50000 composés sont dans ce cas de figure, nombre à opposer aux millions de composés déjà synthétisés.

L'autre point fondamental concerne les tests dits *in vivo* pour l'estimation des propriétés biologiques. Ces tests sont associés à des expériences réalisées sur des animaux de laboratoire. La réalisation de chacun des tests pour un nombre maximal de composés chimiques aboutit d'emblée à des problèmes éthiques majeurs. C'est pourquoi, le développement de nouvelles méthodes alternatives représente un enjeu important dans les prochaines années. On distingue pour ces méthodes, les techniques *in vitro* avec des tests sur des entités cellulaires particulière-

ment (voir EURL-ECVAM [2]) et les tests *in silico* avec une détermination par des méthodes informatiques (QSAR toolbox par exemple [3]). Ces dernières peuvent aller historiquement des méthodes de relations structure-activité quantitatives (QSAR en anglais) à des techniques plus complexes de modélisation des systèmes biologiques (biologie systémique et méthodes *in silico* [4]). Le point crucial concerne évidemment l'accord entre les données issues de méthodes alternatives et les données *in vivo* au plus proche des systèmes biologiques. En effet, même pour les données biologiques (méthodes *in vivo*), des espèces de référence ont été définies, espèces présentant leurs propres caractéristiques (par exemple, des tests sur rat pour estimer une toxicité humaine). Quelque soit la méthode, une marge de sécurité sera appliquée [5] correspondant généralement à près de 1000 fois la concentration d'un composé, concentration associée à un phénotype toxique (*i.e.* si une DL50 est égale à 1 mg/L, la concentration "admise" par une estimation du risque sera autour de 1 ug/L).

Les méthodes QSAR se sont développées sur la base des premières données disponibles. Ces données sont souvent en relation avec la mise en évidence d'une préoccupation particulière sur une classe de composés chimiques(à titre d'exemple, le cas de l'aniline traité par l'Allemagne [6]). Un des objectifs de REACH sera d'élargir la gamme de molécules (diversité moléculaire) pour lesquelles les données biologiques seront accessibles. L'autre point concerne la qualité des données. Les tests biologiques suivent des lignes directrices [7] qui permettent théoriquement une comparaison des données issues de divers laboratoires (point clé pour les méthodes QSAR). On peut estimer ainsi qu'actuellement et dans les prochaines années, le nombre de données disponibles sera multiplié par un facteur conséquent. Le problème sera évidemment leurs accessibilités "informatiques" notamment pour une présentation initiale des données.

L'autre point concernera la notion de classes de composés chimiques. Peut-on comparer des hydrocarbures à des pesticides ? Doit-on définir des classes et par conséquent des limites de

classes pour les composés chimiques. Ce point sera abordé ici à travers la notion de mode d'action en considérant que pour deux composés ayant le même mode d'action, un modèle QSAR devrait permettre d'estimer la valeur de l'un en considérant la valeur de l'autre. Nous ne sommes pas très loin ici de la notion de références croisées (READ-ACROSS). Cette approche *in silico* considère qu'une famille de composés bien identifiés doit avoir des propriétés biologiques très proches [8]. Il suffirait de réaliser des tests sur une partie des composés de la famille pour estimer les propriétés de l'ensemble de la famille (couplage *in vivo* et *in silico*).

Que cela soit pour la définition de familles chimiques ou des modèles QSAR, la discipline de base est la chémoinformatique. Lors de la mise en place de la société française de chémoinformatique, une définition de cette discipline est ressortie en version anglaise : "*Chemoinformatics is a scientific discipline that has evolved in the last 40 years at the interface between chemistry and computer science. It has been realized that in many areas of chemistry, the huge amount of data and information produced by chemical research can only be processed and analyzed by computer methods. Furthermore, many of the problems faced in chemistry are so complex that novel approaches utilising solutions that are based on informatics methods are needed. Thus, methods developed for building databases on chemical compounds and reactions, for the prediction of physical, chemical and biological properties of compounds and materials, for drug design, for structure elucidation, for the prediction of chemical reactions and for the design of organic syntheses. Research and development in chemoinformatics is essential - For increasing our understanding of chemical phenomena - For industry to remain competitive in a global economy Chemoinformatics methods can be applied in any field of chemistry, from analytical chemistry to organic chemistry. It is of particular importance in drug design and development*".

Au CERMN, deux programmes ANR ont été financés durant ces 8 dernières années. Un en relation avec la problématique REACH (ANR Innotox 2007) et le deuxième centré sur le

problème du relargage des médicaments dans l'environnement (ANR Pharm@ecotox, 2010 [9]). Le premier programme nous a permis de définir en interne une base de données importante sur un certain nombre de endpoints en relation avec les propriétés écotoxicologiques et toxicologiques des dérivés chimiques. Le deuxième programme est clairement orienté écotoxicologie et vise à estimer les toxicités aiguës et chroniques des médicaments en considérant les espèces de références décrites dans REACH (poissons, crustacées, algues). Le programme présentait à côté des aspects *in silico* un certain nombre de tests biologiques à réaliser sur ces médicaments. Parmi l'ensemble des données en écotoxicologie et en considérant aussi les modèles QSAR précédemment développés, l'analyse des données sur une espèce clé (les algues) était curieusement absente (nombre de modèles QSAR faibles et orientés vers une classe de composés chimiques). Une des difficultés venait vraisemblablement de la faible présence de données fiables et des caractéristiques du test sur algue, à l'interface entre les données de toxicités aiguës et de toxicités chroniques (analyse de l'impact sur la croissance de l'algue en considérant trois générations).

L'objectif de cette thèse sera de s'intéresser à ce dernier endpoint (algue verte) afin d'établir de nouveaux modèles intégrant un maximum de dérivés. Pour cela, la notion de modèles globaux sera abordée avec dans ce cadre une distinction entre un mode d'action non spécifique d'un mode d'action spécifique (interaction avec une ou des macromolécules cibles). Ces modèles globaux nécessitent une analyse particulière et nous nous sommes intéressés aux méthodes à noyaux conjuguées aux méthodes quantiles pour les régressions.

L'objectif général de cette thèse est à plusieurs niveaux :

- déterminer le potentiel des méthodes à noyau couplées à des méthodes quantiles sur un ensemble très divers de composés chimiques,
- analyser le domaine de validité de ces modèles (notion de outliers),
- donner ou permettre une interprétation pour les résultats statistiques obtenus,

— déterminer de nouveaux modèles QSAR pertinents pour l'écotoxicité sur algues.

Nous articulerons cette thèse autour de 5 chapitres. Le Chapitre I donne un état de l'art des méthodes d'apprentissage et de modélisation statistique de type QSAR utilisées dans le domaine de la chémoinformatique. Les descripteurs structuraux des composés chimiques sont aussi présentés. Le Chapitre II est dédié à la description de modèles statistiques robustes utilisés dans les chapitres suivants. Dans le Chapitre III, le niveau de toxicité de composés chimiques est modélisé et estimé à partir des données collectées sur les algues. Ce travail a été publié en 2015 et est inclus dans l'article intitulé "Quantile regression model for a diverse set of chemicals : application to acute toxicity for green algae". Le chapitre IV est consacré à l'estimation de la toxicité de médicaments. Cet article a été publié en 2016 et est aussi inclus sous l'intitulé "Acute toxicities of pharmaceuticals toward green algae. mode of action, biopharmaceutical drug disposition classification system and quantile regression models". Le Chapitre V est consacré à la construction d'une méthode séquentielle robuste en régression dans le cas univarié et multivarié. Enfin, nous concluons cette thèse par une conclusion générale et nous décrivons également quelques perspectives de recherche.

Chapitre I

État de l'art : modélisation QSAR

Pour classer et utiliser les composés chimiques, il est indispensable de connaître leurs propriétés physico-chimiques et leurs activités biologiques, mais les essais expérimentaux notamment biologiques s'avèrent très onéreux. De plus, pour des raisons d'éthique ils doivent être limités. Aussi, a-t-on recours à des méthodes alternatives comme les méthodes chémoinformatiques qui ont pour but d'anticiper le comportement des molécules ou des systèmes moléculaires. Les techniques QSAR (Quantitative Structure-Activity Relationships) et QSPR (Quantitative structure property relationship) sont parmi les plus utilisées. La méthode QSAR modélise la relation entre la structure et l'activité de la molécule. Cette méthode a été introduite au XIXème siècle. En effet, dès 1868-1869, Crum-Brown chimiste et Fraser pharmacologiste écrivaient déjà dans leur papier [10] "il ne peut y avoir d'objection raisonnable contre le fait que la relation entre l'effet physiologique d'un composé et sa constitution chimique existe ... ". Cependant, à cette époque, les structures moléculaires n'étaient pas encore connues. Par la suite, il a fallu attendre les années 60 avec les travaux de Free et Wilson [11] et les travaux de Hansh et Fujita [12] pour obtenir les premiers modèles de régression reliant les caractéristiques des structures chimiques à leurs propriétés physico-chimiques et/ou biologiques. Pour l'analyse de Hansch [12], trois types de descripteurs sont fondamentaux dans cette approche afin de décrire les activités biologiques : les descripteurs hydrophobiques, stériques et électroniques. Dans Free et Wilson [11], des fragments moléculaires dérivés de l'analyse 2D des structures chimiques sont plutôt considérés. Par la suite comme décrits ci-dessous, de nombreux descripteurs moléculaires ont

été définis avec des techniques d'analyses statistiques extrêmement variées. Pour les travaux QSAR dans le domaine de l'écotoxicologie et ceci à partir des années 70, on peut citer le travail remarquable de Veith [13].

1 Données

L'estimation des propriétés toxicologiques et écotoxicologiques de produits chimiques est une préoccupation environnementale majeure actuellement [14]. Cette préoccupation a été à la base de la mise en place au niveau européen du programme REACH [1]. Au sein de la réglementation REACH, les industriels de la chimie doivent fournir des informations sur un certain nombre de critères (endpoints) concernant les propriétés physico-chimiques et (éco)toxicologiques. Cette législation a mené à la création d'une base de données dans laquelle ces informations sont regroupées (European CHemical Agency : ECHA [15]).

En chimie thérapeutique, on cherche souvent à déterminer l'activité biologique d'un composé chimique sur un certain nombre de récepteurs à travers les notions d'affinités (données de liaisons) ou d'effets pharmacologiques (activation ou non du récepteur pour un récepteur membranaire par exemple). L'activité cherche donc à traduire de façon quantitative l'effet du composé chimique sur la cible. En écotoxicologie, on cherche à mettre en avant l'effet des composés chimiques de façon globale sur une espèce donnée (on ne s'intéresse pas à une cible particulière). Cela se traduit par des notions de mortalité ou d'arrêt de croissance. Dans le cadre de notre étude on cherche à déterminer l'impact de composés [16] sur la croissance d'une algue verte (*Pseudokirchneriella subcapitata*). Cette activité est donnée par la concentration du composé chimique conduisant à 50 % d'inhibition de la croissance de l'algue, concentration classiquement exprimée en mole par litre (mol/L) et notée EC_{50}.

En ce qui concerne les substances enregistrées actuellement au niveau de REACH, l'ECHA [15] nous indique qu'au 6 février 2016, 13876 substances uniques sont enregistrées. Ces substances pourront faire l'objet d'une validation externe de nos modèles. Dans cette thèse, les données recueillies en 2007 ont porté sur une seule espèce d'algue verte (*P. subcapitata*) et proviennent de 5 bases de données différentes. La première base de données enregistre les résultats de tests d'écotoxicité conduits par le ministère de l'environnement du Japon en mars 2014. Elle peut être obtenue sur le site web de l'OECD (Organisation for Economic Cooperation and Development) [17]. Cette base de données est composée de 1026 composés chimiques pour lesquels le protocole OECD-GLP (Good Laboratory Practice) standard a été suivi pour collecter les informations sur l'écotoxicité des composés chimiques. Pour ces tests, 2 types de concentration effective à 50% ont été déterminés sur une durée de 72 heures : le taux de croissance pour les algues (EC50r) et l'aire sous la courbe de croissance (AUG / EC50b). Pour les données collectées les plus récentes, seules les valeurs de EC50r ont été déterminées. Nous avons par conséquent montré la corrélation significative entre les deux variables ($n = 249$, $r = 0,967$, $p < 0.05$). Nous avons considéré au final 277 composés chimiques associés aux valeurs de concentration les plus basses. Nous avons aussi réalisé une sélection sur la base des valeurs hydrophobiques et hydrophiliques (*vide infra*) associées à ces composés. La seconde base de données correspond à la base de données de l'ECB [18]. Dans cette base de données, les tests associés à 2782 composés chimiques étaient initialement référencés (programme ANR Innotox pour l'analyse). Parmi ces 2782 composés, 1749 composés ont une structure chimique unique (ce n'est pas un mélange complexe). Les métaux ou les composés organo-métalliques ont été aussi éliminés. Pour les composés chimiques restants seulement 47 possèdent une valeur pour la croissance sur algue en accord avec les lignes directrices précédentes (EC50r pour 72 heures sur la même espèce d'algue). Comme pour les données provenant de la base de données OECD, les 47 composés appartiennent à l'intervalle de valeur obtenu pour les valeurs hydrophobiques et hydrophiliques.

La troisième base de données correspond à la base de données AQUatic Information REtrieval (AQUIRE) [19]. Sur ces données, 60 composés respectant les critères de sélection sont ressortis (même espèce et même protocole expérimental). La quatrième base de données correspond aux données du Ministère de l'environnement (France). Elle a été discutée dans la publication de Faucon et al. [20], données organisées suite à une collaboration entre le Ministère de l'Environnement et le CERMN. Elle est composée de 94 composés chimiques. Enfin la dernière base de données regroupe 36 médicaments dont les tests ont été réalisés dans le cadre du programme ANR Pharm@ecotox au sein du laboratoire CERMN. Après avoir éliminé tous les doublons, nous nous retrouvons avec un échantillon de 437 composés chimiques.

Afin de décrire la structure des différents composés chimiques, en chémoinformatique, un large panel de descripteurs a été caractérisé depuis plusieurs dizaines d'années. L'objectif de ces descripteurs est de permettre de relier une ou des propriétés particulières des composés à des propriétés physiques ou biologiques enregistrées [21]. Ces descripteurs peuvent être globaux comme le poids moléculaire mais aussi associés à un élément particulier de la structure (une fonction chimique par exemple). On peut distinguer des grandes familles de descripteurs comme les descripteurs dits 1D, 2D et 3D.

Les descripteurs 1D servent essentiellement à décrire les caractéristiques des composés chimiques en considérant plutôt des paramètres généraux comme des propriétés physico-chimiques (poids moléculaire, logP, pKa, solubilité...). Mais cela peut être aussi de simplement compter des fonctions chimiques particulières ou décrire la présence d'atomes comme les halogènes par exemple (un simple comptage). Les contributions à des propriétés particulières de certains groupements peuvent rentrer dans cette catégorie de descripteurs (à la limite de la notion de descripteurs 1D et 2D). En effet, des fonctions chimiques ou des atomes auront potentiellement un effet particulier que l'on peut chiffrer par mécanique quantique (impact électronique sous

forme de charge par exemple) ou par un accord entre l'évolution de données thermodynamiques et la présence / absence de ces substituants. Dans ce dernier cas, ces fonctions / atomes sont décrites sous forme de valeurs comme les constantes de Hammett σ pour les propriétés électroniques, les constantes Π associées aux propriétés hydrophobiques ou les constantes stériques MR représentant les contributions atomiques à la réfractivité moléculaire ou de Taft [22]. La logique par la suite pour ces constantes est de travailler sur un ensemble de dérivés très homogènes et de voir l'impact d'un groupement sur l'évolution de cette propriété (relation type régression linéaire entre les valeurs des constantes et la propriété). Ils sont aussi, suite à une analyse topologique de la structure (descripteurs $2D$) à l'origine du calcul théorique de certaines propriétés physico-chimiques comme notamment le logP.

Les descripteurs 2D sont plus complexes et plus précis et correspondent aux descripteurs topologiques. Ils sont basés sur la théorie des graphes avec une définition historique de la notion d'indice décrivant une structure (travaux fondamentaux de Balaban, Kier, Hall ...). On distingue notamment les indices de connectivité, connectivité exprimée sous forme de matrice (matrice d'adjacente et de distance). Ils permettent de décrire la forme, la taille et les ramifications du composé chimique au niveau 2D. La théorie des graphes est aussi à la base de la notion d'empreintes moléculaires allant de la recherche de sous-structure fixe (MDL-keys en relation avec le code SMART par exemple) jusqu'à un analyse complète de la structure avec notamment les hashed fingerprint. Nous nous sommes intéressés dans cette thèse à un type d'empreintes moléculaires nommé FCFP6 (Functional Class FingerPrint : FCFP) [23] et ceci à titre exploratoire. Les descripteurs 3D tiennent compte de la structure en 3 dimensions et sont logiquement une extension des descripteurs topologiques avec cette fois-ci l'intégration de distances réelles entre les sous-structures. Ils sont à l'origine notamment de la définition des pharmacophores ou toxicophores. Le problème principal vient de la notion de conformation et

de l'importance potentielle vis à vis des descripteurs. De même, pour des calculs quantiques associés à des propriétés électroniques, un minimum conformationnel doit servir de base. Dans nos études, un focus a été particulièrement réalisé sur les interactions de surface. Par conséquent, nos descripteurs sont principalement en accord avec les propriétés de surface. Nous avons évidemment considéré aussi des descripteurs clés en écotoxicologie comme le coefficient de partage octanol/eau Log(K_{OW}) ou la solubilité moléculaire dans l'eau. Les propriétés de surface sont associées à des structures 3D. Nous avons opté vu le nombre de dérivés pour une procédure type workflow pour les générer. Le logiciel utilisé est Pipeline Pilot avec une génération automatique des structures 3D suivie d'une première minimisation d'énergie en utilisant un champ de force particulier (clean force field). Par la suite une nouvelle optimisation a été réalisée par méthode quantique exploitant la théorie de la fonctionnelle de densité (DFT). Pour l'hamiltonien, le basis set DND a été sélectionné avec comme fonctionnelle de densité PWC [24] et une convergence de 1.0e-5 (SCF density convergence) pour l'optimisation. Nous obtenons ainsi une seule structure 3D pour chaque dérivé. Les descripteurs suivants ont été considérés sur une base 1D 2D 3D. Les coefficients de partage octanol/eau (LogK_{OW}) ont été obtenus par deux méthodes. L'une (AlogP) est complètement théorique et est basée sur une contribution atomique à la valeur de LogK_{OW}, l'autre intègre une méthode de calcul proche de la précédente mais aussi des données expérimentales réelles quand elles sont présentes.

La solubilité moléculaire (logS) a été déterminée par un modèle de régression linéaire sur une base de descripteurs décrits par Tetko et al. [25]. Le poids moléculaire, le volume moléculaire, le rayon de giration (la racine carrée des distances moyennes enregistrées pour tous les atomes dans le système moléculaire vis à vis du centre de gravité de la molécule), la globularité (analyse du caractère sphérique ou non d'une molécule (valeur entre 0 et 1)), la somme des polarisabilités atomiques (Apol), les moments principaux d'inertie (valeur globale, valeurs selon les trois

axes X, Y, Z), les descripteurs associés au moment dipolaire (valeur globale, valeurs selon les trois axes X, Y, Z) sont obtenus avec Pipeline Pilot [23]. Des descripteurs sont issus des calculs de mécanique quantique : les valeurs de HO et BV pour les orbitales frontières, la différence d'énergie entre HO et BV, l'énergie de solvatation, le volume de la cavité (phénomène de solvatation / désolvatation), l'énergie diélectrique (phénomène de polarisation). Les descripteurs de type JURS déterminent les surfaces totales, polaires at accessibles aux solvants en partant des structures 3D. La méthode est décrite par Rohrbaugh et Jurs en 1987 [26]. Ils sont au nombre de 30 en considérant les notions de surface partielle, le type de charge pour la polarité (charges positives ou négatives), les différences entre les surfaces polaires notamment. Les indices Shadow, très proches des descripteurs précédents, considèrent une projection de la forme moléculaire (volume) sur trois axes avec dans chaque cas une définition d'un plan d'analyse selon un axe. Ces indices vont faire ressortir les dimensions maximales d'une structure donnée selon chaque axe.

Ehresman et al. [27] ont décrit de nouveaux descripteurs moléculaires basés sur l'analyse de propriétés locales au niveau des surfaces moléculaires et ceci par une approche quantique [28]. Les surfaces générées pour ce programme correspondent à un type particulier nommé "shrink-wrapped isodensity surface" en considérant pour la densité électronique une valeur de $10^{-4}.e^{-}.\text{Å}^{-3}$. Quatre propriétés locales ont été calculées sur la surface moléculaire : le potentiel électrostatique moléculaire, l'énergie d'ionisation locale (IE_L), l'affinité électronique locale (EA_L), la polarisabilité locale (α_L). Deux propriétés, le "hardness" local (η_L) et l'électronégativité locale (χ_L) sont dérivées des valeurs de IE_L et EA_L. À partir de ces propriétés locales, 81 descripteurs ont été calculés.

2 Méthodes statistiques QSAR

2.1 Modélisation statistique

Avec l'apparition des outils informatiques et l'évolution des capacités de stockage informatique, le volume des données à traiter est devenu de plus en plus important. Dans le même temps la capacité des outils de calculs est devenue de plus en plus puissante et offre aux utilisateurs les moyens de traiter ces données. Dans différents domaines comme en chémoinformatique, il est souvent indispensable de connaître les relations qui lient une variable que l'on cherche à expliquer par le biais d'autres variables explicatives de cette dernière. Le but de la modélisation statistique est de prédire une variable à expliquer à partir de variables explicatives.

Il existe trois grandes classes de modèle de régression : les modèles paramétriques, les modèles non paramétriques et les modèles semi paramétriques [29].

Pour les modèles de régression paramétrique, la fonction de lien, entre la variable à expliquer $Y \in \mathbb{R}$ et une variable explicative $X \in \mathbb{R}^p$, dépend d'un nombre fini de paramètres à estimer. Les modèles sont de la forme :

$$Y = f_\theta(X) + \varepsilon,$$

où f_θ appartient à une famille de fonctions paramétrées par θ, vecteur de paramètres réels, et où ε est un terme d'erreur aléatoire. L'objectif concerne l'estimation du paramètre θ et les techniques d'estimation (maximum de vraisemblance, moindres carrés, ...) donnent des bons résultats lorsque la famille f_θ est correctement spécifiée tout en fournissant une interprétation de la ou des variables explicative(s) sur la variable à expliquer. Cependant, le choix d'un bon modèle paramétrique au vu des données peut s'avérer difficile et ainsi conduire dans le cas d'un mauvais choix à des conclusions erronées. Les modèles de régression non paramétrique apparaissent comme une alternative qui offre une flexibilité dans la modélisation du fait qu'aucune hypothèse

paramétrique n'est imposée dans le modèle et que seules des hypothèses de régularité de la fonction de lien sont imposées. La variable à expliquer Y est maintenant reliée à la variable explicative X en utilisant une fonction de lien f inconnue que l'on doit estimer :

$$Y = f(X) + \varepsilon.$$

Le thème commun de la régression fonctionnelle pour estimer $f(.)$ est l'idée d'un lissage local. Les méthodes existantes sont les méthodes à noyaux (estimateurs de Nadaraya-Watson) [30, 31], splines de lissage, polynômes locaux, etc. Dans le cas où la dimension de X devient importante, le nombre d'observations nécessaires pour le lissage local croît avec cette dimension. Les modèles semi-paramétriques ont alors été développés pour conjuguer les avantages des approches paramétriques et non paramétriques, à savoir la capacité d'interprétation des modèles paramétriques et la souplesse des modèles non paramétriques. Dans de tels modèles, la variable à expliquer Y dépend généralement de X par le biais d'un nombre fini de paramètres euclidiens $\theta_1, \ldots, \theta_k$ et d'un paramètre fonctionnel f.

Comme nous l'avons décrit dans le chapitre I.1, le nombre de variables présent en chémoinformatique est très important, chacune d'entre elles apportant une information sur le plan des structures ou des propriétés physico-chimiques. La sélection des variables est cruciale dans ce domaine pour l'interprétation des résultats tout en évitant le problème de sur-apprentissage. Nous présenterons ci-dessous les modèles de régression QSAR ainsi que des méthodes de sélection de variables.

2.1.1 Régression linéaire

Les modèles de régression linéaire sont les modèles paramétriques les plus utilisés du fait de leur facilité d'interprétation. En régression linéaire simple et multiple, les estimations au sens des moindres carrés et du maximum de vraisemblance sont souvent utilisées du fait des facilités

de calcul et de leurs bonnes propriétés pour le modèle linéaire gaussien. Le modèle de régression linéaire s'écrit :

$$Y = X\beta + \epsilon \tag{I.1}$$

où $Y = (Y_1, \ldots, Y_n)^\top$ est le vecteur des observations, X est une matrice connue de dimension $n \times p$ ayant pour lignes $x_i^\top \in \mathbb{R}^p$, $i = 1, \ldots, n$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ un vecteur d'erreurs indépendantes et $\beta = (\beta_1, \ldots, \beta_p)^\top$ les paramètres de la régression à estimer. Toutefois, ces estimateurs sont très sensibles à la présence de valeurs atypiques. En effet, la présence de valeurs atypiques peuvent écarter fortement le modèle et le rendre inapproprié.

Pour des modèles de régression gaussiens, la sélection des variables peut être effectuée en utilisant le coefficient de détermination R^2 et le test de Fisher pour les modèles emboités [32–36]. À titre d'exemple dans [32] sur un total initial de 381 descripteurs, le modèle de régression considère 5 descripteurs directement interprétables. On retrouve également des algorithmes de sélection de variable similaire (backward, forward et stepwise) dans [37–41].

Dans [42], sur 710 descripteurs, une méthode heuristique conduit à la sélection de 191 descripteurs. Par la suite, en utilisant le coefficient de détermination R^2 et le test de Fisher les auteurs obtiennent un modèle de régression avec 4 variables. Ce modèle permet de prédire l'écotoxicité de 91 composés aliphatiques et aromatiques ainsi que d'un sous-ensemble de pesticide pour une algue verte *Chlorella vulgaris*. Cette méthode de sélection de variables est très utilisée [43–46].

Dans [47], à partir d'un modèle de régression linéaire, la sélection de variables est effectuée par un algorithme génétique sur 522 descripteurs qui permet de mettre en évidence les relations structure-activité de prodiginines avec l'activité antipaludique en considérant un modèle de régression à 4 variables. D'autres articles se réfèrent également à ces méthodes [48–51].

Dans le cas particulier où le nombre de variable p est très grand, il est possible de procéder à une analyse en composantes principales ou en composantes principales à noyau et effectuer ensuite la régression sur un nombre réduit de composantes. Le problème est dans une telle configuration que le lien entre les descripteurs et l'activité peut être perdu si le pourcentage de variabilité expliqué dans l'espace de projection est faible. Il est donc préférable dans ce cas d'utiliser d'autres modèles de régression comme par exemple la régression PLS (Partial Least Squares).

2.1.2 Régression des moindres carrés partiels (PLS)

La régression PLS est une méthode non-paramétrique proposée par Wold [52]. Elle réalise un compromis entre la régression multiple de Y sur X et l'analyse en composantes principales de X. Le nombre de composantes est déterminé par validation croisée. L'algorithme classiquement utilisé s'inspire de l'algorithme NIPALS (Nonlinear Iterative Partial Least Squares) qui est itératif. La régression PLS cherche, pour un ensemble de composants (appelés vecteurs latents), à effectuer une décomposition de X et Y sous contrainte que les composantes utilisées expliquent le plus possible la covariance entre X et Y.

Dans [53], les auteurs utilisent la régression PLS afin d'estimer la stimulation du récepteur β induite par 16 composés chimiques à partir de 8 descripteurs physicochimiques et morphologiques. Le problème majeur traité ici est celui de la colinéarité entre les différents descripteurs. Les résultats montrent la performance de la régression PLS dans ce type de configuration comparé à la régression linéaire.

Afin d'améliorer les performances de la régression PLS, cette méthode a été combinée avec d'autres méthodes statistiques permettant de diminuer le sur-apprentissage et l'effet de la colinéarité entre les variables. Dans [54], Wold introduit une méthode de correction orthogonale

du signal (souvent notée OSC-PLS (Orthogonal Signal Correction) ou OPLS) qui permet de supprimer les directions de X qui sont orthogonales à Y. Afin de réduire la complexité des modèles, cet algorithme a été combiné avec de nombreux modèles de régression. Dans [55], cette méthode est utilisée pour mettre en place un modèle entre des descripteurs de peptides et une activité esterase. D'autres études [56–59] se réfèrent également à cette méthode pour construire des modèles QSAR/QSPR.

Dans [60], la régression PLS est combinée avec un algorithme génétique afin de prédire l'inhibition du récepteur 5-lipoxygenase et ainsi tester l'activité de nouveaux composés sur ce récepteur. L'algorithme génétique est souvent utilisé avec la régression PLS afin de diminuer le sur-apprentissage [61–67].

Dans [68], une analyse en composantes principales est mise en place afin de mettre en évidence par le biais d'une régression PLS le lien entre des descripteurs (topologiques, structurels, physico-chimiques, électroniques et spatiaux) et la capacité du composé à être un inhibiteur de la transcriptase inverse (pIC_{50}). Dans cet article, les auteurs mettent en évidence grâce à cette démarche le modèle suivant basé sur 7 descripteurs

$$\begin{aligned} plC_{50} &= 4.161 + 0.401 \times A\log P + 0.021 \times MolRef - 0.254 \times Dipole_{mag} \\ &\quad -18.170 \times Jurs\_FPSA\_3 + 0.014 \times Jurs\_WNSA\_1 \\ &\quad -0.069 \times Jurs\_WNSA\_3 - 0.607 \times Chiralcenters. \end{aligned}$$

Cette approche est également présentée et mise en place dans [69–71].

2.1.3 Régression inverse par tranches (SIR)

La régression inverse par tranche (Sliced Inverse Regression (SIR)) est une méthode de régression semi-paramétrique proposée par Li [72] qui s'écrit :

$$Y = f(\theta_1^\top X, \theta_2^\top X, \ldots, \theta_K^\top X, \epsilon),$$

où Y est une variable unidimensionnelle à expliquer, X est une variable explicative elliptique de dimension p ayant une espérance mathématique μ, une matrice de variance covariance $\mathbf{\Sigma}$ définie positive, $\theta_1, \ldots, \theta_K$ sont des vecteurs de paramètres réels de dimension p, inconnus et linéairement indépendants, avec $K < p$, ε est un terme d'erreur aléatoire indépendant de X, aucune hypothèse n'est faite sur la distribution de ε, f est le paramètre fonctionnel à valeur dans $\mathbb{R}$, inconnu et arbitraire.

La valeur de K étant supposée strictement inférieure à p, ce modèle est donc aussi un modèle permettant une réduction de dimension de l'espace des variables explicatives. Dans un tel modèle, on appelle indices les termes $(\theta_k^\top X)$, $k = 1, \ldots, K$. Ainsi pour $K > 1$, on considère un modèle multi-indices tandis que pour $K = 1$ ce modèle a un seul indice. Ainsi, l'étude de la relation entre Y et X se réduit à l'examen de la structure qui existe entre Y et les K indices $\theta_1^\top X, \ldots, \theta_k^\top X$.

Le point important des modèles SIR se trouve dans la définition et la caractérisation de l'espace EDR (Effective Dimension Reduction) qui est à la base des méthodes SIR [73–75].

Une étude a montré des performances analogues à celle de la régression PLS (modèles QSPR [76]). Dans le domaine des QSAR, d'autres études ont montré des performances intéressantes de SIR sur cette fois-ci des propriétés biologiques [76, 77].

2.1.4 Régression LASSO et Ridge

Considérons le modèle de régression linéaire de la section 2.1.1 défini par V.1. En régression Ridge, on cherche à estimer le modèle en minimisant la somme des carrés des termes d'erreurs pénalisés. L'estimation du vecteur des paramètres de régression se fait donc par la minimisation

de :

$$\sum_{i=1}^{n}\left(Y_i-\beta_0-\sum_{j=1}^{p}X_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

avec $\lambda \in \mathbb{R}^+$. Ici, plus la valeur de λ est élevée plus le niveau de contrôle est important ce qui amène à une plus forte réduction des coefficients. La solution à ce problème de minimisation est

$$\hat{\beta}_0=\frac{1}{n}\sum_{i=1}^{n}y_i$$

et

$$\hat{\beta}=(\mathrm{X}^\top\mathrm{X}+\lambda\mathrm{I})^{-1}\mathrm{X}^\top\mathrm{Y}$$

avec I la matrice identité de taille $p \times p$ pour les coefficients 1 à p. Pour la suite, nous allons considérer que la matrice X est centrée.

En régression LASSO, on s'intéresse à éliminer les variables non pertinentes pour le modèle en les ramenant à 0. Afin d'estimer les paramètres β, il faut alors minimiser la fonction :

$$\|\mathrm{Y}-\sum_{j=1}^{p}\beta_j\mathrm{X}_j\|^2+\lambda\sum_{i=1}^{p}|\beta_j|$$

où comme pour la régression ridge $\lambda \geq 0$ est le paramètre de pénalisation.

Ces 2 méthodes ont été mises en avant dans de nombreux articles en chémoinformatique. Les méthodes de régression LASSO et Ridge ont été utilisées et comparées pour modéliser l'activité des hormones juvéniles de 304 composés à partir de 1201 descripteurs [78]. On trouve également d'autres exemples d'application de ces méthodes dans [79–83].

2.1.5 Régression par Projection Pursuit ou Directions Révélatrices

La régression par Projection Pursuit est un modèle semi-paramétrique introduit par Friedman et Stuetzle en 1981 [84] basé sur un algorithme additif et une réduction dimentionnelle

par projection des entrées X sur un nombre de m facteurs. L'objectif est donc d'étudier un ensemble de données de dimension élevée, par recherche automatisée de projections "intéressantes" suceptibles de révéler sa structure sur des sous espaces de dimension réduite. Pour ce faire, les algorithmes de régression par projection pursuit [84] visent à approcher la fonction de régression $f(.)$ par une somme de fonction tel que :

$$g^p(x) = \sum_{i=1}^{p} g_i(\alpha_i^\top x) \tag{I.2}$$

où α_i est une matrice orthonormale $(m \times n)$ et p est le nombre de fonctions utilisées pour approximer la fonction $f(.)$ en x.

À titre d'exemple, une étude comparative de 3 méthodes (régression Projection Pursuit, Support Vector Machine et régression linéaire) basée sur 8 descripteurs sélectionnés par une méthode heuristique montre une bonne performance, performance comparable avec la régression SVM [85]. Pour d'autres exemples d'applications en chémoinformatique, voir [34, 86–89].

2.1.6 Régression fonctionnelle

Ce modèle est un modèle de régression non-paramétrique. Nous disposons d'un échantillon composé de n couples indépendants de variables aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$ et nous considérons le modèle de régression non paramétrique donné, pour $i = 1, \ldots, n$, par

$$Y_i = m(X_i) + \varepsilon_i. \tag{I.3}$$

où m est une fonction inconnue à estimer et un terme aléatoire d'erreur ε de loi inconnue et indépendant de X. Nous décrivons deux estimateurs non paramétriques de la fonction $m(.)$. Le premier estimateur est l'estimateur de Nadaraya-Watson [30, 31]. Il est construit à partir d'une fonction noyau K et d'une fenêtre h_n, de manière similaire à l'estimateur à noyau de la

fonction de densité de probabilité [90]. Cet estimateur de la densité f de T s'écrit :

$$\hat{f}_n(x) = \frac{1}{n\,h_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{I.4}$$

ou dans sa forme récursive :

$$\tilde{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right). \tag{I.5}$$

La fenêtre h_n désigne une suite de nombres réels strictement positifs vérifiant $(C1)$ $h_n \to 0$ et $n\,h_n \to \infty$ lorsque $n \to \infty$. Le noyau est une fonction mesurable, positive et bornée satisfaisant $(C2)$ $\int K(x)\,dx = 1$, $\int x\,K(x)\,dx = 0$, $\int |x|\,K(x)\,dx < +\infty$ et $\int K^2(x)\,dx = \tau^2$.

L'estimateur de Nadaraya-Watson s'écrit sous la forme d'une moyenne pondérée des valeurs $(Y_1, \ldots, Y_n)$. Il est donné par :

$$\hat{m}_n(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} K\left(\dfrac{x - X_i}{h_n}\right) Y_i}{\displaystyle\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)} & \text{si } \displaystyle\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \neq 0, \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} Y_i & \text{sinon.} \end{cases} \tag{I.6}$$

On propose également d'utiliser l'estimateur de Nadaraya-Watson récursif [91] défini par :

$$\tilde{m}_n(x) = \begin{cases} \dfrac{\displaystyle\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right) Y_i}{\displaystyle\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right)} & \text{si } \displaystyle\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right) \neq 0, \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} Y_i & \text{sinon.} \end{cases} \tag{I.7}$$

Ces deux estimateurs de m sont donc dépendants du choix de la fenêtre et du noyau. Le noyau K détermine "la forme du voisinage" autour du point x et la fenêtre h_n contrôle "la taille de ce voisinage", c'est-à-dire grossièrement le poids des observations pris pour effectuer

le calcul de l'estimateur en x. Le choix du paramètre h_n est par conséquent un point crucial pour la qualité de l'estimation. Cependant, le choix du noyau permet aussi de réduire le biais des estimateurs en se basant sur les propriétés de régularité de la fonction de lien.

Nous rappelons ici les principales propriétés asymptotiques des estimateurs de Nadaraya-Watson et Nadaraya-Watson récursif. Nous introduisons tout d'abord les notations $h_n = n^{-\alpha}$ et $\sigma^2(x) = \mathrm{var}\,(Y \mid X = x)$. Nous ajoutons deux conditions de régularité : (*C3*) la fonction de lien m et la densité f sont bornées et deux fois continûment dérivables sur R et (*C4*) $E(Y^2) < \infty$.

Théorème I.1. *Sous les conditions C1 – C4 et pour tout $\alpha \in [1/5, 1[$, à chaque point de continuité de $\sigma^2(x)$ et pour tout $x \in R$ tel que $f(x) > 0$, nous avons quand $n \to \infty$:*

1. $$\hat{m}_n(x) \xrightarrow{ps} m(x).$$

2. $$\sqrt{nh_n}\,(\hat{m}_n(x) - m(x)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(B(\hat{m}_n(x)), \frac{\sigma^2(x)\,\tau^2}{f(x)}\right),$$

   où le biais de $\hat{m}_n(x)$ est

$$B(\hat{m}_n(x)) = \int u^2 K(u)\,du \left(m''(x) + 2m'(x)\frac{f'(x)}{f(x)}\right).$$

Théorème I.2. *Sous les conditions C1 – C4 et pour tout $\alpha \in]1/3, 1[$, à chaque point de continuité de $\sigma^2(x)$ et pour tout $t \in R$ tel que $f(x) > 0$, nous avons quand $n \to \infty$:*

1. $$\tilde{m}_n(x) \xrightarrow{ps} m(x).$$

2. $$\sqrt{nh_n}\,(\tilde{m}_n(x) - m(x)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2(x)\,\tau^2}{f(x)(1+\alpha)}\right).$$

Il est possible d'estimer la variance de la loi normale limite. Pour cela, la densité marginale f est estimée par (I.4) pour l'estimateur de Nadaraya-Watson et par (I.5) pour l'estimateur de Nadaraya-Watson récursif. La variance conditionnelle $\sigma^2(x)$ est estimée respectivement pour

l'estimateur de Nadaraya-Watson et Nadaraya-Watson récursif par :

$$\hat{\sigma}^2(x) = \frac{1}{\hat{f}_n(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)(Y_i - \hat{m}_n(x))^2 \quad \text{et} \quad \tilde{\sigma}^2(x) = \frac{1}{\tilde{f}_n(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_i}\right)(Y_i - \hat{B}_n(x))^2,$$

avec $\hat{B}_n(x) = \tilde{m}_n(x)$ et $\hat{B}_n(x) = \check{m}_n(x)$.

Le choix de ce paramètre est crucial pour nos trois estimateurs. En pratique, ce paramètre est choisi comme un compromis entre la variance et le biais de l'estimation. Une importante littérature est consacrée à ce sujet, et en particulier aux méthodes de sélection automatique par minimisation d'un critère. Nous utilisons comme critère la méthode de la validation croisée [92, 93] qui consiste à minimiser par rapport à h la fonction

$$CV(h) = \sum_{i=1}^{n} \left(Y_i - \hat{m}_{(-i)}(X_i; h)\right)^2$$

où $\hat{m}_{(-i)}(X_i; h)$ désigne l'estimateur de Nadaraya-Watson ou Nadaraya-Watson récursif de la fonction de lien au point X_i calculé sur l'échantillon privé du couple (X_i, Y_i).

Dans [94], une étude comparative des capacités prédictives de plusieurs modèles (régression linéaire multiple, réseaux de neurones, régression polynomiale locale et régression à noyau) a été réalisée. En partant de 538 inhibiteurs de la tyrosine kinase, la comparaison montre l'interêt de l'estimateur de Nadaraya-Watson. D'autres articles se réfèrent à ce type de modèle [95–100].

2.2 Deep learning et Machine learning

2.2.1 Réseaux de neurones

Afin d'ajuster des données à un modèle, les réseaux de neurones traitent les informations d'entrée et génèrent des modèles de relations cachés. L'un des avantages des réseaux de neurones est qu'ils sont capables de modéliser des systèmes non linéaires. Les inconvénients comprennent une tendance à sur-ajuster les données et à avoir des difficultés à déterminer quels descripteurs

sont les plus importants dans le modèle. Dans les récentes études QSAR/QSPR, RBFNN (réseaux de neurones à fonction de base radiale) et GRNN (General Regression Neural Network), les réseaux de neurones sont fréquemment utilisés [101].

Les réseaux de neurones à fonction de base radiale se composent de trois couches : une couche d'entrée, une couche cachée et une couche de sortie. La couche d'entrée ne traite pas l'information, elle distribue les vecteurs d'entrée vers la couche cachée. Chaque neurone sur la couche cachée emploie une fonction de base radiale comme une fonction de transfert non linéaire pour opérer sur les données d'entrée. En général, il existe plusieurs fonctions de base radiale (RBF) : linéaire, cubique, fonction sline du type "plaque mince", gaussien, multi-quadratique et inverse multi-quadratique. La transformation non linéaire avec RBF gaussien dans la couche cachée est donnée par :

$$k_j(x) = \exp\left(\frac{-(x-\mu_j)^2}{\sigma_j^2}\right)$$

où, k_j dénote la sortie de l'unité $j$ème RBF, μ_j le centre pour la $j$ème fonction RBF et σ_j la largeur pour la $j$ème fonction RBF. Dans une telle méthode, les fonctions de sortie sont linéaires et sont données par :

$$y_l(x) = \sum_i w_{li} k_i(x) + b_l$$

où, y_l est la $l$ème unité de sortie du vecteur d'entrée x et w_{li} est le rapport entre la $l$ème unité de sortie, la $l$ème couche, et b_l est le biais. La procédure d'apprentissage pour l'utilisation d'une fonction RBF implique une sélection des centres, de la largeur et des poids.

À titre d'exemple, en plus de la référence générale [101], [102] montre l'intérêt de l'utilisation des réseaux de neurones dans les études QSAR pour la prédiction des propriétés physicochimiques et (éco)toxicologiques dans le cadre de la législation REACH. Elle donne de meilleures performances que les modèles de régressions linéaires. D'autres articles se réfèrent aussi à cette

méthode en chémoinformatique [42, 103–111].

2.2.2 Classification et régression SVM

En 1964, Vapnik et Chervonenkis [112] ont mis en place une méthode déterminant un hyperplan séparateur dit de marge optimale pour la séparation de deux classes dans un espace Hilbertien associé au produit scalaire noté $\langle ., . \rangle$.

Classification SVM Pour deux classes, l'objectif est de trouver un classificateur qui va séparer les données en maximisant la distance entre ces deux classes. En classification SVM, ce classificateur est un classificateur linéaire appelé hyperplan qui permet la séparation des deux ensembles de points. Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support (Figure I.1). Il existe une multitude d'hyperplans

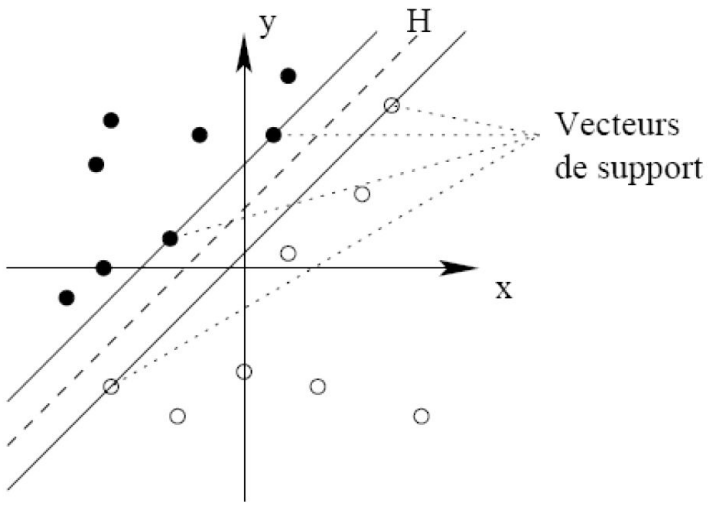


Figure I.1 – Hyperplan séparateur et vecteurs de support.

mais la propriété remarquable de la classification SVM est que cet hyperplan doit être optimal. L'idée intuitive consiste à chercher un hyperplan dont la distance minimale aux données d'apprentissage est maximale. Cette distance est appelée "marge" et comme on cherche à maxi-

miser cette marge, on parlera de séparateurs à vaste marge comme représenté dans la Figure I.2 ci-dessous. Nous considérons n couples $(x_1, y_1), \ldots, (x_n, y_n)$ où $x_i \in \mathbb{R}^p$ et $y_i \in \{-1, 1\}$ pour

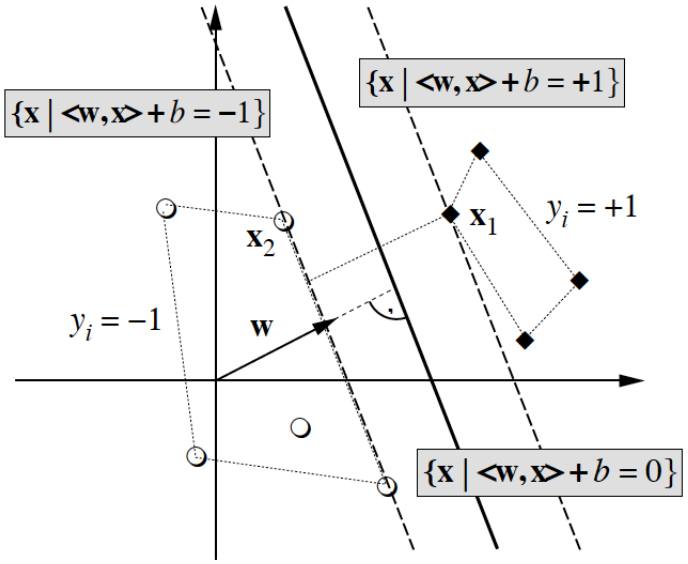


Figure I.2 – Séparateur à vaste marge.

$i = 1, \ldots, n$. Un hyperplan canonique séparateur H est défini par :

$$\langle w, x \rangle + b = 0.$$

On peut alors choisir w et b tel que le point x le plus proche de H satisfasse :

$$\langle w, x \rangle + b = \begin{cases} 1 & \text{si } x \text{ est positif,} \\ -1 & \text{sinon.} \end{cases}$$

On en déduit que $x_1 > 0$ et $x_2 < 0$:

$$\langle w, x_1 - x_2 \rangle = \langle w, x_1 \rangle - \langle w, x_2 \rangle = (1 - b) - (-1 - b) = 2,$$

et par conséquent la marge de H, notée M est donnée par l'expression :

$$M = \left\langle \frac{w}{\|w\|}, x_1 - x_2 \right\rangle = \frac{2}{\|w\|},$$

où $\|w\| = \sqrt{w_1^2 + \ldots + w_n^2}$. On voit donc que plus $\|w\|$ est petite, plus la marge de l'hyperplan canonique correspondante est grande. Pour cette raison, afin de trouver l'hyperplan qui sépare le mieux les données, il faut trouver celui qui respecte les conditions d'un hyperplan canonique et pour lequel $\|w\|$ est minimale. Par conséquent, afin de trouver l'hyperplan canonique qui sépare les données avec la plus grande marge possible, il faut minimiser

$$\frac{1}{2}\|w\|^2$$

sous les contraintes $y_i\left(\langle w, x_i\rangle + b\right) \geq 1, i = 1, \ldots, n$.

Ces contraintes assurent d'une part que l'hyperplan sépare les données correctement, et d'autre part qu'il est canonique. En effet pour $i = 1, \ldots, n$, $y_i\left(\langle w, x_i\rangle + b\right) > 0$ si et seulement si $\text{signe}\left(y_i\left(\langle w, x_i\rangle + b\right)\right) = y_i$, donc si et seulement si x_i est du bon côté de l'hyperplan. Ainsi, l'hyperplan doit correctement séparer les données. Ensuite, on peut montrer qu'imposer $y_i\left(\langle w, x_i\rangle + b\right) \geq 1$ assure que pour toutes les données qui ne sont pas sur la marge, $|\langle w, x_i\rangle + b| > 1$ et que $|\langle w, x_i\rangle + b| = 1$ pour les données sur la marge, donc que l'hyperplan est canonique.

Une propriété intéressante de ce problème est que la fonction :

$$f(w) = \|w\|^2 = w_1^2 + \ldots + w_n^2$$

est une fonction strictement convexe. Ceci assure qu'il n'y a pas de minimum local et qu'il n'existe qu'une unique solution optimale. On note, pour $i = 1, \ldots, n$, $\alpha_i \geq 0$ les multiplicateurs

de Lagrange associés aux contraintes. Le Lagrangien de la fonction s'écrit alors :

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\langle w, x_i\rangle + b\right) - 1\right). \tag{I.8}$$

On cherche alors à minimiser $\mathcal{L}(w, b, \alpha)$ par rapport aux variables primales w, b et à le maximiser par rapport aux variables duales α_i. Lorsque l'on atteint le point optimal de ce problème de minimisation on doit avoir $\frac{\partial \mathcal{L}(w,b,\alpha)}{\partial b} = 0$ et $\frac{\partial \mathcal{L}(w,b,\alpha)}{\partial w} = 0$ et donc $\sum_{i=1}^{n} \alpha_i y_i = 0$ et $\sum_{i=1}^{n} \alpha_i y_i x_i = w$. En remplaçant ces valeurs dans le Lagrangien, on obtient alors le problème dual :

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle$$

sous les contraintes $\alpha_i \geq 0$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$. Dans le problème dual, le coefficient b n'apparaît pas. Or on sait que lorsque $\langle w, x\rangle + b > 1$ le coefficient $\alpha_i = 0$ et que $\langle w, x\rangle + b = 1$ pour les vecteurs supports. On calcule alors une moyenne de ces paramètres pour les vecteurs supports pour estimer b. L'estimation de la marge est $M = \frac{2}{\|w\|} = \left(\sum_{i \in SV} \alpha_i\right)^{-1/2}$ où SV désigne l'espace des vecteurs supports. La frontière de décision s'écrit alors :

$$\langle w, x\rangle + b = \sum_{i \in SV} \alpha_i y_i \langle x_i, x\rangle + b$$

Nous nous sommes jusqu'à présent intéressés au cas où les données sont linéairement séparables. Lorsque ce n'est plus le cas, on introduit une variable de relâchement pour les contraintes ξ_i qui pénalisera les relâchements dans la fonction objective. Le problème devient alors

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

sous les contraintes $y_i \left(\langle w, x_i\rangle + b\right) \geq 1 - \xi_i$ et $\xi_i \geq 0$ avec C un terme de pénalisation. Le Lagrangien s'écrit alors

$$\mathcal{L}(w, b, \xi_i, \alpha, \nu) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\langle w, x_i\rangle + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{n} \nu_i \xi_i.$$

Sous des conditions d'optimalité on sait que $\sum_{i=1}^{n} \alpha_i y_i = 0$, $w = \sum_{i=1}^{n} \alpha_i y_i x_i = 0$ et $C - \alpha_i - \nu_i = 0$ pour $i = 1, \ldots, n$. En remplaçant ces valeurs dans le Lagrangien, on obtient le problème dual suivant :

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

sous les contraintes $0 \leq \alpha_i \leq C$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Considérons maintenant le cas d'une séparation non linéaire. Dans ce cas, la complexité de l'hyperplan séparateur mis en place dans le cas linéaire n'est pas suffisant pour séparer correctement les données. Afin de pallier à ce problème, on considère une transformation non-linéaire $\phi(.)$ qui projette les données dans un espace de plus grande dimension de façon à ce que les données dans l'espace transformé soient linéaires. Dans ce cas l'hyperplan séparateur s'écrit :

$$\langle w, \phi(x) \rangle + b.$$

En remplaçant dans le Lagrangien les données par la transformation $\phi(.)$ on obtient :

$$\mathcal{L}(w, b, \xi_i, \alpha, \nu) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\langle w, \phi(x_i) \rangle + b \right) - 1 + \xi_i \right) - \sum_{i=1}^{n} \nu_i \xi_i.$$

Sous les conditions $\sum_{i=1}^{n} \alpha_i y_i = 0$, $w = \sum_{i=1}^{n} \alpha_i y_i x_i = 0$ et $C - \alpha_i - \nu_i = 0$, $i = 1, \ldots, n$ on obtient :

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle$$

sous les contraintes $0 \leq \alpha_i \leq C$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$. Il est difficile de construire l'hyperplan dans l'espace transformé. En 1992, Boser et al. [113], en utilisant le théorème de Mercer ont trouvé un moyen de construire l'hyperplan optimal dans l'espace transformé sans utiliser une forme explicite de l'espace transformé.

Theorem I.1 (Théorème de Mercer)**.** *Considérons $x \in X$ reporté dans un vecteur $z \in Z$ un espace de Hilbert. Alors il existe dans l'espace X une fonction symétrique définie positive notée*

$K(x_i, x_j)$ tel que :

$$\langle z_i, z_j \rangle = K(x_i, x_j). \tag{I.9}$$

Et pour toute fonction définie positive $K(x_i, x_j)$ dans l'espace X, il existe une transformation de X vers Z telle que cette fonction décrive un produit scalaire dans l'espace Z.

Le problème devient alors :

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

sous les contraintes $0 \leq \alpha_i \leq C$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Nous précisons quelques noyaux utilisés en pratique :

— Noyau linéaire

$$K(x_i, x_j) = \langle x_i, x_j \rangle,$$

— Noyau polynomiale

$$K(x_i, x_j) = (\beta \langle x_i, x_j \rangle + \tau)^d,$$

où β est le paramètre d'échelle, τ le décalage et d le degré.

— Noyau Gaussien radial (RBF= Radial Basis Function)

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\,\sigma^2}\right),$$

où σ est la largeur de fenêtre.

— Noyau tangente hyperbolique

$$K(x_i, x_j) = \tanh\left(\beta \langle x_i, x_j \rangle + \tau\right),$$

— Noyau ANOVA radial basis

$$K(x_i, x_j) = \left(\sum_{k=1}^{q} \exp\left(-\sigma (x_{ik} - x_{jk})^2\right)\right)^d.$$

— Noyau de Tanimoto

$$K^T(x_i, x_j) = \frac{K(x_i, x_j)}{K(x_i, x_i) + K(x_j, x_j) - K(x_i, x_j)}$$

avec $K(.,.)$ un produit scalaire.

Régression SVM

En 1998, Vapnik a introduit dans [112] une fonction de perte

$$u_\epsilon = \begin{cases} |u| - \epsilon & \text{if} \quad |u| \geq \epsilon \\ 0 & \text{if} \quad |u| < 0 \end{cases} \tag{I.10}$$

qui permet de transférer certaines propriétés de SVM au problème de régression. Considérons les couples $(x_1, y_1), \ldots, (x_n, y_n)$ où $x \in X$ un vecteur et $y \in \mathbb{R}$. Comme pour la classification, on utilise une transformation $\phi(.)$. On cherche alors à approximer les régressions linéaires par

$$y = \langle w, \phi(x) \rangle + b$$

où w et b sont les paramètres à définir. Afin de résoudre le problème de régression, on cherche à minimiser

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} |y_i - \langle w, \phi(x) \rangle - b|_\epsilon.$$

afin de minimiser cette fonction, on cherche à minimiser la fonction équivalente

$$\frac{1}{2}\langle w, w \rangle + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

sous les contraintes

$$y_i - \langle w, \phi(x_i) \rangle - b \leq \epsilon + \xi_i^*, \quad \xi_i^* \geq 0, \quad i = 1, \ldots, n,$$

$$\langle w, \phi(x_i) \rangle + b - y_i \leq \epsilon + \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$$

Pour résoudre ce problème, on construit le Lagrangien

$$\begin{aligned}&\frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i\left[y_i - \langle w, \phi(x_i)\rangle - b + \epsilon + \xi_i\right]\\&-\sum_{i=1}^{n}\alpha_i^*\left(\langle w, \phi(x_i)\rangle + b - y_i + \epsilon + \xi_i^*\right) - \sum_{i=1}^{n}(\beta_i\xi_i + \beta_i\xi_i^*).\end{aligned} \tag{I.11}$$

La minimisation de cette fonction par rapport à w, b, ξ et ξ^* conduit aux équations

$$w = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\phi(x_i), \tag{I.12}$$

$$\sum_{I=1}^{n}(\alpha_i - \alpha_i^*) = 0 \tag{I.13}$$

et

$$\alpha_i^* + \beta_i^* = C, \quad \alpha_i + \beta_i = C, \tag{I.14}$$

où α, α^*, β, $\beta^* > 0$ désignent les multiplicateurs de Lagrange. La fonction s'écrit :

$$y = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)K(x_i, x) + b. \tag{I.15}$$

En introduisant (I.12), (I.13), (I.14) dans (I.11), on obtient

$$-\sum_{i=1}^{n}\epsilon(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n}y_i(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i,j}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j). \tag{I.16}$$

Afin de déterminer α_i et $\alpha_i^\star$ dans l'équation (I.15), il faut maximiser (I.16) sous les contraintes

$$\sum_{I=1}^{n}\alpha_i^* = \sum_{i=1}^{n}\alpha_i, \quad 0 \leq \alpha_i \leq C, \quad 0 \leq \alpha_i^* \leq \epsilon, \quad i = 1, \ldots, n.$$

Afin d'améliorer les performances de cette méthode, la régression SVM est combinée à d'autres approches. Dans [114–119], la LS-SVM (Least Square Support Vector Machine) développée par Suykens et al. [120] est utilisée pour les performances de la classification SVM sur des problématiques liées aux modèles QSAR. Dans [121–124], les auteurs mettent en place des modèles QSAR basés sur une approche SVM utilisant des algorithmes génétiques pour la

sélection composante dans l'espace transformé.

2.2.3 Forêt aléatoire

En apprentissage, les méthodes de forêt aléatoire [125] sont très utilisées en régression et en classification. Elles fonctionnent en plusieurs étapes.

- Tirage de n échantillons par bootstrap en partant des données,
- Pour chaque tirage, mettre en place m arbres de régression non élagués, et sélectionner l'arbre pour lequel l'erreur est minimum.
- Prédiction pour les données en agrégeant les prédictions des n arbres.

Une estimation de l'erreur peut être obtenue sur les composés n'étant pas présents dans le bootstrap et permet d'avoir une interprétation de l'importance des variables dans le modèle.

De nombreux articles se réfèrent à cette méthode, pour analyser les propriétés des composés chimiques. Par exemple, dans [126], les forêts aléatoires sont appliquées à l'analyse QSAR afin de déterminer la toxicité aquatique de 1093 composés répartis en un ensemble d'apprentissage de 644 composés, un ensemble de validation de 339 composés et un ensemble de tests de 110 composés sur une algue unicellulaire (*Tetrahymena pyriformis*). Ces composés sont décrits par 6000 descipteurs dont le log(P) et la réfraction moléculaire. On se réfère également à cet algorithme dans [127–129] et plus récemment dans [130–133].

Chapitre II

Méthodes statistiques d'apprentissage robuste

Du fait de la présence attendue de valeurs atypiques ou extrêmes dans les données, il est important de considérer des estimateurs *robustes*. Nous pouvons trouver différentes définitions de la robustesse mais toutes procèdent du même esprit. Ainsi, P. Huber écrit en 1972 [134] : "La robustesse est une sorte d'assurance : je suis prêt à payer une perte d'efficacité de 5 à 10% par rapport au modèle idéal pour me protéger de mauvais effets de petites déviations de celui-ci : je serai bien sûr heureux que ma procédure statistique fonctionne bien sous de gros écarts, mais je n'y prête pas réellement attention car faire de l'inférence à partir d'un modèle aussi faux n'a que peu de significations concrètes".

Dans beaucoup d'applications (astronomie, biologie, chimie, médecine, physique, etc), les données sont contaminées par des valeurs atypiques qui proviennent d'erreurs dues à l'environnement expérimental ou de tout autre cause, tout aussi triviale qu'une erreur d'enregistrement ou de lecture. L'estimation au sens des moindres carrés (estimation L^2) est souvent utilisée du fait des facilités de calcul et de ses bonnes propriétés pour le modèle linéaire gaussien (estimateur sans biais de variance minimale) ; toutefois, ces estimateurs sont très sensibles à la présence de valeurs atypiques.

En revanche la robustesse de la *médiane* (archétype d'estimation L^1) est connue de longue date. Plus récemment, en 1964, Huber [135] a publié un article de référence sur l'estimation

robuste du paramètre de location. Ces vingt dernières années, un effort théorique considérable a été déployé pour construire des méthodes statistiques robustes. Mentionnons simplement ici que le travail de Huber a été étendu aux modèles linéaires par Andrews en 1974 [136], Bickel en 1975 [137], Huber en 1973 et 1981 [138, 139], Hampel et al. en 1986 [140] et Jurečková et Sen en 1996 [141].

1 Quantile de régression linéaire et notation

Considérons le modèle de régression linéaire qui s'écrit :

$$Y = X\beta + \epsilon \tag{II.1}$$

où $Y = (Y_1, \ldots, Y_n)^\top$ est le vecteur des observations, X est une matrice connue de dimension $n \times p$ ayant pour lignes $x_i^\top \in \mathbb{R}^p$, $i = 1, \ldots, n$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ un vecteur d'erreurs indépendantes et $\beta = (\beta_1, \ldots, \beta_p)^\top$ les paramètres de la régression à estimer.

1.1 Notations

Nous introduisons tout d'abord quelques notations et présentons les conditions de régularité sur la fonction de répartition F ($C1-C3$) de l'erreur ε_1, sur la suite de matrices $\boldsymbol{X}_n$ ($C4-C7$) et sur l'estimation ($C8-C10$) permettant la construction de cette procédure séquentielle. Soit :

- $\boldsymbol{D}_n = n^{-1}\boldsymbol{X}_n^\top \boldsymbol{X}_n \equiv \left(d_{ij}^n\right)_{1\leq i,j\leq p}$ et $\boldsymbol{D}_n^{-1} \equiv (d_n^{ij})_{1\leq i,j\leq p}$ son inverse dont on suppose l'existence.
- $\mathbb{I}(\mathcal{P})$ prend la valeur 1 ou 0 selon que la condition $\mathcal{P}$ est vérifiée ou non.
- θ désigne un nombre réel de l'intervalle $]0, 1[$.
- $\psi_\theta(x) = \theta - \mathbb{I}(x < 0)$ et $\rho_\theta(x) = x\psi_\theta(x)$.
- Q la *fonction quantile* de ε_1 vérifiant $Q(1/2) = 0$.
- q la *fonction densité du quantile* définie par $q(\theta) = Q'(\theta)$.

— $\sigma_\theta = \sqrt{\theta\,(1-\theta)}\,q(\theta)$ et $\sigma = \sigma_{1/2}$.

— $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée réduite.

— $\|.\|$ dénote la norme euclidienne.

C1. F est de classe $\mathcal{C}^2$ et sa dérivée f est strictement positive sur $\{x \ : \ 0 < F(x) < 1\}$.

C2. Il existe $c > 0$, $\theta_0 \in]0, 1/2[$ et $a \in]0, 1/4[$ tel que $q(\theta) \leq \dfrac{c}{(\theta\,(1-\theta))^{1+a}}$ pour $\theta \in]0, \theta_0[\cup]1-\theta_0, 1[$.

C3. Il existe c et K positifs tel que $\left|\dfrac{f'(x)}{f(x)}\right| \leq c\,|x|$ pour $|x| \geq K$.

C4. Soit a le paramètre défini dans la condition *C2*. Nous supposons que $\max\limits_{1\leq i\leq n} \|\boldsymbol{x}_i\| = O\left(n^{(2(b-a)-\delta)/(1+4b)}\right)$ pour b et δ positifs tel que $0 < b-a < \epsilon/2$ quand $n \to \infty$. Dans ce qui suit nous notons $\theta_n^* = n^{-1/(2(1+4b))}$.

C5. $\lim\limits_{n\to\infty} \boldsymbol{D}_n = \boldsymbol{D}$, matrice définie positive.

C6. $n^{-1}\sum\limits_{i=1}^{n} \|\boldsymbol{x}_i\|^4 = O(1)$ quand $n \to \infty$.

C7. $x_{i1} = 1$ pour $i = 1, \ldots, n$.

Soit $(\nu_n)_{n\geq 1}$ une suite décroissante de nombres positifs (*tailles de fenêtre*). Les conditions que l'on demandera à $(\nu_n)_{n\geq 1}$ appartiendront à la famille suivante :

C8 . $n\,\nu_n \to \infty$ quand $n \to \infty$.

C8'. $n\,\nu_n^2 \to \infty$ quand $n \to \infty$.

C9 . $n\,\nu_n^3 \to 0$ quand $n \to \infty$.

La dernière condition concerne la fonction noyau.

C10. k est une fonction continue à support compact telles que $\int k(v)\,dv = 0$ et $\int v\,k(v)\,dv = -1$; on note aussi $K(x) = \int_{-\infty}^{x} k(y)\,dy$ et $\overline{K} = \int K^2(x)\,dx$.

La condition *C2* indique en particulier que les queues de distribution des erreurs ne sont pas trop épaisses, et implique l'existence de moments d'ordre 4. Toutes ces conditions sont raisonnables du point de vue des applications et sont entre autres satisfaites par les lois uniforme,

normale, logistique, double exponentielle, Student avec 5 degrés de liberté ou plus, etc.

En 1978, Koenker et Basset ont proposé le concept de "quantile de régression" permettant la détermination des $L-$estimateurs. On appelle *θ-quantile de régression* toute solution du problème de minimisation

$$\widehat{\boldsymbol{\beta}}(\theta) \equiv \widehat{\boldsymbol{\beta}}^n(\theta) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\theta \left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right). \tag{II.2}$$

Un cas particulier de cette classe d'estimateurs (obtenu pour $\theta = 1/2$) est l'estimateur L^1 qui s'obtient par résolution du problème de minimisation (II.2). Ce problème de minimisation peut être résolu en particulier par résolution du programme linéaire (ou de son dual)

$$\widehat{\boldsymbol{\beta}}^n(\theta) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{\theta\, \mathbf{1}^\top \boldsymbol{r}^+(\boldsymbol{\beta}) + (1-\theta)\, \mathbf{1}^\top \boldsymbol{r}^-(\boldsymbol{\beta})\right\},$$

sous les contraintes

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{r}^+ - \boldsymbol{r}^- = \boldsymbol{Y},$$

$$(\boldsymbol{\beta}, \boldsymbol{r}^+, \boldsymbol{r}^-) \in \mathbb{R}^p \times \mathbb{R}_+^{2n},$$

où

$$\boldsymbol{r}^-(\boldsymbol{\beta}) = \max\left(\mathbf{0}, \boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{Y}\right) \quad \text{et } \boldsymbol{r}^+(\boldsymbol{\beta}) = \max\left(\mathbf{0}, \boldsymbol{Y} - \boldsymbol{X}\,\boldsymbol{\beta}\right).$$

Une autre méthode consiste à résoudre ce problème de minimisation par la méthode itérative "IRLS" (*Iteratively Reweighted Least Squares*). Considérons le problème de minimisation II.2. La solution à ce problème de minimisation est équivalente à celle d'un système à p équation linéaire

$$\sum_{i=1}^{n} x_{ij} \psi_\theta \left(Y_i - \sum_{k=1}^{p} x_{ik}\beta_k\right) = 0 \tag{II.3}$$

avec $j = 1, \ldots, p$. En appliquant la transformation $w(x) = \psi(x)/x$ on obtient

$$\sum_{i=1}^{n} x_{ij} w \left(Y_i - \sum_{k=1}^{p} x_{ik}\beta_k\right)\left(Y_i - \sum_{k=1}^{p} x_{ik}\beta_k\right) = 0$$

où les poids w dépendent de la valeur de β. Pour la suite, nous allons considérer que $\beta = \beta^0$ avec $\beta^0 \in \mathbb{R}^p$ un nombre fixé et V une matrice diagonale décrivant les poids tel que

$$V \equiv V_n(\beta^0) = \begin{bmatrix} w\left(r_1(\beta^0)\right) & 0 & \dots & 0 \\ 0 & w\left(r_2(\beta^0)\right) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w\left(r_n(\beta^0)\right) \end{bmatrix}$$

alors

$$\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{Y}$$

Si on note par $\hat{\beta}_0$ la solution initiale de II.3 et $V(\hat{\beta}_0)$ la matrice de poids correspondante, alors on peut écrire$\left(\hat{\beta}_k\right)$ la suite de solution de l'équation

$$X^\top V(\hat{\beta}_{k-1}) X\beta = X^\top V(\hat{\beta}_{k-1}) Y.$$

La suite des valeurs de $\left(\hat{\beta}_k\right)_k$ converge vers la vraie valeur β. Un critère d'arrêt est donné par

$$\|\hat{\beta}_{l+1} - \hat{\beta}_l\|_{L^1} < \delta$$

où δ est une valeur positive fixée.

Les deux méthodes de calcul précédentes sont coûteuses en temps calcul pour des grands échantillons. En effet, comme la méthode du simplexe consiste à parcourir l'ensemble des sommets d'un polyèdre convexe en choisissant à chaque fois l'arête correspondant à la pente la plus forte, cet algorithme de résolution n'est pas applicable lorsque n est grand. Typiquement, il est conseillé d'utiliser pour $n > 10^5$ la méthode du point intérieur développée par Portnoy et Koenker [142]. Cette méthode consiste à se ramener à un programme d'optimisation standard facilement résoluble en utilisant la méthode de Newton.

L'étude du comportement asymptotique de cet estimateur ou plus généralement de cette classe d'estimateurs repose sur la décomposition suivante :

Théorème II.1 (Gutenbrunner et Jurečková, 1992)**.** *Considérons le modèle (II.1). Soit* $\epsilon \in [0, 1/2]$ *; on suppose que les conditions C1, C5 – C7 sont satisfaites. Alors, on a*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^n(\theta) - \widetilde{\boldsymbol{\beta}}(\theta)\right) = \frac{1}{\sqrt{n}} q(\theta)\, \boldsymbol{D}_n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \left(\theta - \mathbb{I}\left(\varepsilon_i - Q(\theta) \leq 0\right)\right) + O_p\left(n^{-1/4}\right).$$

uniformément pour $\theta \in [\epsilon, 1-\epsilon]$*, où*

$$\widetilde{\boldsymbol{\beta}}(\theta) = \left(\beta_1 + Q(\theta), \beta_2, \ldots, \beta_p\right)^\top.$$

La variance asymptotique dépendant de la densité de probabilité des erreurs (*inconnue*), nous avons besoin de "bons" estimateurs de la variance asymptotique (notée S_n) invariants par régression ($S_n(\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\beta}) = S_n(\boldsymbol{Y})$, $\forall\ \boldsymbol{Y}, \boldsymbol{\beta}$) et "invariants" par homothétie ($S_n(c\,\boldsymbol{Y}) = c\, S_n(\boldsymbol{Y})$, $\forall c > 0$). L'estimateur de Welsh [143] vérifie ces deux conditions d'invariance. Nous allons décrire deux types d'estimateurs de $q(\theta)$ dans le cas du modèle de régression (II.1) basés sur le concept de quantile de régression introduit par Koenker et Basset en 1978 [144] qui sont invariants par régression et homothétie. Ceci est en quelque sorte une extension de l'approche de Falk (1986) pour la fonction quantile pour un échantillon $X_1, \ldots, X_n$.

Quand les données sont contaminées il est important que l'interprétation des résultats reste invariante. Dans [144], Koenker et Basset ont donné des propriétés d'invariance de l'estimateur de type quantile de régression.

Théorème II.2. *Considérons* A *une matrice non singulière* $p \times p$*,* $\omega \in \mathbb{R}^p$*, et* $a > 0$*. Alors pour tout* $\theta \in [0, 1]$*,*

1. $\hat{\beta}(\theta, ay, X) = a\hat{\beta}(\theta, y, X)$,
2. $\hat{\beta}(\theta, -ay, X) = -a\hat{\beta}(1-\theta, y, X)$,
3. $\hat{\beta}(\theta, y + X\omega, X) = \hat{\beta}(\theta, y, X) + \omega$,
4. $\hat{\beta}(\theta, y, XA) = A^{-1}\hat{\beta}(\theta, y, X)$.

Preuve. Ce théorème a été démontré dans [144]. □

Nous donnons maintenant la normalité asymptotique de l'estimateur du vecteur $\widehat{\boldsymbol{\beta}}(\theta)$. Nous notons la loi de probabilité conditionnelle de Y_i sachant $X = x_i$ par $P(Y_i < y | X = x_i) = F_{Y_i}(y | X = x_i) = F_i(y)$.

Théorème II.3. *Sous les conditions :*

1. *Pour $i = 1, \ldots, n$, les fonctions $F_i(.)$ sont continues de densités $f_i(.)$ uniformément bornées sur l'intervalle $[0, \infty]$ pour $\xi_i(\theta) = Q_\theta(Y|x_i)$,*
2. *Il existe des matrices définies positives D_0 et D_1 tel que $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n x_i x_i^\top = D_0$ et $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n f_i(\xi_i(\theta)) x_i x_i^\top = D_1$,*
3. *$\max_{i=1,\ldots,n} \|x_i\|/\sqrt{n} \to 0$.*

On a quand $n \to \infty$:

$$\sqrt{n}\left(\hat{\beta}(\theta) - \beta(\theta)\right) \xrightarrow{\mathcal{D}} N_p\left(0, \boldsymbol{\Sigma_\theta}\right), \tag{II.4}$$

avec

$$\boldsymbol{\Sigma_\theta} = \theta(1-\theta) D_1^{-1} D_0 D_1^{-1}. \tag{II.5}$$

Preuve. Ce théorème a été démontré dans [145], page 121. □

Pour une erreur i.i.d, la variance asymptotique s'écrit :

$$\boldsymbol{\Sigma_\theta} = \frac{\theta(1-\theta)}{f^2(0)} D_0^{-1}. \tag{II.6}$$

Dans le cas unidimensionnel pour la première composante du vecteur de régression $\boldsymbol{\beta}$, il s'ensuit que lorsque $n \to \infty$

$$\sqrt{n}\left(\widehat{\beta}_1^n(1/2) - \beta_1\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{q^2(1/2)}{4}\right).$$

La quantité $1/f(Q(\theta))$ est la densité du quantile qui est la dérivée de la fonction quantile $Q(\theta)$. En effet, en dérivant par rapport à θ, l'identité

$$F(F^{-1}(\theta)) = \theta,$$

nous obtenons

$$\frac{d}{d\theta} F(F^{-1}(\theta)) = f(F^{-1}(\theta)) \frac{d}{d\theta} F^{-1}(\theta) = 1,$$

et par conséquent, nous en déduisons que :

$$\frac{d}{d\theta} F^{-1}(\theta) = \frac{1}{f(F^{-1}(\theta))} = q(\theta).$$

Les variances asymptotiques (II.6) et (II.5) dépendant de la densité de probabilité des erreurs (inconnue), nous avons besoin de "bons" estimateurs de la variance asymptotique.

• **L'estimateur de type histogramme** noté $\widehat{T}_n(\theta)$ est donné par

$$\widehat{T}_n(\theta) = \frac{\widehat{\beta}_1^n(\theta+\nu_n) - \widehat{\beta}_1^n(\theta-\nu_n)}{2\nu_n}, \tag{II.7}$$

où $(\nu_n)_n$ est une suite de taille de fenêtre bien choisie. Le comportement asymptotique de cet estimateur est donné par le théorème qui suit.

Théorème II.4 (Dodge et Jurečková, 1995)**.** *Sous les conditions C1, C4 – C7, C8 et C9, nous avons, quand* $n \to \infty$*,*

1. $\sqrt{n\nu_n}\left(\widehat{T}_n(\theta) - q(\theta)\right) = O_p(1)$;
2. $\sqrt{n\nu_n}\left(\widehat{T}_n(\theta) - q(\theta)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{q^2(\theta)}{2}\right)$.

Cela implique bien sûr la consistance de l'estimateur $\widehat{T}_n(\theta)$, c'est-à-dire

$$\widehat{T}_n(\theta) \xrightarrow{P} q(\theta) \text{ quand } n \to \infty. \tag{II.8}$$

• **L'estimateur de type noyau**. Cet estimateur de $q(\theta)$, $0 < \theta < 1$, est défini par

$$\widehat{Z}_n(\theta) = \frac{1}{\nu_n^2}\int_0^1 \widehat{\beta}_1^n(w)\, k\left(\frac{\theta - w}{\nu_n}\right) dw, \tag{II.9}$$

où les conditions *C8'*, *C9* sur la taille de la fenêtre sont satisfaites et k est une fonction (noyau) vérifiant *C10*. On a

Théorème II.5 (Dodge and Jurečková, 1995)**.** *Sous les conditions C1, C4 – C7, C8', C9 et C10, nous avons, quand* $n \to \infty$*,*

1. $\sqrt{n\nu_n}\left(\widehat{Z}_n(\theta) - q(\theta)\right) = O_p(1)$;
2. $\sqrt{n\nu_n}\left(\widehat{Z}_n(\theta) - q(\theta)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, q^2(\theta)\overline{K}\right)$.

Là aussi, cela implique la consistance de $\widehat{Z}_n(\theta)$, c'est-à-dire

$$\widehat{Z}_n(\theta) \xrightarrow{P} q(\theta) \text{ quand } n \to \infty. \tag{II.10}$$

2 Quantile de régression par machine à vecteurs supports

Le Quantile de régression par machine à vecteurs supports est une adaption du modèle de régression SVM (Support Vector Machine) pour lequel nous appliquons un niveau de quantile de Y_i conditionnellement à $X = x_i$ pour $i = 1, \ldots, n$. Rappelons tout d'abord l'expression de la fonction quantile

$$\mathbf{Q}(\theta|x_i) = \omega_\theta^\top \phi(x_i) \quad \text{pour} \quad i = 1, \ldots, n \tag{II.11}$$

avec ω_θ est le quantile de régression d'ordre θ. En régression SVM on cherche à mettre en place une régression linéaire dans un espace transformé défini par $f(x, w) = \langle w, \phi(x) \rangle + b$ avec $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ et $\langle ., . \rangle$ un produit scalaire. En régression SVM, l'objectif est donc de minimiser

$$\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} (\xi_i^+ + \xi_i^-) \tag{II.12}$$

sous les contraintes

$$y_i - f(x_i, w) \leq \delta + \xi_i^-, \quad f(x_i, w) - y_i \leq \delta + \xi_i^+, \quad \xi_i^-, \xi_i + \geq 0, \quad i = 1, \ldots, n,$$

avec ξ_i^-, ξ_i^+ les variables d'ajustement associées respectivement à une sous-estimation et une sur-estimation de la variable réponse pour une entrée x_i, δ détermine la marge d'approximation et C est un paramètre de pénalité associé à une erreur supérieure à la marge δ. En régression quantile SVM, on cherche à estimer la régression au niveau de quantile θ plutôt que d'utiliser une régression par vaste marge. La fonction à minimiser devient alors

$$\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \rho_\theta(y_i - w_i^\top \phi(x_i)), \tag{II.13}$$

où C désigne le degré de pénalisation qui contrôle le degré de tolérance à s'écarter du lissage au niveau du quantile θ. Le quantile de régression de niveau θ peut alors s'écrire pour x^* comme

$$\mathbf{Q}(\theta|x^*) = \sum_{i=1}^{n}(\lambda_i^- - \lambda_i+)K(x_i, x^*) \tag{II.14}$$

et

$$w_\theta = \sum_{i=1}^{n}(\lambda_i^- - \lambda_i^+)\phi(x_i), \tag{II.15}$$

avec λ_i^+, λ_i^- les multiplicateurs de Lagrange et $K(x_i, x_j)$ la fonction noyau.

3 Quantile régression par forêt aléatoire

L'algorithme de régression quantile par forêt aléatoire est une alternative robuste de la régression par forêt aléatoire. La régression par forêt aléatoire est basée sur une approche de construction d'arbres décisionnels [146] par bootstrap [147] et bagging [148]. Cet algorithme de régression fonctionne en plusieurs étapes. Nous notons τ le paramètre qui détermine la façon dont les arbres sont construits, c'est à dire pour chaque noeud les variables et valeurs des séparations et $T(\tau)$ l'arbre correspondant. Soit $\mathcal{B}$ l'espace dans lequel X est défini, $X : \Omega \to \mathcal{B} \subseteq \mathbb{R}^p$, où $p \in \mathbb{N}^+$ est la dimension de X. Chaque feuille $l = 1, \ldots, L$ de l'arbre correspond à un sous espace rectangulaire de $\mathcal{B}$. Nous notons cet espace par $R_l \subseteq \mathcal{B}$.

L'algorithme est le suivant :

— tirage de n échantillons par bootstrap à partir des données,

— pour chaque tirage, construction de m arbres de régression non élagués, et sélection de l'arbre pour lequel l'erreur est minimum,

— prédiction par agrégation des n arbres.

La prédiction d'une nouvelle observation donnée par un arbre $T(\tau)$ est déterminée par la moyenne pondérée des observations de la feuille dans laquelle se situe l'observation (on notera $l(x, \tau)$ la feuille prédite pour x par l'arbre $T(\tau)$). Les poids $w_i(x, \tau)$ correspondent à une

valeur constante positive si $x \in l(x,\tau)$ et 0 sinon. Le poids est alors donné par :

$$w_i(x,\tau) = \frac{\mathbb{I}_{\{X_i \in R_{l(x,\tau)}\}}}{\text{Card}\{j : X_j \in R_{l(x,\tau)}\}} \qquad \text{et} \qquad \sum_{i=1}^{n} w_i(x,\tau) = 1.$$

Pour un arbre, la prédiction est alors donnée par

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x,\tau) Y_i.$$

En forêt aléatoire, on cherche à approcher la moyenne conditionnelle $E(\mathbf{Y}|\mathbf{X} = x)$ par une prédiction obtenue sur les m arbres. Les pondérations s'écrivent alors comme une moyenne des poids pour chaque arbre tel que :

$$w_i(x) = \frac{1}{m} \sum_{t=1}^{m} w_i(x,\tau_t)$$

où τ_t désigne les paramètres de construction du $t$ème arbre ($t = 1, \ldots, m$). Les prédictions par forêt aléatoire sont alors obtenues par

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x) \mathbf{Y}_i.$$

On sait que la distribution conditionnelle de $\mathbf{Y}$ donnée par $\mathbf{X} = x$ est donnée par

$$F_Y(y|\mathbf{X} = x) = P(\mathbf{Y} = y|\mathbf{X} = x) = E\left(\mathbb{I}_{\{\mathbf{Y} \leq y\}}|\mathbf{X} = x\right).$$

En quantile de régression par forêt aléatoire [149] approxime $E\left(\mathbb{I}_{\{\mathbf{Y} \leq y\}}|\mathbf{X} = x\right)$ par la moyenne pondérée des observations de $\mathbb{I}_{\{\mathbf{Y} \leq y\}}$ par

$$\hat{F}_Y(y|\mathbf{X} = x) = \sum_{i=1}^{n} w_i(x) \mathbb{I}_{Y_i} \leq y\}$$

en utilisant les mêmes pondérations que celles utilisées en forêt aléatoire. En quantile de régression par forêt aléatoire, on cherche à estimer la distribution conditionnelle de la fonction

$$F_Y(y|X = x) = P(Y \leq y|X = x).$$

Le quantile d'ordre α, $0 \leq \alpha \leq 1$, est donné par

$$Q_\alpha(x) = \inf\{y : F_Y(y|X = x) \geq \alpha\}. \tag{II.16}$$

Afin d'estimer la fonction quantile Q_α, il suffit de remplacer dans (II.16) $F_Y(y|X = x)$ par $\hat{F}_Y(y|X = x)$. La consistance de cette méthode est montrée dans [149].

Chapitre III

Modélisation robuste de l'écotoxicité de composés chimiques

Si on considère la législation REACH [1], trois types de tests sont requis pour définir une écotoxicité aigüe de substances chimiques. Les tests sont réalisés sur trois espèces de références : poissons, crustacées et algues. Actuellement, les études QSAR sur poissons sont les plus nombreuses [20]. Les études QSAR sur daphnies (crustacées) ont fait l'objet de nombreuses études dont une au niveau du CERMN [20]. Par contre, la situation pour les algues est complètement différente avec très peu d'études QSAR (voir [150] et [151]). La raison vient de l'absence d'un ensemble consistant de données et d'une variabilité importante des mesures des tests sur algues provenant de l'utilisation de méthodes variées et de différentes espèces d'algues utilisées. Ainsi, peu de modèles QSAR pour une narcose non-polaire ont été définies pour les algues : un modèle pour *Selenastrum capricornutum* avec seulement dix produits chimiques, deux modèles pour *Chlorella vulgaris* avec 34 et 91 produits chimiques, et un autre pour les algues vertes avec 51 composés chimiques [42, 152, 153].

La notion de Mode d'Action (Mode Of Action : MOA) apparaît comme fondamentale pour les modèles QSAR en écotoxicologie [154]. On distingue 2 grands types de MOA, un MOA spécifique et un MOA non spécifique, qui se répartissent en 5 classes. Pour le MOA non spécifique [155], on définit 3 classes :

— la narcose non polaire ou "Baseline narcosis",

— la narcose polaire ou "Polar narcosis",

— une réactivité non spécifique ou "Reactive compounds".

Historiquement la narcose non polaire et la narcose polaire se différencient par rapport à la relation linéaire entre le $\log(K_{OW})$ et l'activité, le $\log(K_{OW})$ correspond au logarithme du coefficient de partage d'un composé entre l'octanol et l'eau (l'octanol mimant un milieu membranaire). La narcose est en relation avec une accumulation des composés dans une membrane cellulaire aboutissant à un phénomène type anesthésie (narcotique) réversible. On observe pour la narcose polaire un excès de toxicité qui se traduit par une valeur plus élevée de l'ordonnée à l'origine dans le modèle de régression [155] (page 443). On retrouve la même logique pour la réactivité non spécifique avec une valeur encore plus élevée de l'ordonnée à l'origine. Pour un MOA spécifique on définit une seule classe. Ce mode se traduit par une toxicité supérieure à celle enregistrée pour un MOA non spécifique associé aux narcoses non polaires et polaires. Cette toxicité est le résultat d'une interaction directe avec les mécanismes biochimiques majeures de l'espèce étudiée. Dans ce cas, la modélisation QSAR classique est basée sur des familles chimiques restreintes possédant le même type d'action biologique. La dernière est indéfinie pour le MOA (ni spécifique, ni non spécifique) et est souvent associée à une toxicité inférieure à la toxicité dite de base correspondante à la narcose non polaire. Sur cette classification a été défini un critère appelé le Ratio de Toxicité (TR) permettant de positionner théoriquement un composé dans un mode d'action en connaissant les modèles associés à un MOA non spécifique [156]. Une erreur d'un facteur 10 nous permet de différencier les composés ayant soit un MOA spécifique, soit aucun MOA, des autres composés possédant un MOA spécifique. Un des objectifs est donc de construire un modèle robuste pour les composés ayant un MOA non spécifique.

Notre étude s'est concentrée sur une espèce algue (EC_{50} à 72 heures, *P. Subcapitata*). Les données sont décrites dans l'état de l'art (Chapitre I) et regroupent les informations de 401

composés chimiques associées à 153 descripteurs. Parmi les 153 descripteurs, 3 descripteurs sont apparus comme étant fondamentaux, 2 descripteurs classiques (le coefficient de partage octanol eau ($\log(K_{OW})$), la solubilité moléculaire dans l'eau ($\log(S)$)) pour les études QSAR en écotoxicologie et un descripteur électronique (la polarisabilité (Apol)).

À notre connaissance, les approches robustes basées sur des quantiles n'ont pas été développées pour modéliser les relations quantitatives entre la structure et l'activité de composés chimiques. La thèse et l'article ci-dessous s'articulent principalement autour de ces modèles. Plusieurs noyaux ont été utilisés et nous avons sélectionné le noyau Gaussien en régression quantile et noyau radial Gaussien en régression quantile SVM pour lesquels les résultats sont les meilleurs.

1 Quantile regression model for a diverse set of chemicals : application to acute toxicity for green algae

Ce paragraphe concerne un article publié en 2014 dans le Journal of Molecular Modeling. Cet article a été écrit par Jonathan Villain, Sylvain Lozano, Marie-Pierre Halm-Lemeille, Gilles Durrieu et Ronan Bureau.

Abstract The potential of quantile regression (QR) and quantile support vector machine regression (QSVMR) was analyzed for the definitions of quantitative structure-activity relationship (QSAR) models associated with a diverse set of chemicals toward a particular endpoint. This study focused on a specific sensitive endpoint (acute toxicity to algae) for which even a narcosis QSAR model is not actually clear. An initial dataset including more than 401 ecotoxicological data for one species of algae (*Selenastrum capricornutum*) was defined. This set corresponds to a large sample of chemicals ranging from classical organic chemicals to pesticides. From this original data set, the selection of the different subsets was made in terms of the

notion of toxic ratio (TR), a parameter based on the ratio between predicted and experimental values. The robustness of QR and QSVMR to outliers was clearly observed, thus demonstrating that this approach represents a major interest for QSAR associated with a diverse set of chemicals. We focused particularly on descriptors related to molecular surface properties.

Keywords : Algae species . Ecotoxicology . Molecular surface . Outliers . Quantile regression . Support vector machine

1.1 Introduction

Under REACH legislation [157], quantitative structure-activity relationship (QSAR) models are expected to be used as an alternative to save resources and to accelerate hazard and risk assessments. For algae, one of the three major endpoints in ecotoxicology, even QSAR models [158, 159] associated with a non-specific mode of action (MOA) are not clearly defined. The reason is explained by Aruoja and al. [150] and Netzeva and al. [151]. The issue comes from the lack of a consistent dataset with more than 100 values and the variability of algal test results due to the different methods and algae species used. Therefore, few non-polar narcotic QSAR models were defined for algae : one for *Selenastrum capricornutum* with only ten chemicals, one for *Chlorella vulgaris* with 34 chemicals, and one for green algae with 51 chemicals [152, 153]. For a global model, to our knowledge, only one study has been published (45 chemicals [160]) regarding the prediction of acute toxicity of chemicals to *Selenastrum capricornutum*. From these first QSAR studies, it appears that this endpoint is characterized by a particular sensitivity toward chemicals, and the presence of outliers affects the estimated models [161, 162]. An important characteristic of quantile regression (QR), compared to classical least squares regression, is its robustness to distribution assumptions and to outlying observations [144]. Experimental conditions associated with our study have integrated some measurement errors and systematic biases difficult to control (particularly true as soon as several MOA are related to

our set). The use of QR should make the inference less biased and less sensitive to outliers. A major aspect associated with QSAR is the relationship between MOA and chemical derivatives [163, 164]. This assumption is relied upon when applying read-across analysis to data from REACH. Read-across information considers that the toxicity of a derivative could be estimated from the real toxicological data of a second derivative based on the chemical similarity between the two structures and by assuming that they interact through the same MOA. In aquatic ecotoxicology, four MOAs are classically differentiated. Two are directly related to the relationship between one major descriptor (logKOW) and the biological activities, i.e., baseline and polar narcosis mechanisms for the MOA. The other two correspond to chemical reactions with macromolecules (reactive functions related to native structures or metabolites) and to specific intermolecular interactions with macromolecules (modulation of biological pathways). The notion of toxic ratio (TR) represents one parameter to differentiate a non-specific (narcosis) from a specific MOA [164]. In this study, a large and diverse set of derivatives is considered with the integration of several major sources of ecotoxicological data. One was provided by the Japanese Ministry of Environment [17]. In this case, the biological tests were carried out according to the OECD test guidelines performed under Good Laboratory Practices (GLP). The others correspond to data extracted from the registration files accessible from the ECB website [18], the AQUIRE database [19], and an internal database called MATE [20]. The objective of our study is to define quantile QSAR models for this endpoint, models integrating a large number of chemicals. The sensitivity of the regressions to outliers is analyzed, and QR was used in combination with support vector machine (QSVMR [165]). A comparison of QSVMR with the classical SVM regression (SVMR) is also provided.

1.2 Methods

1.2.1 Training set

The ecotoxicological data were downloaded from the OECD website [17]. The biological tests on a specific algal species named *Pseudokirchneriella subcapitata* (*Selenastrum capricornutum*) followed the OECD-GLP standard and OECD test guidelines. Two types of 72 h-EC_{50} values (OECD TG 201) were recorded in this set corresponding to the growth rate (EC_{50r}) and/or to the area under growth curve (AUG / EC_{50b}). A high correlation between the two types of EC_{50} values ($n = 249, r = 0.967$) was observed. For the most recent data, only the values associated with the growth rate method are displayed. When examining the overall data and particularly the correlation between the two values, we chose to consider the lowest acute toxicity values recorded for the 72 h EC_{50} regardless of the methods. With the cut-off values associated with hydrophobic and hydrophilic properties (*vide infra*), 277 chemical derivatives were part of the training set.

1.2.2 Testing set

A first dataset of biological data relating to algal growth effects of 2782 high-production volume chemicals was taken from the ECB website [18]. Among these 2782 chemicals, 1749 structures can be downloaded, and only 47 have 72-h EC_{50r} value(s) for the same species as the training set : *Selenastrum capricornutum* (the lowest EC_{50r} value is taken in the case of several values displayed). A second data set was obtained from AQUatic Information REtrieval (AQUIRE) [19]. An advanced query was carried out on AQUIRE, and a set of 60 chemical structures with 72-h EC_{50r} value for the *Selenastrum capricornutum* was extracted. A third data set of 94 chemicals was retrieved from our internal database [20]. These three data sets were gathered in a testing set. By considering the cut-off values associated with hydrophobic and hydrophilic properties (*vide infra*) and the suppression of duplicates (comparison with the

training set), 124 derivatives composed the testing set.

1.2.3 Descriptors

The octanol-water partition coefficients ($\log K_{OW}$) were determined by two *in silico* methods leading to a descriptor named logP (open-source software KOWWIN, [166]) and a second one named ALogP, an atom-based method [167]. Molecular solubility, expressed as logS with S in M, was estimated from multiple linear regression models defined by Tetko et al. [25]. Only derivatives with ALogP ≥ 0 and logS values ≥ -6 were retained. The 3D atomic coordinates were generated for each structure and a first energy minimization was performed with Pipeline Pilot [23] using a clean force field [168]. Then, another optimization was carried out with DMol3 [169] by considering PWC [170, 171] for the functional (DFT exchange correlation potential) and medium for the convergence. For the descriptors, special attention was given to molecular surface properties. The first ones correspond to molecular surface areas and their associated descriptors. In this case, total, polar, and solvent accessible surface areas were computed for each molecule using a 2D approximation. The fractional polar surface areas were also determined using the ratio between polar and total surface areas (the same process was applied to solvent accessible surface areas). For topological descriptors associated with molecular shapes, shadow indices and Jurs descriptors were calculated. Shadow indices [26] project the molecular shapes on three mutually perpendicular axes : XY, XZ, and YZ. The associated lengths (shadow Xlength, Ylength, Zlength) correspond to the maximum dimensions of the molecular surface projections. The ratio between the largest and the smallest dimension corresponds to the last descriptor (shadow_nu). The 30 Jurs descriptors [172] combine shape and electronic information. It is impossible to detail all these descriptors, but we can mention a descriptor named Jurs_PPSA_1 corresponding to the sum of the solvent accessible surface areas of all positive atomic charges or Jurs_PNSA_1, calculated in the same way but for negative atomic charges. Ehresmann and al. [27] have described new molecular descriptors based on local properties at

the molecular surface. This surface corresponds to a shrink-wrapped isodensity surface [173], with $10^{-4}.e^{-}.\hat{A}^{-3}$ for the electronic density, generated from semi-empirical molecular orbital calculations (VAMP [174] in this case). Four local properties, the molecular electrostatic potential (MEP), the local ionization energy (IE_L), the local electron affinity (EA_L), and the local polarizability (α_L, polarizability for the name of the descriptor) were calculated at the points on the surface. Two properties, the local hardness (η_L) and the local electronegativity (χ_L), were derived from IE_L and EA_L. Starting from these local properties, 81 descriptors were determined. Other descriptors were associated with steric and electronic properties (Pipeline pilot and Dmol3) like dipole moment descriptors [175, 176], the radius of gyration, the sum of atomic polarizabilities (Apol), principal moments of inertia (PMI), molecular weight, globularity, count descriptors (H bond acceptor and donor), pKa_acide (most acidic site), pKa_basic (most basic site), HOMO, LUMO, band gap energy (LUMO–HOMO), dielectric energy, solvation energy, molecular volume, and cavity volume.

1.2.4 Quantile regression

We consider the linear regression model :

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{III.1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is the vector of observations, $\mathbf{X}$ is the design matrix of dimension $n \times p$ where, for $i = 1, \ldots, n$, $x_i^\top \in \mathbb{R}^p$ is the ith line of the matrix $\mathbf{X}$ with $\top$ denoting the transpose of the vector, $\varepsilon = (\varepsilon_\mathbf{1}, \ldots, \varepsilon_\mathbf{n})^\top$ is a vector of independent errors with an unknown distribution function f and $\beta = (\beta_\mathbf{1}, \ldots, \beta_\mathbf{p})^\top$ denotes the vector of unknown regression parameters to be estimated. The classical least squares linear regression estimator is ineffective if the errors are non-normal. To overcome this problem, in 1978, Koenker and Basset [144] proposed a quantile-based approach for linear regression models. The quantile regression estimator is more robust to non-normal errors and outlier observations. In this case, instead of focusing on the changes

in the mean of $\mathbf{Y}$, the QR approach tests whether there is a change in the θth-quantile of $\mathbf{Y}$ for any given $\theta \in (0,1)$. So, QR gives better characterization of the data, since it enables estimating the impact of a covariate on the entire distribution of the response variable rather than on its conditional mean. The least squares estimators in regression are designed to estimate the mean of the response variable $\mathbf{Y}$ conditional on $\mathbf{X}$ whereas in QR, the estimators are designed to estimate the relation of $\mathbf{X}$ with $\mathbf{Y}$, conditional on quantiles of $\mathbf{Y}$. QR can be viewed as an extension of the classical least squares estimation of conditional mean models for the estimation of models associated with several conditional quantile functions. Therefore, QR estimates the conditional median or other quantiles of the response variable, unlike the ordinary least squares regression, which estimates the conditional mean. We also note that QR is invariant to monotonic transformations, such as logarithmic transformation, and QR algorithms are now available in most statistical packages. The quantile regression estimators :

$$\hat{\beta}(\theta) = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\theta(Y_i - x_i^\top \beta), \tag{III.2}$$

are defined as a solution of the minimization problem where $\rho_\theta(z) = |\rho_\theta(z)| = z(\theta - I(z < 0))$ and $I(P)$ takes the value 1 or 0 depending on whether the condition P is satisfied or not. The QR loss function denoted by the function ρ_θ is an absolute loss function, that is a weighted sum of absolute deviations where the $(1-\theta)$ weight is assigned to the negative deviations and the θ weight is used for the positive absolute deviations. More specifically, it can be shown that this loss function makes it possible to determine the θ-quantile, $\theta \in (0,1)$. A special case of this class of estimator (obtained for $\theta = 1/2$) is the least absolute deviation (LAD) estimator or the median regression, which is obtained by resolution of the minimization problem (III.2). LAD is often chosen as an alternative to least squares estimators. It performs better in the presence of heavy tail distributions. Under the regularity conditions given in [145], the asymptotic normality of the quantile regression estimator $\hat{\beta}(\theta)$ was proven by [144] and [145] under the assumption of

independent and identically distributed (iid) errors, that is, ϵ_i are iid variables in the model. The asymptotic representation was given in [177] for independently, but not necessarily, identically distributed errors. The asymptotic variance of the estimator $\hat{\beta}(\theta)$ can be obtained by direct estimation using the non-parametric estimation of the sparsity function [178, 179]. When the observations are independent but not identically distributed, as often experienced in practical chemical applications, it is possible to extend the iid theory to produce a version of the Huber-Eicker-White sandwich formula for the limiting covariance matrix of $\hat{\beta}(\theta)$. Several estimators have been proposed for this problem, including a rank test as described in [180, 181] and following the work of [182], [183] and bootstrap methods [184, 185]. Statistical tests in quantile regression models need an estimator of the unknown nuisance sparsity function. A Wald test for the null hypothesis can be applied using the consistency of the sparsity function estimator and the asymptotic normality of the quantile regression estimator. The regression rank score [182] also provides an interesting approach to many inference problems while avoiding the density function estimation. The inference on quantile regression can also be considered using Khmaladze's extension [186, 187] of the Doob-Meyer construction. For more information on quantile regression methods, see Briollais and Durrieu publications [178, 188].

1.2.5 Segmented linear regression model

This approach led to the definition of the optimum number of descriptors for the equations. We consider the simple linear regression model with only one change point given by :

$$Y_i = \begin{cases} a_1 + b_1 X_i + \epsilon_i & \text{if } X_i \leq \tau \\ a_2 + b_2 X_i + \epsilon_i & \text{if } X_i > \tau \end{cases} \tag{III.3}$$

where ϵ, a_1, a_2, b_1, b_2 and τ denote respectively the random error term, the unknown intercepts, slopes and change-point in the coefficient of the two linear regression models. The objective is to test, using the likelihood ratio test [189], the "no change in the regression coefficient" null

hypothesis against the "one change in the regression coefficient" alternative hypothesis. For general information on the bilinear model applied to biological systems, see the initial work of Kubinyi [190].

1.2.6 Support Vector Machine Regression (SVMR)

For the linear regression (in feature space) defined by $f(x, w) = \langle w, \phi(x)\rangle + b$ with $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ and $\langle w, \phi(x)\rangle$ is the dot product in the feature space, the objective is to minimize :

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-), \tag{III.4}$$

subject to :

$$y_i - f(x_i, w) \leq \delta + \xi_i^-, \quad f(x_i, w) - y_i \leq \delta + \xi_i^+, \quad \xi_i^-, \xi_i^+ \geq 0, \quad i \in \{1, \ldots, n\},$$

where ξ_i^+ and ξ_i^- are respectively the slack variables associated with an overestimate and an underestimate of the calculated response for the input vector x_i, δ determines the limits of the approximation, and C is a positive constant that controls the penalty associated with deviation larger than δ. The minimization problem can be formulated in its dual quadratic optimization form, which involves maximizing

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\lambda_i^- - \lambda_i^+)(\lambda_j^- - \lambda_j^+)\langle x_i, x_j\rangle - \delta\sum_{i=1}^{n}(\lambda_i^- + \lambda_i^+) + \sum_{i=1}^{n}(\lambda_i^- - \lambda_i^+)y_i$$

under the constraint

$$\sum_{i=1}^{n}(\lambda_i^- - \lambda_i^+) = 0 \quad \text{and} \quad \lambda_i^-, \lambda_i^+ \in [0, C],$$

where λ_i^-, λ_i^+ denote the Lagrange multipliers. Once the dual problem is solved for λ_i^- and λ_i^+, the solution for a given x is obtained by :

$$w = \sum_{i=1}^{n}(\lambda_i^- - \lambda_i^+)x_i$$

and therefore

$$f(x) = \sum_{i=1}^{n}(\lambda_i^- - \lambda_i^+)\langle x_I, x\rangle + b,$$

where λ_i^- and $\lambda_i^+ \in [0, C]$. For non-linear regression, the support vector machine algorithm can be performed by simply transforming the x_i by a non-linear mapping ϕ from the input space to some highdimensional feature space (sometimes even infinite-dimensional). The optimization problem involves finding the flattest function in future space, not in input space. The solution for x^* is obtained by :

$$f(x^*) = \sum_{i=1}^{n}(\lambda_i^- - \lambda_i^+)K(x_i, x^*) + b$$

SVMR performance depends on a correct setting of the hyper-parameters C, δ and the kernel function.

1.2.7 Quantile Support Vector Machine Regression (QSVMR)

The quantile function Y_i conditionally to $X = x_i$ is given for $i = 1, \ldots, n$ by :

$$Q(\theta|x_i) = \omega_\theta^\top \phi(x_i) \quad \text{for} \quad \theta \in (0, 1) \tag{III.5}$$

where ω_θ denotes the θ-quantile regression. QSVMR can be defined by minimizing for $\theta \in (0, 1)$

$$\frac{1}{2}\|\omega_\theta\|^2 + C\sum_{i=1}^{n}\rho_\theta(y_i - \omega_\theta^\top \phi(x_i)), \tag{III.6}$$

where C denotes the degree of penalization controlling the trade-off between the flatness of the quantile function estimate and the amount up to which deviations larger than zero are tolerated. A solution to the minimization problem (III.4) for $\theta \in (0, 1)$ is obtained by optimizing its quadratic dual version. The θ-quantile regression for x^* can be written :

$$\omega_\theta = \sum_{i=1}^{n}(\lambda_i^- - \lambda_i^+)\phi(x_i) \quad \text{for} \quad Q(\theta|x^*) = \sum_{i=1}^{n}(\lambda_i^- - \lambda_i^+)K(x_i, x^*), \tag{III.7}$$

where λ_i^-, λ_i^+ are Lagrange multipliers and $K(x_i, x_j)$ denotes a kernel function.

1.2.8 Parameters and function

We consider the kernel Gaussian radial basis function (RBF) given by the equation :

$$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right), \tag{III.8}$$

where σ corresponds to the bandwidth parameter. The bandwidth parameter is estimated using the procedure developed in [191]. We also used the cross-validation method to determine the value of the bandwidth. We denote in the sequel by $\hat{\sigma}$ the bandwidth estimator of σ. The parameter C determines the trade-off between the model complexity and the degree to which deviations larger than δ are tolerated in the optimization phase. The parameter δ controls (only SVMR) the width of the δ-insensitive zone used to fit the data. Its values affect the number of support vectors. Larger values result in fewer support vectors and more flat regression estimates. To estimate C, we considered the approach of Cherkassy and Ma [192] given by :

$$\hat{C} = \max(|\bar{Y} + 3S|, |\bar{Y} - 3S|), \tag{III.9}$$

where $\bar{Y}$ and S are respectively the empirical estimators of the mean and the standard deviation of the biological activities. This choice of C is more robust than the approach of Mattera and Haykin [193] when the data contains outliers. The choice of δ in SVMR should be proportional to the variability of Y. Cherkassy and Ma [192] propose :

$$\hat{\delta} = 3S\sqrt{\frac{\log(n)}{n}}, \tag{III.10}$$

to determine δ where S corresponds to the empirical estimation of the standard deviation associated with a biological data error of 5 %, 10 % and 15 %. The regression was performed on the training set (277 derivatives) with a threefold cross-validation process to determine the optimum number of variables. The equation selected from SVMR or QSVMR was applied to the testing set (134 derivatives) leading to R^2_{test} values. Afterward, the two sets were reunified

($n = 401$).

1.2.9 Statistical computation

The R statistical environment was used for the overall calculations. Principal component analysis was carried out with the ade4 package [194]. Stepwise regressions were carried out by examining the best descriptors step by step.Within each fold of the cross-validation experiment, an arbitrary division of the dataset into training set (70 %) and testing set (30 %) was fixed. QR, SVMR, and QSVMR were applied using the kernlab package [195]. The coefficients of determination R^2_{cross}, R^2_{test}, R^2_{train} denote respectively the cross-validation, the testing and training determination coefficients respectively. To select the optimum number of variables, we applied a segmented regression model using one unknown change point to be estimated. For all the statistical results, after checking the condition of application of statistical tests (normality, independence, homogeneity, etc.), a probability of $p < 0.05$ was considered significant.

1.3 Results

1.3.1 Comparison of the biological activities (training and testing set)

There is no significant difference in distribution between the training and testing sets when considering the biological activities (p=0.46 with the Kolmogorov-Smirnov test). Therefore, principal component analysis was performed on the joined training and testing sets. The first two principal component axes explain 55 % of the total variability. The following variables have a significant correlation coefficient ($p < 0.05$ with Spearman and Kendall tests) greater than 0.4 in relation to the biological activities : Surface_Area_DMol3, Cavity_Volume_DMol3, Apol, Jurs_TASA, Shadow_XY, ALogP, logP, Molecular_Solubility, Molecular_Weight, Molecular_SurfaceArea, Molecular_SASA, polarizability, and POLint. The representation of the projection of these variables into the correlation circle associated with the first two component axes (see Figure III.1) summarizes the correlations between the variables.

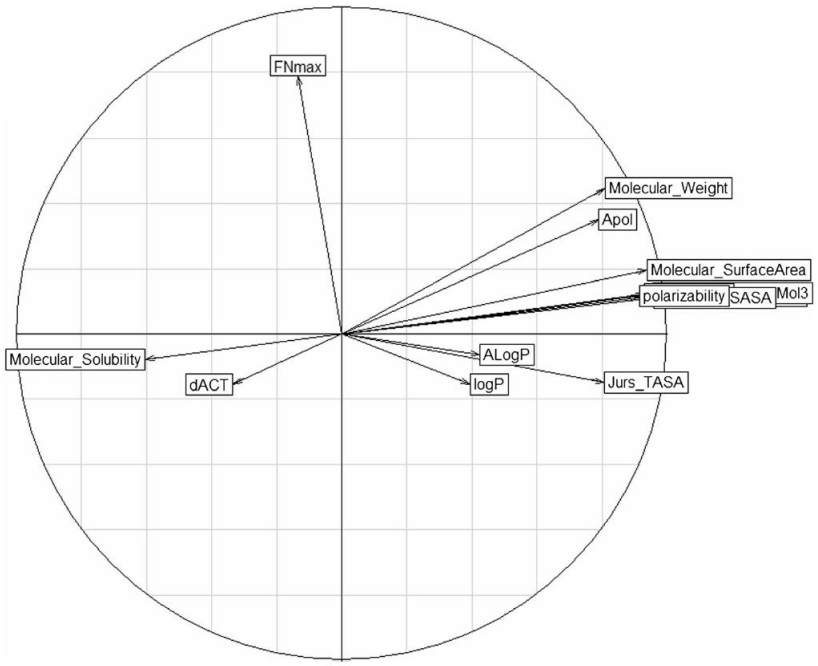


Figure III.1 – Projection of the variables into the plane spawned by the first two principal component axes (dACT for biological activities)

1.3.2 SVMR and QSVMR

SVMR ($\hat{C} = 8.03$,$\hat{\delta} = 0.2$, $\hat{\sigma} = 0.11$) and QSVMR ($\hat{C} = 8.03$, $\hat{\sigma} = 0.11$) were carried out on the training set, and the models were computed on the testing set (see Figure III.2 and Table III.1). Figure III.2 shows the variations of R^2_{cross}, R^2_{test}, R^2_{cross}, and SE_{test} as a function of the number of components (descriptors). SVMR ($\hat{C} = 8.24$, $\hat{\delta} = 0.18$, $\hat{\sigma} = 0.14$) and QSVMR ($\hat{C} = 8.24$, $\hat{\sigma} = 0.14$) were also carried out on the whole dataset (see Table III.1). With QSVMR, when considering the addition of new descriptors in the regression, a stability of statistical values (R^2_{cross}, SE_{cross}) was observed, unlike SVMR for which a sensibility of R^2_{cross} and R^2_{test} values to this number was recorded. For QSVMR (277/124, Table III.1), the three descriptors correspond to logP, solubility, and Apol.

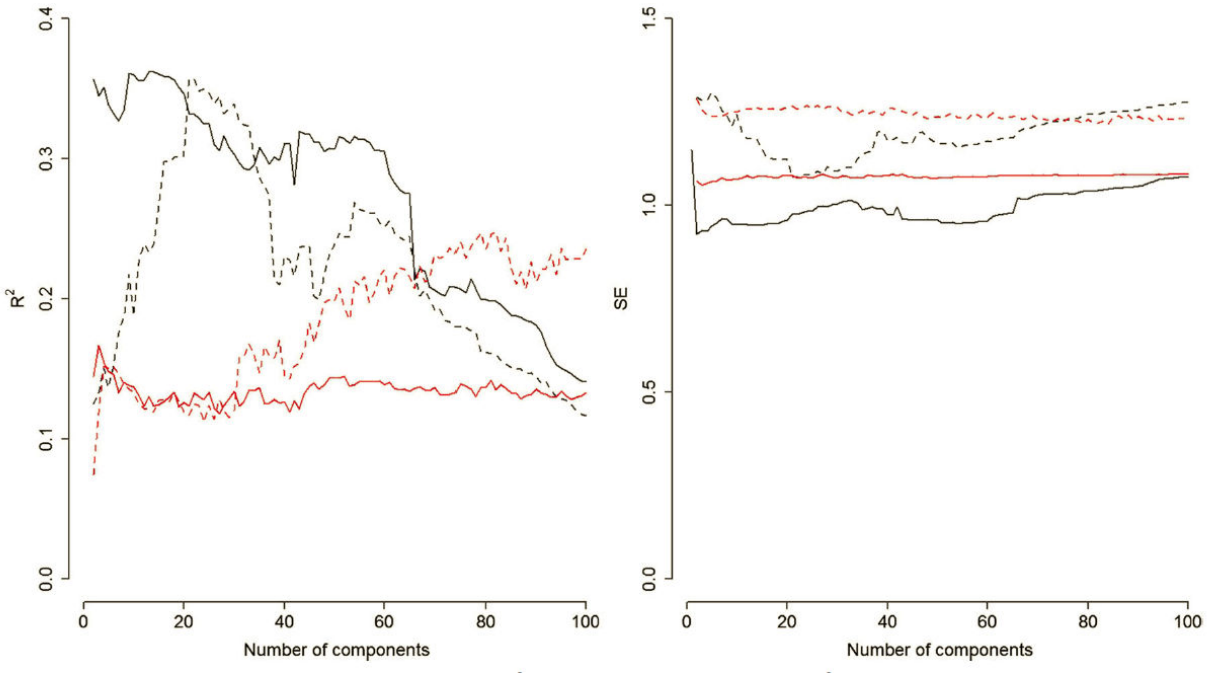


Figure III.2 – Variation in function of the number of components of the R^2_{cross}, SE_{cross} (in solid lines) and R^2_{test}, SE_{test} (in dotted lines). The QSVMR results are in red and the SVMR results are in black

1.3.3 TR from external QSAR and MOA

To analyze the weakness of the initial global model and a potential relation with the association, in the same set of compounds acting with different MOA, the TR was calculated [164]. TR represents the ratio between the predicted (QSAR approach) and the experimental values. The predicted values are determined from a QSAR model associated with a baseline narcosis for the MOA. A cutoff was fixed at ten for differentiating the two MOA (non-specific vs. specific). The model

$$\log\left(\frac{1}{\mathrm{EC}_{50}[M]}\right) = 0.95\log(D_{lipw})(pH7) + 1.16 \qquad \text{(III.11)}$$

of Escher and al. [196] was chosen for this definition, keeping in mind the various remarks concerning the lack of a precise equation for this MOA. The relationship [197] between $\log(D_{lipw})$

Tableau III.1 – Statistical results from SVMR and QSVMR for the training and testing datasets. In the first row, the sample size of the training and the testing sets is 277 and 124 respectively. In the second row, we consider the joined training and testing sets (n=277+124=401)

n	$\hat{\sigma}$	$\hat{C}$	$\hat{\delta}$	θ	Variables	R^2_{train}/SE_{cross}	R^2_{train}/SE_{cross}	R^2_{test}/SE_{test}
SVMR								
(277/124)	0.11	8.03	0.2		22	0.77/0.55	0.33/0.98	0.36/1.08
(401)	0.14	8.24	0.18		19	0.78/0.95	0.35/0.95	
QSVMR								
(277/124)	0.11	8.03		0.5	3	0.61/0.95	0.16/1.03	0.15/1.28
(401)	0.14	8.24		0.5	29	0.60/1.02	0.20/1.09	

and $\log(K_{OW})$ was fixed by considering the model

$$\log(D_{lipw}) = 0.997 \log(K_{OW}) + 0.0851. \tag{III.12}$$

Our values correspond to a 72-h growth rate as opposed to a 24-h growth rate for model (III.11) ; hence, accounting for the expected time dependence of effect, EC_{50} are considered to be three times lower. The integration of the different points and our descriptor logP for $\log(K_{OW})$ led to the final model given by

$$\log\left(\frac{1}{\mathrm{EC}_{50}[M]}\right) = 0.947 \log(P) + 0.77 \tag{III.13}$$

By Eq. (III.13), 294 derivatives out of 401 had a $TR < 10$. Interestingly, 64 % of the derivatives ($n = 254$) were found with a TR between 0.1 and 10. Accordingly, a significant regression model ($R^2_{train} = 0.77$, $R^2_{cross} = 0.76$, $SE_{cross} = 0.48$, $n = 254$) given by

$$\log\left(\frac{1}{\mathrm{EC}_{50}[M]}\right) = 0.76 \log(P) + 2.22 \tag{III.14}$$

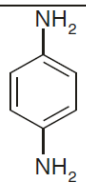
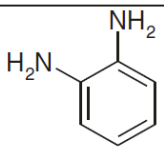
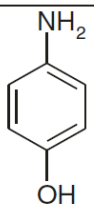
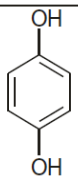
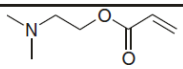
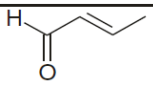
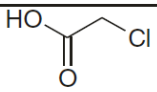
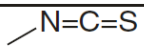
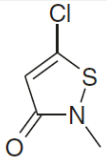
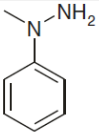
was obtained on this set with logP as a descriptor. This equation represents our first general narcotic equation (associated with baseline and other narcosis like polar narcosis) for the algae endpoint. A SVM classification [112, 198] was carried out to differentiate the chemical charac-

teristics of the two groups ($TR \geq 10$ vs $TR < 10$). The optimum separation was obtained with ALogP, logP, molecular solubility. This classification led to a group called group A ($n = 69$) for which 75 % of chemicals have a $TR > 10$ and a group called group B ($n = 332$) for which 17 % (55 derivatives) of chemicals have a $TR > 10$. In fact, and logically when considering the intercept in regression model (III.13), most of the derivatives in group A have a low value of logP (logP$<$ 1). Our initial validity domain based on ALogP and logS (ALogP$\geq$ 0 and logS values≥ -6) is justified and amplified, starting from this differentiation based on the agreement with the baseline narcosis model given in (III.13). In group A, the highly toxic derivatives ($-\log(\mathrm{EC}_{50}) > 5$) with low values of logP correspond mostly to reactive derivatives (see Table III.2). For phenylenediamine (ortho and para), aminophenol (ortho and para), and hydroquinone, the toxicities are related directly to their redox equilibrium with quinones. Alpha beta unsaturated carbonyl, activated halides, and isothiocyanate can also react with macromolecules. In fact, for reactive chemicals, the interval between predicted and experimental values based on $\log K_{OW}$ was described previously to be higher for hydrophilic derivatives than for hydrophobic derivatives.

1.3.4 TR from external QSAR and QSVMR

A QSVMR analysis ($\theta = 0.5$) was carried out on group B after the suppression of the 13 derivatives with logP$<$ 1. The optimum relationship was obtained for 21 variables ($\hat{C} = 8.27$; $R^2_{cross} = 0.6$; $SE_{cross} = 0.79$, $n = 319$, see Figure III.3). A TR determination was carried out by considering the equation associated with QSVMR as the basis for the predicted values. A total of 18 derivatives were found with a TR$>$ 10. By discarding these 18 derivatives, no real increase in the statistical quality of the equation was observed, thus showing the stability of QSVMR ($\theta = 0.5$) toward potential outliers ($R^2_{cross} = 0.64$, $SE_{cross} = 0.85$, $n = 301$, $\hat{C} = 7.97$, see Figure III.3). However, a decrease in the optimum number of descriptors was recorded with three descriptors (ALogP, molecular solubility, polarizability) instead of the previous 21

Tableau III.2 – Structural analysis of some derivatives (group A) with high toxicities and low logP values

 5.78	 5.12	 6.03	 6.31
 5.85	 5.17	 6.15	 5.71
 6.06	 7.15	 6.28	

descriptors (for $n = 319$). The ecotoxicity of most of these outliers is clearly associated with a specific MOA [199] well recorded in the literature (see Table III.3). Phenylurea, triazinone, and bipyridylium derivatives are inhibitors of photosynthesis (photosystem I for bipyridylium and photosystem II otherwise). Chloracetamides are long-chain fatty acid inhibitors and inhibitors of mitosis and cell division. Diphenylethers are inhibitors of protoporphyrinogen oxidase or potentially uncouplers by considering the phenol function. Quinoline derivatives represent a specific class of antifungic drugs. For reactive chemicals, two unsaturated derivatives, one peroxide and one phenyl nitro derivative appear in this set. As always with reactive chemicals, the highest toxicities of alkyl thiols toward the corresponding nearest alcohol are clearly observed with the formation of free radical species for the explanation [200]. Two polyaromatic derivatives with aniline and nitro functions have a very high toxicity, which must be associated with a specific MOA, but in these cases, no explanation is provided in the literature.

1.3.5 TR/QR/QSVMR

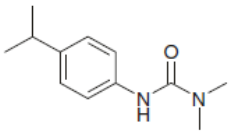
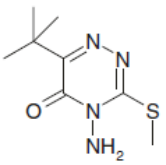
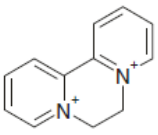
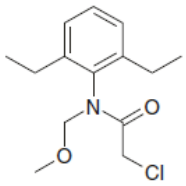
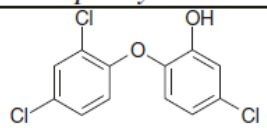
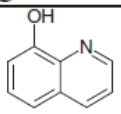
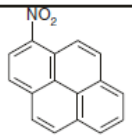
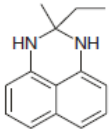
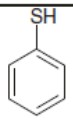
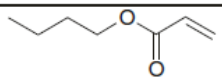
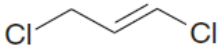
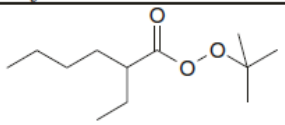
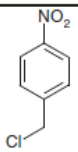
Starting from the initial dataset (401 derivatives), a linear QR was done with logP as descriptor for the definition of the TR. The median quantile regression ($\theta = 0.5$) is given by

$$\log\left(\frac{1}{\mathrm{EC}_{50}[M]}\right) = 0.43\log(\mathrm{P}) + 3.35 \tag{III.15}$$

with $R^2 = 0.48$. Depending on the quantiles, an evolution of the intercepts (2.41($\theta = 0.1$) to 5.25 ($\theta = 0.9$)) and an evolution of the slopes were observed (0.43 ($\theta = 0.1$) to 0.33 ($\theta = 0.9$)). So, logP exerts a change in the conditional distribution of the biological activities. The variance of activity decreases with an increase of logP. From Eq. (III.15), a TR value for each derivative was determined, leading to 336 derivatives with a TR$<$ 10 (as compared to 294 in the previous case). Starting from this set, we obtained the median quantile regression

$$\log\left(\frac{1}{\mathrm{EC}_{50}[M]}\right) = 0.45\log(\mathrm{P}) + 3.08 \tag{III.16}$$

Tableau III.3 – Description of some outliers associated with group B ($n = 310$)

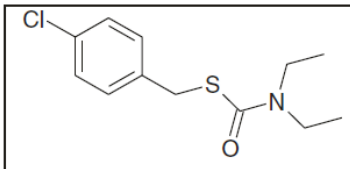
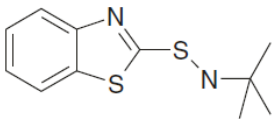
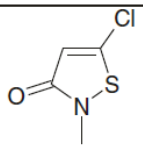
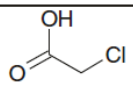
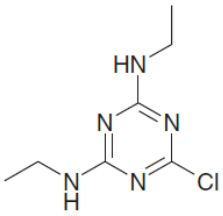
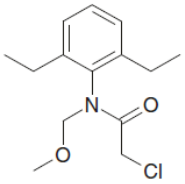
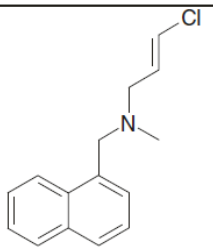
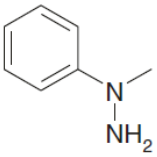
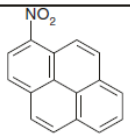
phenyl urea	*Triazinone*	*Bipyridylium*	*Chloroacetamides*
 6.16	 7.15	 6.4	 7.37
Diphenylether	*Quinolines*	*Polyaromatics with aniline and nitro functions*	
 8.14	 5.65	 8.16	 7.02
Thiols			
 7.01	 5.83		
Reactive species for MOA			
 5.15	 5.66	 5.97	 6.66

with a $R^2 = 0.64$. Depending on the quantile, the slope is nearly stable (0.41 for $\theta = 0.1$ and 0.44 for $\theta = 0.9$) with an evolution of the intercept between 2.35 ($\theta = 0.1$) and 3.90 ($\theta = 0.9$). A QSVMR was applied to this set ($\hat{C} = 7.04$, $\theta = 0.5$) leading to interesting results, particularly for the value associated with the standard error of the estimates SE*cross* ($R^2_{cross} = 0.66$, $\text{SE}_{cross} = 0.4$, see Figure III.3). This relationship was obtained with three descriptors (ALogP, molecular solubility, Apol). When examining the 95 % confidence intervals, we obtained less than one logarithmic unit (plus or minus) for the interval associated with predictions. A SVM classification was carried out to understand the differences between the two groups of derivatives (336 with TR< 10 and 65 with TR> 10). No interesting result was obtained from our descriptors. However, for the isodensity surfaces and the properties (molecular shapes, electrostatics, donor and acceptor properties, polarizability) associated with these surfaces, a particular type of fingerprint named the rotationally invariant fingerprint (RIF) can be calculated [28]. Using SVM classification on the overall fingerprint (RIF), an error of classification of 1.75 % was observed on the training and 14.95 % on the cross-validation processes. After the cross-validation process, the dataset was separated into two groups with 368 derivatives and 33 derivatives (TR> 10) respectively. QSVMR was carried out ($\hat{C} = 7.18$, $\theta = 0.5$) on the set of 368 derivatives, leading to correct statistical results with 67 descriptors ($R^2_{cross} = 0.67$, $\text{SE}_{cross} = 0.41$, see Figure III.3 and III.4). Concerning the 33 derivatives associated with outliers, the most toxic derivatives are displayed in Table III.4.

1.3.6 Discussion

The definition of a global QSAR model for a dataset of chemicals is a worthwhile endeavor because, for most chemical derivatives, the classification into a specific set is not obvious. The interval of prediction of this model must cover the experimental value and must be a basis for the estimation of risk. For acute toxicities to algal species, even a narcosis model associated with a large set of derivatives is needed. A previous study related to hydrocarbon derivatives, and

Tableau III.4 – Description of the most toxic chemicals among the set of 33 derivatives. The number of derivatives associated with a typical scaffold is indicated in brackets

 6.79 (2)	 6.62 (4)	 6.06 (1)	 6.15 (1)
 6.30 (7)	 7.37 (3)	 6.08 (1)	 7.14 (1)
 8.16 (1)			

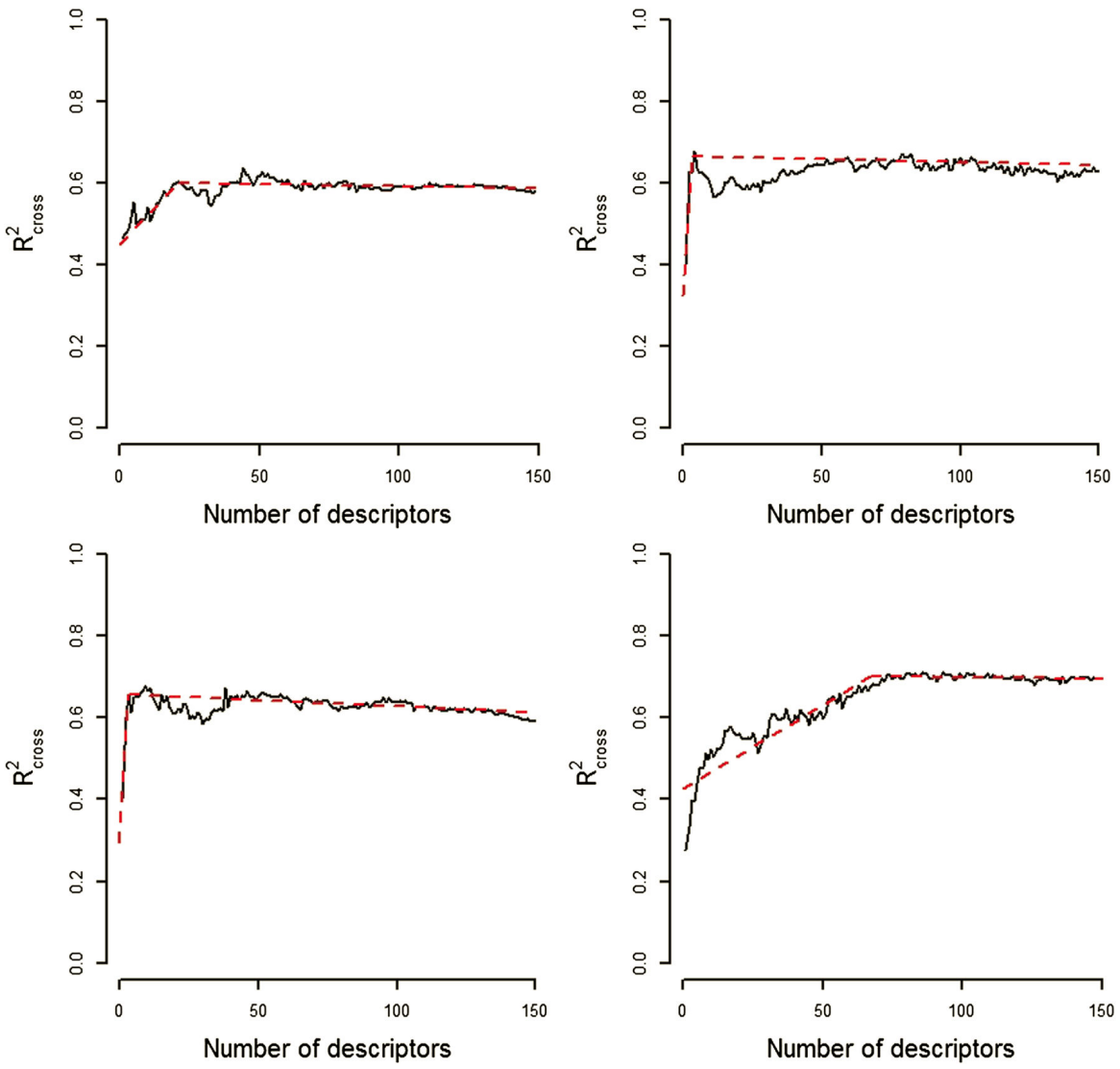


Figure III.3 – Variation of R^2_{cross} as function of the number of components. QSVMR with $n = 319$ (up left), $n = 301$ (up right), $n = 336$ (down left), $n = 368$ (down right)

particularly industrial chemicals such as petroleum products, has demonstrated that a narcosis target lipid model could be built for algae [162]. A final R^2 value of 0.85 was obtained with a standard deviation of 0.34 for the residuals after exclusion of four outliers. For our dataset, the initial correlation (equal to 0.4) with $\log K_{OW}$ was clearly observed.With the notion of TR (external equation), 319 derivatives out of 401 were found with a TR< 10 and 254 derivatives

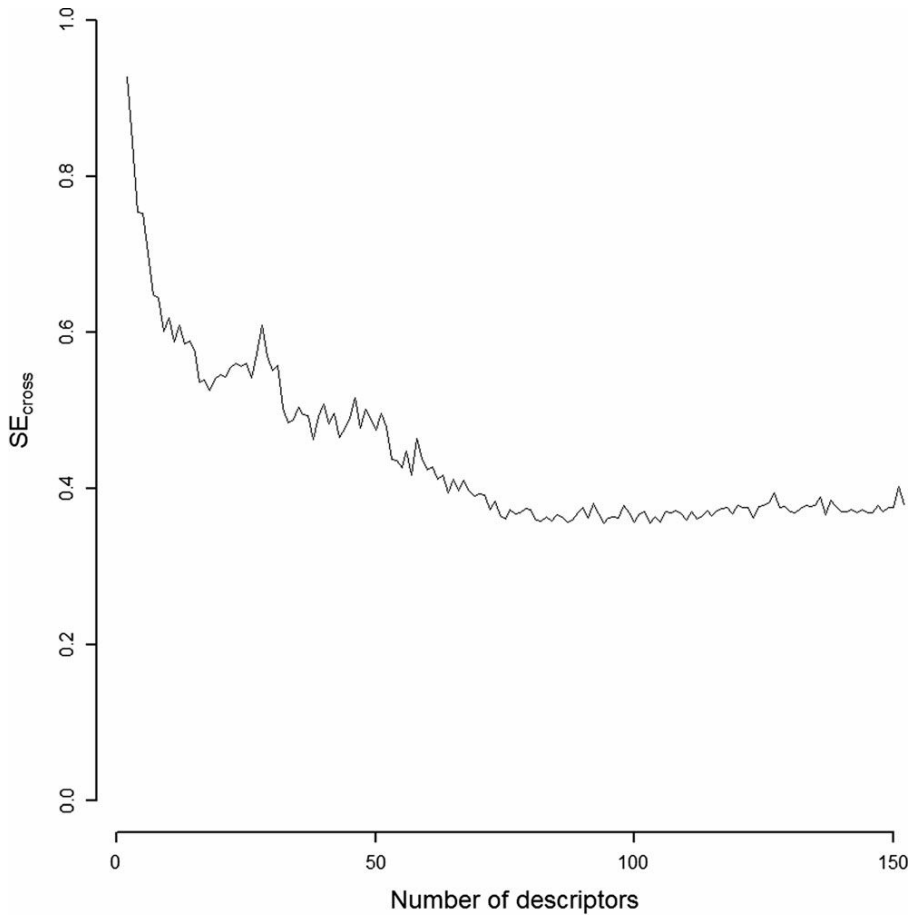


Figure III.4 – Variation of SE_{cross} as a function of the number of components (QSVMR, $n = 368$)

with a TR between 0.1 and 10. The slope estimate in model (III.14) was found to be slightly lower than the slopes observed from the previous narcotic equations with an increase in the intercept. As always with $\log K_{OW}$, the QR regression on the overall set ($n = 401$) then on the subset of 336 derivatives (84 %) led to a higher value of the intercept, namely a value around 3 with an interval between 2.35 and 3.9 for different quantiles. These last equations fit the data associated with some classes of derivatives appropriately, such as aniline, which follows a polar narcosis mechanism for the toxic action (aniline ($-\log EC_{50} = -3.35$, $\log P = 0.9$) ; p-methyl aniline ($-\log EC_{50} = -4.03$, $\log P = 1.39$)). With QSVMR, for the initial training set ($n = 277$), a low predictive quality of the model was obtained but with three descriptors corresponding

to logP, molecular solubility, and Apol. The same type (logP or ALogP, molecular solubility, Apol or polarizability) of optimum descriptors was observed for the set associated with $n = 310$ and with $n = 336$. A narcotic MOA is a function of the relationship between the activities and the water-octanol partitioning ($\log K_{OW}$). However, octanol is not the optimal middle (as an analog of biological membrane), and the partitioning of the chemical derivatives is very sensitive to the polarity-polarizability factor of the structures [61]. Therefore, it is expected that the polarizabilities of the derivatives are one of the main descriptors associated with $\log K_{OW}$ values. Molecular solubility completes the information by describing the thermodynamic equilibrium between solute and solvent. Thus, for the last two sets ($n = 310$ and $n = 336$), the optimum relationships were obtained with these three descriptors and with a stability of the predictive quality of the regressions, regardless of the number of descriptors (see Fig. 3). With QSVMR and in relation with the three descriptors, we always observed two types of evolution in terms of the initial approach. With TR associated with an external equation, derivatives with high toxicities combined with low logP values were suppressed in a first step. Thus, the resulting set (group B) still included hydrophobic reactive derivatives and series of pesticides. QSVMR on this set led to a relationship with 21 descriptors (optimum value of R^2_{cross}) and a high standard error of estimates (SE_{cross}). The suppression of 18 derivatives from this set ($n = 301$) led to a relationship with no difference, excepting the number of descriptors with an optimum for the same three descriptors. With TR defined from QR, the relationships between logP and the ecotoxicities were analyzed on the basis of the QR slopes and intercepts. With 401 derivatives, the variance of the biological activities in function of logP was observed, but after the selection of the subset of 336 chemicals, no real modification of the variance was observed. With this new set, we obtained a model based on the same three descriptors (QSVMR) with the standard errors of the estimates decreasing by a factor of two, as compared to the previous models. The difference between the two sets (chemicals with TR$>$ 10 and others) was

understood with a molecular fingerprint (RIF) integrating the molecular shape and the properties on the molecular surfaces. After this classification, a set integrating 91 % of the initial chemicals gave a significant QSAR model with 67 descriptors. To analyze this last situation, different observations could be made : a) 83 % of the derivatives should follow a narcosis mechanism for the MOA with a QSVMR correct for three descriptors ($n = 336$) ; b) the selection (SVM classification with RIF) led to elimination of more than 50 % of the most toxic derivatives ($-\log EC_{50} > 6.5$) ; c) the remaining derivatives ($n = 368$) share a few main properties associated with the molecular surface (RIF descriptors) ; and d) our descriptors are strongly related to the analysis of the molecular surfaces. Therefore, the QSVMR modeling approach improves the results with an optimum of 67 descriptors. The three descriptors were again integrated to the first descriptors following those initially correlated with biological activities such as molecular surface area, molecular_SASA, Molecular_Weight, Cavity_Volume_DMol3, shadows descriptors (shadow_XY), Jurs descriptors (Jurs_TASA), followed by other descriptors (PMI_Y, PMI_Z, PMI_Mag). The descriptors associated with Parasurf are also selected in the first ones with a series of properties integrated over the surface : polarizability (POLint), local electro-negativity (ENEGint), local ionization energy (IELint), local hardness (Hardint), local electron affinity (EALint), and molecular electrostatic potential (MEPint). A differentiation between some classes of compounds could be found rapidly with these last descriptors such as bipyridilium (MEPint, the highest value), phthalates and phosphates (IELint, high values), and alkyl halides (IELint, the lowest values with one or two carbons).

1.3.7 Conclusion

These analyses provide evidence for a robust modeling approach based on QR and QSVMR to define a global model. QR, based on a fundamental descriptor in ecotoxicology ($\log K_{OW}$), allowed selecting directly a subset of derivatives for which a correct predictive quality was found with QSVMR and three descriptors. Stability of the model, due to its robustness against

outliers, was observed with QSVMR, particularly from the R^2_{cross} values. Based on this result, 83 % of our chemicals should have a narcosis mechanism for the MOA. When examining a subset of derivatives, differentiated by the distances from the rotational invariant fingerprint, a correct relationship was obtained for 91 % of our initial dataset by integrating in this case a large number of descriptors related to the molecular surface properties.

Acknowledgments M. Jonathan Villain was supported by a grant from the Région de Bretagne (Region of Brittany) and the French Ministry of Education. The authors thank Laurent Briollais for the improvement of the narrative style of the manuscript and the Agence Nationale de la Recherche (French National Research Agency) (ANR, ANR-07-CP2D-09-02 and Pharm@ecotox) for financial support.

Chapitre IV

Régression quantile et mode d'action : application aux médicaments

Dans le cadre du programme ANR pharm@ecotox, l'objectif est de cerner l'impact des médicaments sur l'environnement et notamment sur les espèces de référence. Les médicaments étant conçus pour agir directement sur des récepteurs biologiques, la probabilité d'un mode d'action spécifique est plus importante. L'estimation de la présence des médicaments dans l'environnement (PEC : Predict Environnemental Concentration) a été déterminée en partant des données de consommation pharmaceutique en Basse-Normandie (officines). Ceci a abouti à la sélection de 36 médicaments préoccupants. Les valeurs d'écotoxicité (EC_{50} à 72 heures) sur les algues *P. Subcapitata* ont été déterminées en suivant les lignes directrices de l'OECD (Organisation for Economic Co-operation and Development).

Dans le chapitre 3, nous avons pour la première fois établi des modèles robustes associés à un mode d'action non spécifique. Dans cette nouvelle étude sur les médicaments, un faible écart entre les valeurs prédites et réelles doit nous permettre de préciser leur mode d'action et dans le cas d'un mode d'action non spécifique d'étudier la performance et la robustesse des différents modèles.

Un des problèmes concerne la notion d'espace chimique et la similarité entre les médicaments et les 401 dérivés chimiques de référence. Nous avons proposé une méthode statistique robuste

permettant de décrire cette similarité à travers la notion de détection de nouveauté (novelty detection). Cette approche statistique a permis de révéler que sur les 36 médicaments, 12 sont dans le domaine de validité. Parmi les 24 composés se situant hors du domaine de validité, 3 médicaments ont un mode d'action spécifique (2 antibiotiques et 1 antifongique). Les médicaments étant dans la classe 1 (classe associée aux médicaments : classe BBDS) sont très bien prédits en considérant nos modèles robustes QSAR et les dérivés de la classe 4 (classe associée aux médicaments : classe BBDS) n'ont aucun mode d'action au niveau de la toxicité et ne présentent au niveau expérimental aucune toxicité. Les résultats obtenus ont permis de montrer la capacité de nos modèles à prédire la toxicité de médicaments.

1 Acute toxicities of pharmaceuticals toward green algae. mode of action, biopharmaceutical drug disposition classification system and quantile regression models

Ce paragraphe concerne un article publié en 2016 dans le Journal of Ecotoxicology and Environmental Safety. Cet article a été écrit par Jonathan Villain, Laetitia Minguez, Marie-Pierre Halm-Lemeille, Gilles Durrieu et Ronan Bureau.

Abstract The acute toxicities of 36 pharmaceuticals towards green algae were estimated from a set of quantile regression models representing the first global quantitative structure–activity relationships.The selection of these pharmaceuticals was based on their predicted environmental concentrations. An agreement between the estimated values and the observed acute toxicity values was found for several families of pharmaceuticals, in particular, for antidepressants. A recent classification (BDDCS) of drugs based on ADME properties (Absorption, Distribution, Metabolism and Excretion) was clearly correlated with the acute ecotoxicities towards algae. Over-estimation of toxicity from our QSAR models was observed for classes 2, 3 and 4 whereas

our model results were in agreement for the class 1 pharmaceuticals. Clarithromycin, a class 3 antibiotic characterized by weak metabolism and high solubility, was the most toxic to algae (molecular stability and presence in surface water).

Keywords Pharmaceuticals, Acute toxicity for algae, Quantile regression, Mode of action, Novelty detection, , Machine learning.

1.1 Introduction

Among ecosystems highly impacted by human activities, freshwater systems are exposed to a "cocktail" of contaminants. The toxicity of most of these contaminants is still poorly known, particularly for emerging contaminants such as pharmaceutical compounds [201, 202]. Nowadays, the occurrence of pharmaceuticals in aquatic environments is a well-established issue and has become a matter of both scientific and public concern [203]. Indeed, contrary to some conventional pollutants (e.g. pesticides, hydrocarbons, metals), pharmaceutical residues are continuously discharged into superficial waters involving for aquatic organisms an exposure during their entire life cycle [201]. Moreover, these pharmaceutical compounds are designed to act on a known biological target with a particular mode of action (MOA). They are designed to highly interact with biological systems and to resist to inactivation before exerting their intended therapeutic effect. Nevertheless, for algae this notion of specific interaction is not well known. Thus, in this context of multi-contamination, it is very important to be able to quickly estimate toxicity levels of pharmaceutical compounds. QSAR studies (Quantitative Structure Activity relationship) should lead to an estimation of these ecotoxicological characteristics.

To define the most concern drugs, a first step was to characterize the exposure by predicting the aquatic concentrations (PEC). The determination of PEC was obtained from the following

formulae described in the publication of Liebig and al. (2006) [204].

$$PEC = \frac{A \times (100 - R)}{365 \times P \times V \times D \times 100}$$

where the value A corresponds to the real prescription amount by year in Basse-Normandie (France) ; R for the removal rate ($R = 0$ in our case, no biodegradation and absorption whatever the drug) ; P for number of inhabitants of the geographic area considered ($1,450,000$ in Basse Normandie) ; V for the volume V of waste water per capita and day (200 L see [204]) and D for a dilution factor 10 [204].

For instance, the highest value was associated to acetaminophen with an amount of 4.0545e+4 kg for the prescription leading to a PEC value of $38,304$ ng/L. If PEC values exceed the threshold of 10 ng/L, environmental fate and effects are analyzed (Directive 2001/83/ EC amended by Directive 2004/27/EC). Environmental concentrations in freshwaters ($\mathrm{PEC_{FW}}$) were extrapolated for 513 pharmaceuticals using the worst-case assumption [205] with a pipeline pilot workflow [206]. $\mathrm{PEC_{FW}}$ values higher than the threshold of 10 ng/L were selected. Starting from this set, different therapeutic classes of pharmaceuticals were pointed out and particularly antidepressants (selective serotonin-reuptake inhibitors, serotonin-norepinephrine reuptake inhibitor, tricycli antidepressant), neuroleptic (serotonin antagonist), anti-inflammatory (inhibitors of cyclooxygenase), antitussive (NMDA receptor antagonist), anti-histaminics ($\mathrm{H_1}$ receptor antagonists), lipid lowering substances (activation of peroxisome proliferator-activated receptor-alpha), beta-blockers (beta1-selective adrenergic antagonists), antibiotics (binding to the bacterial 50S ribosomal subunit), antifungic (inhibitors of $14-\alpha$ demethylase), antihypertensor (angiotensin II antagonists), anticholesterol (inhibitors of HMG-CoA reductase). Finally, this analysis led to a set of 36 pharmaceuticals assessed for their toxicity to the green algae *Pseudokirchneriella subcapitata*.

IV.1 Acute toxicities of pharmaceuticals toward green algae. mode of action, biopharmaceutical drug disposition classification system and quantile regression models

In our recent publication [207], new robust QSAR quantile regression models were defined for an estimation of the acute toxicities to algae. The chemicals corresponded mainly to conventional organic chemicals but also to some pesticides. In these studies, the potential of Quantile Support Vector Machine Regression (QSVMR) and Quantile Regression (QR) in the prediction of acute toxicities were pointed out. The main motivation of quantile regression models is the robustness of these models to outliers and to non Gaussian distribution [145, 178, 188]. The selection of different subsets was based on the notion of Toxic Ratio (TR). TR represents the ratio between the predicted (QSAR approach) and the experimental values [164]. These approaches increased the relevance of the QSAR models and identified particularly three descriptors. These descriptors were the water-octanol partition coefficient (log K_{ow}), the molecular solubility in water and the polarizability of chemicals. With log P as descriptor (one of the two descriptors associated to log K_{ow} in our initial study), a Linear Regression (LR, $n = 258$) and a Quantile Regression (QR, $n = 336$) led to a first well-fitting significant regression model ($p < 0.05$). The values of this descriptor (log P) were obtained from KowWin, a software which estimates log K_{ow} values [166] or considers, from an internal database, experimental values. On the same set, but with a Quantile Support Vector Machine Regression (QSVMR) and three descriptors log P, molecular solubility [25], and Apol (the sum of Atomic polarizabilities [175, 176] for a descriptor associated to polarizability), a crossvalidated Standard Error (SE_{cross}) value equal to 0.40 was obtained. After a Support Vector Machine (SVM) classification based on a particular chemical fingerprint Rotational Invariant Fingerprint (RIF [28]), a QSVMR model (n=368) was finally defined with a SE_{cross} of 0.41.

The objectives of the present study are (1) to analyze and to compare the performance of the different models in predicting the toxicities of these 36 pharmaceuticals to algae, (2) to characterize some potential MOA of these pharmaceuticals in function of the difference between

observed and predicted values (based on the notion of TR), and (3) to analyze the relationship between known properties of these drugs (pharmacodynamic and pharmacokinetic properties through the Biopharmaceutical Drug Disposition and Classification System (BDDCS)) and their observed acute toxicities for algae.

1.2 Materials and methods.

1.2.1 Test compounds

This study focused on 36 pharmaceutical compounds (Table IV.1) for which PEC-values were estimated above the threshold of 10 ng/L [205]. All were purchased from Kemprotec Limited® (Maltby, Middlsbrough, U.K.) at analytical grade (purity $\geq$ 99%). For water non-soluble compounds, dimethyl sulfoxide was used as carrier solvent at 0.1% (final concentration).

1.2.2 Algal growth inhibition assay.

Acute toxicity tests on the freshwater algae *Pseudokirchneriella subcapitata* were conducted following the NF EN ISO 8692 guideline (2012). Algae *P. subcapitata* AC152 were obtained from Algobank (Caen, France). All algal growth inhibition tests were conducted at 20 ± 1 °C with continuous shaking at 100 rpm and continuous white light (4000 lx). The toxicity tests were performed in 96-well cell culture plates. Each substance, the medium and the algal inoculums were mixed to obtain an initial algal concentration of 10^4 cells/mL in 0.21 mL of bioassay volume. At least three replicates were used per concentration. Cell density was measured after 72 h of exposure. The results were quantified as average growth rates calculated from cell numbers based on measurements of chlorophyll fluorescence (680 nm, TECAN Infinite® M200 microplate reader). The percentages of inhibition of average specific growth relative to controls were calculated for each concentration. All the controls met the acceptability criteria.

Name	log P	ALogP	Molecular solubility	Apol
Acebutolol	1.71	1.61	-3.86	12,647
Acetaminophen	1.16	0.95	-1.74	5937
Amitriptyline	4.92	4.77	-5.95	12,619
Bezafibrate	4.25	3.87	-5.88	15,022
Bisoprolol	1.87	2.03	-4.02	11,995
Carbamazepin	2.45	2.68	-3.75	10,940
Cetirizine	1.7	0.45	-4.61	15,887
Citalopram	3.74	3.72	-5.54	12,130
Clarithromycin	3.16	2.2	-5.38	24,357
Clindamycin	2.16	1.28	-3.46	14,974
Clomipramine	5.19	5.05	-5.32	13,716
Clozapine	3.23	3.42	-3.88	14,016
Dextromethorphan	3.97	3.67	-4.53	10,538
Diclofenac	4.51	4.37	-5.19	13,148
Duloxetine	4.68	3.85	-6.43	13,623
Econazole	5.61	4.98	-7.51	16,933
Fenofibrate	5.19	5.11	-6.26	15,102
Fluoxetine	3.82	4.03	-5.92	11,443
Fluvoxamine	3.09	2.55	-5.11	10,479
Gemfibrozil	4.77	4.17	-4.37	9787
Hydroxyzine	2.36	3.47	-4.51	15,432
Irbesartan	5.31	4.49	-6.84	17,243
Ketoconazole	4.35	3.61	-5.99	21,276
Ketoprofene	3.12	3.36	-4.39	11,212
Metoprolol	1.88	1.76	-3.43	10,110
Mianserine	4.24	3.70	-3.38	11,571
Miconazole	6.25	5.64	-8.27	18,424
Milnacipran	2.03	1.29	-3.02	9823
Naproxen	3.18	2.85	-4.06	9762
Paroxetine	4.74	3.23	-4.57	12,749
Pravastatine	2.18	2.16	-4.29	14,926
Propranolol	3.48	3.03	-4.16	10,767
Sertraline	5.29	5.00	-6.77	13,748
Valsartan	3.65	4.10	-6.31	17,290
Venlafaxine	3.28	3.02	-3.87	10,651
Verapamil	3.79	5.53	-7.51	16,834

Tableau IV.1 – Descriptors used in Linear Regression model (LR), Quantile Regression model (QR) and Quantile Support Vector Machine Regression (QSVMR$_A$).

1.2.3 QSAR and Descriptors.

The set of descriptors and the QSAR models were described in our previous publication [207]. A cutoff (selection or not) was defined in function of two chemical properties (AlogP ≥ 0 and log S values ≥ -6). All the 36 pharmaceuticals follow this cutoff.

The Linear Regression model (LR) is given by :

$$\log \frac{1}{\mathrm{EC}_{50}[M]} = 0.76 \log \mathrm{P} + 2.22.$$

The Quantile Regression model (QR) is given by :

$$\log \frac{1}{\mathrm{EC}_{50}[M]} = 0.45 \log \mathrm{P} + 3.08.$$

For Support Vector Machine (SVM) regression, we proposed in [207] a robust Quantile Support Vector Machine Regression (QSVMR_A, $R^2_{cross} = 0.66$, $SE_{cross} = 0.4$, $n = 336$) based on three descriptors : ALogP (octanol-water partition coefficients), molecular solubility (molecular solubility in water), Apol (molecular polarizability). In linear regression model, the classical least squares estimator is ineffective if the errors are non-normal distributed and in presence of outliers. To overcome this problem, we proposed in [207] a quantile-based approach for linear regression models. In this case, instead of focusing on the changes in the mean of the response variable (biological activity), the quantile approach tests whether there is a change in the θth quantile for any given $\theta \in (0, 1)$. So, quantile model gives better characterization of the data, since it enables estimating the impact of a covariate on the entire distribution of the response variable rather than on its conditional mean. The least squares estimators in regression are designed to estimate the mean of the response variable conditional on the covariates whereas in quantile regression, the estimators are designed to estimate the relation between the descriptors and the response variable, conditional on quantiles of the response variable. For instance for

$\theta = 0.5$, quantile regression estimates the conditional median of the activity variable know as more robust than the mean, unlike the ordinary least squares regression, which estimates the conditional mean. The quantile regression estimators

$$\hat{\beta} = \underset{\beta \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\theta(Y_i - x_i^\top \beta)$$

are defined as a solution of the minimization problem where the quantile regression loss function $\rho_\theta(x) = |\rho_\theta(x)| = x(\theta - I(x < 0))$ is given in Figure IV.1 and $I(P)$ takes the value 1 or 0 depending on whether the condition P is satisfied or not.

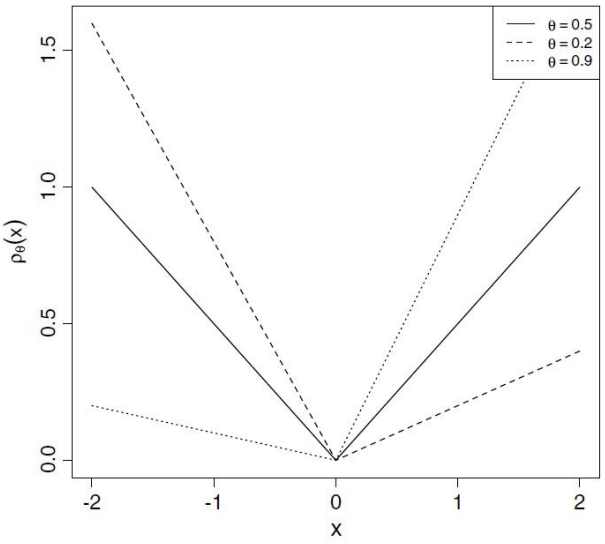


Figure IV.1 – Representation of the function $\rho_\theta(x)$.

The quantile regression loss function $\rho_\theta(x)$ is an absolute loss function, that is a weighted sum of absolute deviations where the $(1 - \theta)$ weight is assigned to the negative deviations and the θ weight is used for the positive absolute deviations (Figure IV.1). More specifically, it can be shown that this loss function makes it possible to determine the θ quantile, $\theta \in (0, 1)$. For $\theta = 0.5$, the regression median is often chosen as an alternative to least squares estimators with better performances in the presence of heavy tail distributions.

After a SVM classification based on RIF descriptors, a second QSVMR (QSVMR$_B$) was carried out with 67 descriptors ($R^2_{cross} = 0.67$, $SE_{cross} = 0.41$, $n = 368$). The complete listing of these 67 descriptors are described in [207].

Table IV.1 displays the values for the descriptors associated to LR, QR and QSVMR$_A$.

1.2.4 Novelty detection

Novelty detection has many practical real-life applications in different domains, and it is of crucial importance in applications that involve large datasets acquired from criticals systems. In the literature, novelty detection is also often called anomaly detection and outlier detection [208]. Barnett and Lewis in 1994 [209] define an outlier as a data point that appears to be inconsistent with the training data. The notion of outlier is also used to describe a small fraction of normal data which lies far away from the majority of normal data in the feature space. The idea of the one-class SVM approach [210] is to define the novelty boundary in the feature space corresponding to a kernel, by separating the transformed training data from the origin in the feature space, with maximum margin. This approach requires to fix a priori the percentage of data allowed to fall outside the description of the training class (here we chose $\nu = 0.2$).

1.2.5 Statistical computation

The R statistical environment version 3.2.0 was used for the overall calculations with the package Kernlab. In the following, for all SVM methods, we consider the Radial Basis Function (RBF) given in (IV.1).

$$K(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right). \tag{IV.1}$$

The hyper-parameter ν was determined using automatic sigma estimation for RBF kernel [191]. This kernel is one of the most used especially in machine learning.

For the novelty detection, the parameter ν is equal to 0.2. Furthermore, for all the statistical results, after checking the condition of application of statistical tests (normality, independence, homogeneity, etc), a probability of $p < 0.05$ was considered significant.

1.3 Results

Table IV.2 and Figure IV.2 display the different estimations based on the four QSAR models for the pharmaceuticals. The best models correspond to QSVMR$_B$ and QR (see Table IV.3). Information concerning the classes associated to the drugs (BDDCS) is also described in Table IV.2.

Name	Therapeutic class	BDDCS	EC_{50} (mg/L)	-log(EC_{50} in mol/L) with 100mg/L for > 100mg/L	LR	QR	$QSVMR_A$	$QSVMR_B$
Acetaminophen	NSAID	1	> 100	3.13	3.10	3.60	3.45	3.23
Carbamazepin	Antiepileptic drugs	2	> 100	3.37	4.08	4.18	4.92	4.49
Clomipramine	Antidepressants	1	0.46	5.83	6.16	5.41	5.73	5.74
Clozapine	Neuroleptic	2	3.13	5.01	4.67	4.53	5.09	5.06
Dextromethorphane	Antitussive	1	2.49	5.03	5.23	4.86	5.57	4.95
Diclofenac	NSAID	1	21.3	4.14	5.64	5.11	5.58	4.75
Gemfibrozil	Lipid lowering substances	2	7	4.55	5.84	5.22	5.43	5.31
Ketoprofene	NSAID	2	49.3	3.71	4.59	4.48	5.32	4.58
Metoprolol	Beta-blockers	1	74.3	3.55	3.64	3.92	4.33	4.74
Mianserine	Antidepressants	1	2.12	5.09	5.44	4.99	4.47	4.99
Naproxene	NSAID	2	44.4	3.71	4.63	4.51	5.02	4.42
Venlafaxine	Antidepressants	1	47.58	3.76	4.71	4.55	4.99	4.68
Acebutolol	Beta-blockers	1	> 100	3.52	3.51	3.85	4.16	4.71
Amitriptyline	Antidepressants	1	0.72	5.58	5.95	5.30	5.38	5.43
Bezafibrate	Lipid lowering substances	2	> 100	3.55	5.45	4.99	4.37	4.45
Bisoprolol	Beta-blockers	3	> 100	3.51	3.64	3.92	4.57	4.37
Citalopram	Antidepressants	2	3.3	4.99	5.06	4.76	5.79	5.37
Fluvoxamine	Antidepressants	1	0.98	5.51	4.56	4.47	4.71	4.70
Paroxetine	Antidepressants	1	0.63	5.72	5.82	5.2	5.14	4.87
Pravastatine	cholesterol-lowering agent	3	> 100	3.63	3.87	4.06	4.17	4.36
Cetirizine	Antihistaminic	3	21.58	4.25	3.52	3.84	3.92	4.39
Clarithromycin	Antibiotics	3	0.23	6.51	4.62	4.50	4.35	4.51
Clindamycin	Antibiotics	1	0.01	7.62	3.86	4.05	5.47	4.86
Duloxetine	Antidepressants	1	0.37	5.90	5.77	5.18	4.46	6.05
Econazole	antifungals	-	1.37	5.44	6.48	5.60	4.35	4.92
Fenofibrate	Lipid lowering substances	2	1.34	5.43	6.14	5.41	4.88	4.98
Fluoxetine	Antidepressants	1	0.2	6.19	5.12	4.80	5.81	5.10
Hydroxyzine	Antihistaminic	1	2.13	5.24	4.01	4.14	5.48	4.95
Irbesartan	Antihypertensor	2	> 100	3.63	6.25	5.47	4.35	4.64
Ketoconazole	Antifungals	2	0.28	6.28	5.52	5.04	4.37	4.46
Miconazole	Antifungals	2	1.35	5.49	6.97	5.89	4.36	4.76
Milnacipran	Antidepressants	3	61.34	3.60	3.76	3.99	3.75	3.88
Propranolol	Beta-blockers	1	1.86	5.20	4.86	4.64	4.15	4.52
Sertraline	Antidepressants	1	0.15	6.31	6.24	5.46	4.66	5.78
Valsartan	Antihypertensor	4	> 100	3.64	6.62	5.69	4.43	4.46
Verapamil	Antiarrythmic drug	1	4.01	5.05	5.10	4.78	4.43	4.41

Tableau IV.2 – Listing of the observed and estimated EC_{50}-values (half maximal effective concentrations) for the 36 pharmaceuticals. The 12 first rows are in the validity domain for $\nu = 0.2$. BDDCS : Biopharmaceutical Drug Disposition and Classification System [211]. NSAID : Non-Steroidal Anti-Inflammatory Drug. Red rows correspond to pharmaceuticals with potentials specific MOA (LR for Linear Model, QR for Quantile Regression, $QSVMR_A$ and $QSVMR_B$ for Quantile Support Vector Machine Regression using 336 and 368 derivatives).

IV.1 Acute toxicities of pharmaceuticals toward green algae. mode of action, biopharmaceutical drug disposition classification system and quantile regression models

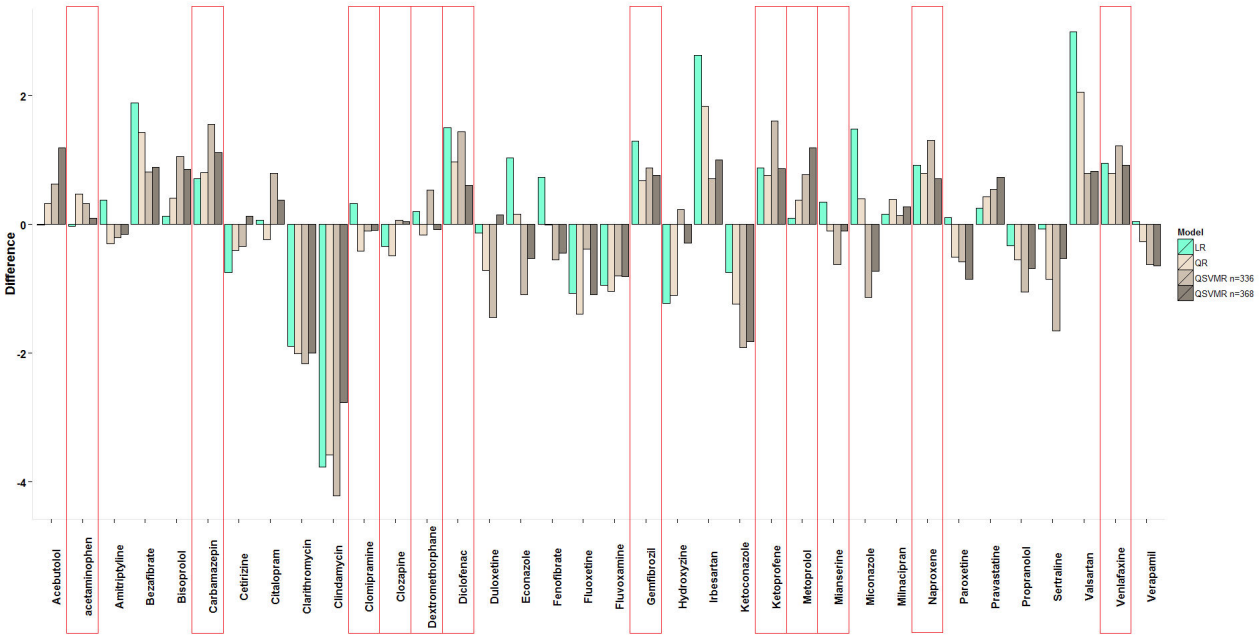


Figure IV.2 – Representation of the difference between observed and predicted EC50-values (half maximal effective concentrations) in function of the four QSAR models. Pharmaceuticals in the validity domain are framed.

In view with these first results, the structural relationships between pharmaceuticals and the chemicals associated to the different training set were analyzed in a first step. By considering our initial SVM classification based on RIF descriptors [207], all the pharmaceuticals are in the same class as the main dataset ($n = 368$). As described previously, three descriptors (log K_{OW}, molecular solubility and polarizability) are important for the different QSAR models, descriptors related to a non-specific MOA. For the definition of a validity domain, the novelty detection method, described in Section 1.2.4 and Appendix 2, was carried out with these three descriptors leading to 24 out of the 36 pharmaceuticals outside the validity domain (see Table IV.2). Due to the complex and non linear relationship between the three descriptors and the compounds, a Kernel Principal Component Analysis (KPCA) [212, 213] using a Radial Basis Function (RBF) kernel was carried out in parallel with the same descriptors to give a graphical representation of this validity domain. The data with KPCA were projected in a space spawned

	LR	QR	$QSVMR_A$	$QSVMR_B$
36 pharmaceuticals				
SE	1.26	1.05	1.05	0.92
12 pharmaceuticals				
SE	0.78	0.62	1.02	0.69
34 pharmaceuticals				
SE	1.02	0.82	0.95	0.75

Tableau IV.3 – Sandard Error (SE) of the estimate for the four QSAR models.

by 29 principal components axes. The graphical representation of the projection of compounds into the first three principal component axes is given in Figure IV.3 (this projection explains 80% of the total variability). Figure IV.2 shows that, for the 12 chemicals in the validity domain, the QSAR models estimate correctly the acute toxicities or overestimate the toxicity (less than one logarithmic unit for the best models). For the 24 compounds outside the validity domain ($\nu = 0.2$), we observed a slight under estimation of the acute toxicities, except for the two sartans, pravastatin, bezafibrate and some beta-blockers. The underestimation is particularly high for the two antibiotics (clarithromycin and clindamycin) and one antifungal agent (ketoconazole). Table 1 and Figure IV.2 display the different estimations based on the four QSAR models for the pharmaceuticals. Table 2 displays the values for the descriptors associated to QR, LR and $QSVMR_A$ (for 67 descriptors associated to $QSVMR_B$ see supporting information). The Figure IV.2 shows that, for the 12 chemicals in the validity domain, the QSAR models tend to estimate correctly the acute toxicities or to overestimate the toxicity (less than one logarithmic unit for the best models). The best models correspond to $QSVMR_B$ and QR (see Table 3 for $n = 12$, $n = 36$ and $n = 34$ (without antibiotics)). For the 24 compounds outside the validity domain ($\nu = 0.2$), we observed an inverse tendency (Figure IV.2) with a clear underestimation of the acute toxicities,except for the two sartans, pravastatin, bezafibrate and some beta-blockers. This underestimation is particularly true for the two antibiotics (clarithromycin and clindamycin)

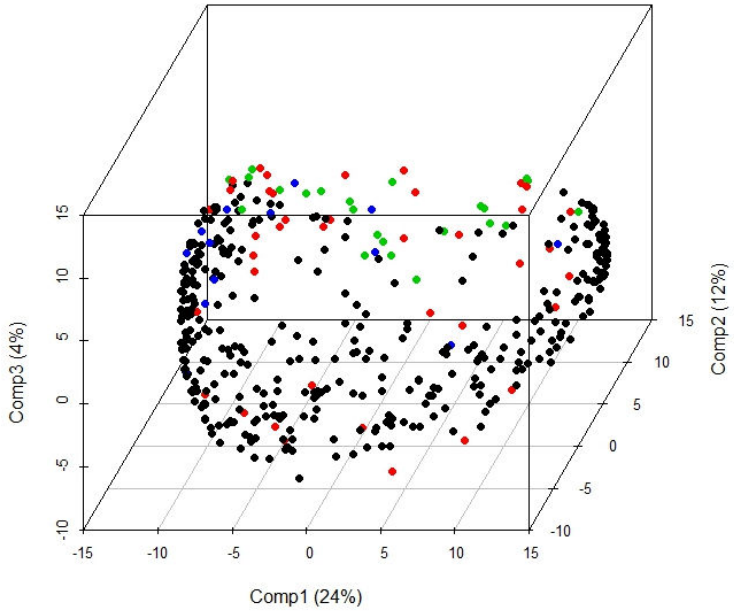


Figure IV.3 – Projection of the compounds ($401 + 36$ derivatives) into the first three components of a KPCA (RBF kernel hyper-parameter estimator of $\sigma = 0.96$) considering ALogP, molecular solubility and Apol descriptors : black (initial data set) and blue points (drugs) correspond to chemicals in the validity domain (novelty detection, $\nu = 0.2$, RBF kernel hyper-parameter estimator of $\sigma = 1.09$), red (initial data set) and green (drugs) points are out of the validity domain. A total of 80% of the total variance is explained by the three principal components axes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and one antifungal agent (ketoconazole).

1.4 Discussion

The objectives of this study are summarized on three points 1) to analyze and to compare the different models in terms of performance 2) to characterize some potential MOA of these pharmaceuticals in function of the difference between real and predicted values (based on the

notion of TR), and 3) to analyze the relationship between known properties of these drugs (pharmacodynamic and pharmacokinetic properties) and their observed acute toxicities for algae.

For the first point, two QSAR models, QSVMR$_B$ and QR, are reliable for the prediction of the acute toxicities of some classes of pharmaceuticals to *P.subcapitata*. However, we also need to understand the divergence between real and estimated toxicities by considering the second (MOA) and the third points (pharmacokinetic data). On the second point, the high difference between estimated (QSAR models) and observed ecotoxicity values for particularly antibiotics and one antifungal should be associated to a specific mode of action for these chemical derivatives. [214] showed a high toxicity of an antibiotic erythromycin to the same algal species used in the present study, *i.e. P. subcapitata.* Clarithromycin is a semisynthetic macrolide antibiotic derived from erythromycin. Clindamycin is a semisynthetic lincosamide antibiotic [215]. These three antibiotics bind to the bacterial 50S ribosomal subunit leading to an inhibition of the bacterial protein synthesis. Our results suggest a specific action on chloroplast ribosomes of these antibiotics which contain homologs of all the bacterial 70S ribosomal proteins [216]. Ketoconazole is an antifungal agent interacting with $14-\alpha$ demethylase, a cytochrome P-450 enzyme necessary for the conversion of lanosterol to ergosterol in fungi. Its action leads to the inhibition of ergosterol synthesis and an increase of the fungal cellular permeability. Ergosterol is the predominant sterol of fungi and green algae. A similar MOA is likely by considering the biological pathway associated to sterol synthesis in green algae [217]. Miconazole and econazole, the two antifungals in the dataset, with the same biological mechanisms and similar structures, should have like ketoconazole a specific MOA. However to explain a correct estimation of their ecotoxicities with log P only, these azole derivatives should have a potential non-specific MOA. In fact, these antifungals are known to exert direct physicochemical cell membrane damage

[218] in function of the concentration (high concentration in this case). So, the determination of the MOA (non specific vs specific) shows some ambiguities for antifungals if we consider the differences between estimated and observed values. Our results for these antifungals are in favor of a non specific MOA.

Beside these results associated to a specific MOA (antibiotics), an overestimation of the toxicity for a set of derivatives with an acidic function was observed (NSAID, valsartan, ibesartan, pravastatin, bezafibrate, cetirizine, gemfibrozil). Nearly half of these derivatives have EC_{50} values superior to 100 mg/L. The overestimation represents on average 0.65 logarithmic unit (by considering 100 mg/L for pharmaceuticals with $EC_{50} > 100$ mg/L). For human medication, pharmaco kinetics of drugs are of ten associated to specific uptake and efflux transporters like the OATP, BCRP or MRP2 [219, 220]. The Biopharmaceutical Drug Disposition Classification System (BDDCS) was introduced in 2005 [211] with as objective an estimation on overall drug disposition, including routes of drug elimination and the effects of efflux and absorptive transporters on oral drug absorption (Absorption Distribution Metabolism Excretion properties (ADME)). From BDDCS, four classes were fixed with notably for each class the potential impact of transporters [221]. Class 1 is characterized by high solubility and permeability with low impact of transporters (a drug substance is considered “highly soluble” when the highest dose strength is soluble in 250 ml or less of aqueous media over a pH range of $1 - 7.5$ at 37 °C). Class 2 is characterized by low solubility and high permeability with high impact of uptake and efflux transporters. Class 3 is characterized by high solubility and low permeability with high impact of absorptive transporter and class 4 is characterized by low solubility and permeability. The last definition of the classes (BDDCS vs BCS for Biopharmaceutic Classification System [211]) considers as a fundamental point the metabolism of the drugs. On the twelve derivatives associated to the validity domain, seven are in class 1. They are correctly predicted and par-

ticularly drugs like clomipramine, dextromethorphan and mianserin. However for metoprolol (beta-blocker) and venlafaxine (antidepressant), an overestimation of the toxicity was observed. We will discuss afterward these cases by considering the complete family of beta-blockers and a structural comparison with venlafaxine. For the class 2 derivatives in the validity domain, an overestimation is observed except for clozapine. This overestimation for the class 2 is pointed out in the overall set ($n = 36$) except for fenofibrate. On this point, the main difference between gemfibrozil and fenofibrate is the acidic function for gemfibrozil vs an ester function for fenofibrate. From these results, it is clear that an acidic function, beside hydrophobic properties, has an impact for the diffusion of these chemicals in algae cells (permeability of the membrane). Always, on the overall set, the class 3 is characterized by a low toxicity (on the five derivatives, three have an $EC_{50} > 100$ mg/L) except for antibiotics. In this last case, the possibility for macrolides to penetrate into the algal cell is clearly demonstrated by their high toxicities (erythromycin and clarithromycin). Valsartan is the only drug in the class 4 [221] with no toxicty for *P. subcapitata*.

For class 1 pharmaceuticals, correct predictions or a low underestimation of acute toxicities were obtained. However, beta-blockers showed a particular behavior. Indeed, only propranolol was correctly predicted, the others are overpredicted. Beta-blockers are associated to a common pharmacophoric fragment (aromatic ring/alkyl chain with a hydroxyl group/basic amine). The difference between the beta-blockers is based on the hydrophilic/hydrophobic characteristic of the chemical fragments connected to the aromatic ring. A long aliphatic chain on para position leads to a class 3 drug (bisoprolol) and a selectivity to β_1 adrenergic receptor. Metoprolol is close to bisoprolol with a shorter aliphatic chain on para position and the same selectivity. Propranolol has no aliphatic chain but a naphthyl group instead of a phenyl group. With this specific structural characteristic, one beneficial effect described for propranolol was a strong

membrane stabilizing action [222]. This is not observed for metoprolol and bisoprolol (or at a very low level). For acebutolol the potency of the membrane-stabilizing action of propranolol is much lower than propranolol by considering data (internal and external) described in the publication of Takeo and al. [222] (the membrane-stabilizing action was found to be exerted at concentrations of more than 10 μM for propranolol in rabbit atrial muscle [223] and 134 μM for acebutolol in guinea pig papillary muscle [224]. In contrast, atenolol and metoprolol can only exhibit their membrane-stabilizing action at concentrations of 442 and 292 μM, respectively [225] and [226]. The agreement between this membrane stabilizing action and the algae toxicity is clearly seen by considering the differences of ecotoxicity for these beta-blockers. Most of the antidepressants (class 1 mainly) in our set have the characteristics associated to an optimum interaction with the biological membrane *i.e.* a lipophilic aromatic group capable of intercalating between phospholipids in the membrane, an amino group capable of associating with an anionic group and an alkyl chain connecting the two features. Antidepressants correspond to the most toxic derivatives for algae, toxicity associated clearly to a non-specific MOA. This is confirmed by a correct QSAR prediction with log P (LR, $R^2 = 0.74$, $n = 12$). For antidepressants in the class 1, only venlafaxine has a different behavior. Venlafaxine is a serotonine-noradrenalin reuptake inhibitor (SNRI) with specific structural characteristics notably a tertiary alcohol and a methoxy group on para position of the phenyl ring.Venlafaxine possesses structural similarities with other beta-blockers like metoprolol but also with antidepressants like selective serotonin reuptake inhibitors (SSRIs : fluoxetine, paroxetine, sertraline). These structural characteristics relatively unique for venlafaxine should explain this different behavior. For antidepressant, this underestimation is particularly true for fluoxetine. But, it is difficult to associate a specific MOA from this observation. Moreover one previous study on fluoxetine demonstrated a non-specific MOA for algae [227].

1.5 Conclusion

Our quantile regression models translate correctly the toxicities of class 1 drugs, displaying a strong interaction with biological membrane (narcosis as MOA) due to high permeability of these derivatives. Without considering the validity domains of the models, BDDCS classification gives a correct idea on the potential ecotoxicity of this class starting from our QSAR models. Pharmaceuticals in the class 1 are considered to be highly soluble and with an extensive metabolism. Indeed, drugs like fluoxetine are extensively metabolized to norfluoxetine and other metabolites [215]. Herein, for the selection of pharmaceuticals, PEC values were extrapolated without considering the metabolism of the drugs. So, in terms of risk, for the class 1, metabolites should be analyzed in priority and the ecotoxicity of the pharmaceuticals can be estimated with our QSAR models. Pharmaceuticals in the class 2 are characterized by a low solubility and a high metabolism. A weak overestimation is classically recorded from our QSAR models. Classes 3 and 4 are characterized by their poor metabolism and low permeability (a relationship between metabolism and permeability is recorded). Most of them have no toxicity ($EC_{50} > 100$ mg/L) leading to a strong overestimation from our models. These pharmaceuticals (class 3 and 4) are really problematic in case of a very high toxicity to algae. Clarithromycin, a class 3 drug, is observed in surface water with relatively high concentration [228] and must be considered clearly as one of the most problematic pharmaceuticals for algae.

1.6 Acknowledgements

M. Jonathan Villain was supported by a grant from the Region of Bretagne and the French Ministry of Education. The authors thank the Agence Nationale de la Recherche (ANR, ANR-07-CP2D-09-02 and Pharm@ecotox) for financial supports.

1.7 Supporting information

The data are accessible through the website http ://www.cermn.unicaen.fr without user registration. The R codes are also available upon request to the authors.

2 Appendix : novelty detection

The novelty detection or one-class classification [229] is a SVM method to detect outliers in dataset by a classification method [210]. In Support Vector Machine classification, we search to estimate classes using multi-classification problems by decomposing the classification problem in multiple two-class classification problems. The approach of one-class classification is to create a decision boundary around the data in a feature space F defined by a mapping transformation $\phi : \mathbb{R}^n \to F$. In practice by the kernel trick [112, 230], we do not need to know the explicit form of the function ϕ. The dot product in the image of ϕ can be simply calculated by evaluating a kernel function

$$K(x, x') = \langle \phi(x), \phi(x') \rangle$$

For instance using the Radial Basis Function (RBF) kernel, we obtain :

$$K(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right). \tag{IV.2}$$

where σ denotes the bandwidth parameter.

The method determines a set of support vectors describing the decision boundary in the feature space. The decision boundary is defined by the solution of the following quadratic minimization problem with respect to $w \in F$, $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$ and $\rho \in \mathbb{R}$

$$\frac{1}{2}\|w\|^2 + \frac{1}{n\nu}\sum_{i=1}^{n}\xi_i - \rho \tag{IV.3}$$

subject to $\langle \phi(x_i), w \rangle \geq \rho - \xi_i$ and $\xi_i \geq 0$ with $i = 1, \ldots, n$ and ν represents an upper bound on the fraction of data that may be outliers. To find a solution to the minimization problem IV.3, we consider the Lagrangian given by

$$\mathcal{L}(w, \xi, \rho, \alpha, \beta) = \frac{1}{2}\|w\|^2 + + \frac{1}{n\nu}\sum_{i=1}^{n}\xi_i - \rho - \sum_{i=1}^{n}\alpha_i\left(\langle w, \phi(x_i)\rangle - \rho + \xi_i\right) - \sum_{i=1}^{n}\beta_i\xi_i,$$

where, for $i = 1, \ldots, n$, α_i, β_i are positive Lagrangian multipliers. Solving the derivatives of the Lagrangian with respect to w, ξ, ρ equal to zero, we obtain

$$w = \sum_{i=1}^{n}\alpha_i K(x_i, x),$$

where, for $i = 1, \ldots, n$, $\alpha_i = \frac{1}{n\nu} - \beta_i \leq \frac{1}{n\nu}$ and $K(x_i, x)$ is a kernel function. The decision function is

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i K(x_i, x) - \rho\right)$$

where sgn is the sign function ($\operatorname{sgn}(x) = 1$ for $x \geq 0$ and -1 otherwise). The dual formulation of the problem can be expressed as

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x_i, x_j).$$

subject to $0 \leq \alpha_i \leq \frac{1}{n\,\nu}$ and $\sum_{i=1}^{n}\alpha_i = 1$.

The overall margin ρ can be calculated for a given x_i by

$$\rho = \langle w, \phi(x_i)\rangle = \sum_{j=1}^{n}\alpha_j K(x_j, x_i).$$

When ν approaches 0, the upper bounds on the Lagrangian multipliers tend to infinity. In that case, the constraint of the minimization problem (IV.3) becomes void. We then have a hard margin problem since the penalization of the errors become infinite. The problem is still feasible since we have no restriction on ρ and so ρ can be a large negative number to satisfy the constraint given in (IV.3). When ν is equal to 1, we have $\alpha_1, \ldots, \alpha_n = 1/n$ and so, the decision

function corresponds to the kernel Parzen-Rosenblatt estimator of the unknown probability density function [231–233].

Chapitre V

Statistique séquentielle et régression quantile

Nous construisons dans ce paragraphe une méthode séquentielle pour les modèles QSAR. Dans la première partie, nous décrivons dans le cas univarié une procédure séquentielle publiée en 2009 par Durrieu et Briollais [178] pour un paramètre réel du vecteur de régression et par conséquent nous nous limitons à une seule variable explicative d'un modèle de régression (cadre unidimensionnel) et dans une seconde partie nous généralisons les résultats au cas multidimensionnel. L'objectif est d'appliquer ensuite ces résultats en chémoinformatique.

Un aspect qui nous a motivé est le nombre et le mode d'acquisition des données. Tout d'abord, l'acquisition des données recueillies par les chimistes et biologistes du centre d'étude et de recherche sur le médicament de Normandie de l'université de Caen est réalisée au fur et à mesure dans le temps. Les méthodes de statistique séquentielle sont tout à fait adaptées à ce type de situation où l'objectif principal est la précision que l'on souhaite obtenir. La définition d'une règle d'arrêt nous indique si nous pouvons arrêter l'expérience après les n premières mesures ou si nous devons continuer avec une nouvelle observation ou avec un nouveau lot d'observations. Cette règle d'arrêt est cruciale et nous nous intéressons ici à l'ajout séquentiel d'une seule observation.

1 Cas univarié

Nous considérons le modèle linéaire :

$$\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n \tag{V.1}$$

où pour tout $n \geq 1$, $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^\top$ est le vecteur des observations, $\boldsymbol{X}_n$ est une matrice connue de dimension $n \times p$ ayant pour lignes $\boldsymbol{x}_i^\top \in \mathbb{R}^p$, $i = 1, \ldots, n$, $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ est un vecteur d'erreurs indépendantes et identiquement réparties (i.i.d.), de fonction de répartition F inconnue et de médiane nulle ($F^{-1}(1/2) = 0$) et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ désigne le vecteur inconnu des paramètres de régression à estimer.

Au départ sont fixés une précision $d > 0$ et un seuil α dans l'intervalle $]0, 1[$. Notre objectif est de construire un intervalle de confiance I_n pour β_1, basé sur un estimateur robuste de type quantile, telle que sa longueur L_n satisfait

$$L_n \leq 2d, \tag{V.2}$$

et qui vérifie

$$P_F\left(\beta_1 \in I_n\right) \geq 1 - \alpha. \tag{V.3}$$

Bien entendu, avec de telles conditions, n ne peut être fixé *a priori* et nous sommes naturellement placés dans le cadre d'une procédure séquentielle. Stein en 1945 [234] a prouvé que, dans le cas où $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ est un vecteur gaussien, ce type d'intervalle peut être construit par une procédure à deux pas. Plus tard, Chow et Robbins en 1965 [235] ont proposé une procédure séquentielle dans le cas d'une population avec variance finie. Pour le modèle linéaire, des méthodes similaires ont été construites par Ghosh et Sen en 1972 [236] en utilisant des statistiques de rang pour estimer le paramètre de régression et par Jurečková et Sen en 1996 [141]. Nous décrivons ici la procédure de construction d'un intervalle de confiance d'un para-

mètre de régression dans un modèle linéaire lorsque la fonction de répartition des erreurs est supposée inconnue et un estimateur L^1 du coefficient de régression est choisi. Plus précisément, nous nous mettons dans le cas de la régression médiane.

Le schéma de construction repose essentiellement sur deux étapes :

— **Première étape** : nous déterminons deux estimateurs $\widehat{\beta}_1$ et $\widehat{\Xi}_n$ tel que $I_n = \left[\widehat{\beta}_1 - \widehat{\Xi}_n, \widehat{\beta}_1 + \widehat{\Xi}_n\right]$ soit un intervalle de confiance de β_1 de coefficient de confiance $1-\alpha$. Cette construction se fait de manière classique en utilisant les propriétés des estimateurs (normalité asymptotique de $\widehat{\beta}_1$ et consistance de $\widehat{\Xi}_n$).

— **Deuxième étape** : en nous plaçant maintenant dans le contexte de l'analyse séquentielle, nous étudions la variable d'arrêt N_d qui correspond au plus petit entier $n \geq n_0$ telle que la longueur de I_{N_d} est inférieure ou égale à $2d$. Nous discuterons l'introduction du paramètre n_0 et le choix de sa taille ultérieurement.

La mise en place de cette méthode permet de déterminer avec la précision souhaitée *a priori* à quel intervalle appartient la valeur du paramètre étudié, avec un nombre minimal de données.

2 Étude théorique

Nous utilisons les notations introduites dans II.1.1 et les résultats de convergence du II.1.

Dans la suite, nous avons choisi l'estimateur L^1, qui est probablement le plus vieil estimateur robuste. Galilée utilisa un tel estimateur en 1632 pour résoudre un différend entre les astronomes de l'époque. Boscovich en 1757 [237] discuta ce mode d'estimation dans le cas du modèle de régression linéaire simple. Dans l'ouvrage de Laplace en 1895 [238] : "Le deuxième supplément à la théorie analytique des probabilités" ; on trouve une étude sur la médiane. Bien plus tard, Edgeworth en 1887 [239] a de nouveau abordé ce sujet pour un modèle de régression linéaire. On peut se reporter à Farebrother en 1987 [240] pour une discussion des différentes contributions de

Laplace, Edgeworth, ... à la solution géométrique, graphique et analytique pour les problèmes d'estimation L^1. Une bibliographie assez exhaustive relative à cette méthode d'estimation est donnée par Gentle en 1977 [241], et Bloomfield et Steiger en 1983 [242].

Nous présentons maintenant les deux étapes nous permettant de déterminer, avec une précision donnée *a priori*, des intervalles de confiance de paramètre de régression.

2.1 Première étape : préliminaires

Dans la section II.1, nous introduisons les "quantiles de régression" permettant la détermination des estimateurs. On appelle θ-quantile de régression toute solution du problème de minimisation

$$\widehat{\boldsymbol{\beta}}(\theta) \equiv \widehat{\boldsymbol{\beta}}^n(\theta) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\theta\left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \tag{V.4}$$

Un cas particulier de cette classe d'estimateurs (obtenu pour $\theta = 1/2$) est l'estimateur L^1 qui s'obtient par résolution du problème de minimisation (V.4). Nous avons également précisé que lorsque $n \to \infty$

$$\sqrt{n}\left(\widehat{\beta}_1^n(1/2) - \beta_1\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{q^2(1/2)}{4}\right).$$

La variance asymptotique dépendant de la densité de probabilité des erreurs qui est inconnue, nous avons besoin de "bons" estimateurs de la variance asymptotique "invariants" par régression et par homothétie. Nous utiliserons dans la suite l'estimateur de type noyau de $q(\theta)$, $0 < \theta < 1$, défini par

$$\widehat{Z}_n(\theta) = \frac{1}{\nu_n^2} \int_0^1 \widehat{\beta}_1^n(w)\, k\left(\frac{\theta - w}{\nu_n}\right) dw,$$

où les conditions *C8'*, *C9* sur la taille de la fenêtre sont satisfaites et k est une fonction (noyau) vérifiant *C10*. Cet estimateur est décrit dans le Chapitre II.

2.2 Deuxième étape : statistique séquentielle

D'après (II.10) du Chapitre II, on obtient : $n \to \infty$, on a

$$\widehat{W}_n(1/2) \xrightarrow{P} \frac{1}{2\, f(0)}.$$

En combinant ceci avec le résultat de normalité asymptotique, le théorème de Slutsky permet d'écrire

$$\frac{\sqrt{n}\,\left(\widehat{\beta}_1^n(1/2) - \beta_1\right)}{\widehat{W}_n(1/2)} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ quand } n \to \infty.$$

On a donc

$$\lim_{n\to\infty} P_F\left\{\left|\frac{\sqrt{n}\,(\widehat{\beta}_1^n(1/2) - \beta_1)}{\widehat{W}_n(1/2)}\right| \leq z_{1-\alpha/2}\right\} = 1-\alpha,$$

et par conséquent

$$I_n = \left[\widehat{\beta}_1^n(1/2) - \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\widehat{W}_n(1/2), \widehat{\beta}_1^n(1/2) + \frac{z_{1-\alpha/2}}{\sqrt{n}}\,\widehat{W}_n(1/2)\right] \tag{V.5}$$

est un intervalle de confiance asymptotique de niveau de confiance $(1-\alpha)$ de β_1.

La longueur L_n de l'intervalle I_n vérifie alors

$$\sqrt{n}\, L_n = 2\, z_{1-\alpha/2}\widehat{W}_n(1/2) \xrightarrow{P} \frac{z_{1-\alpha/2}}{f(0)} \text{ quand } n \to \infty.$$

Pour une taille d'échantillon n et pour α fixés, la longueur de l'intervalle de confiance I_n est une variable aléatoire qui n'a aucune raison d'être comparable à une longueur donnée *a priori*. La procédure séquentielle consiste à ajouter aux n_0 premières mesures une nouvelle observation et à vérifier sur le nouvel échantillon ainsi obtenu si la condition $L_n \leq 2d$ est satisfaite. Cette procédure prend fin pour la plus petite valeur de $n \geq n_0$ pour laquelle $L_n \leq 2d$.

En comparant

$$I_n^* = [\,\widehat{\beta}_1^n(1/2) - d,\, \widehat{\beta}_1^n(1/2) + d\,] \tag{V.6}$$

($d > 0$ fixé) avec (V.5), la longueur de (V.5) est majorée par $2\,d$ dès lors que

$$n \geq \frac{z_{1-\alpha/2}^2 \widehat{W}_n^2(1/2)}{d^2}. \tag{V.7}$$

Pour éviter d'obtenir de mauvais résultats au niveau du calcul des estimateurs nous devons choisir une taille initiale n_0 qui ne soit pas trop petite. Nous n'avons pas de résultats théoriques permettant de choisir cette taille initiale mais des simulations ont montré que ce paramètre ne doit pas être choisi trop petit [178].

Ainsi, la variable d'arrêt N_d est définie par

$$N_d = \min \left\{ n \geq n_0 \;\middle|\; n \geq \frac{z_{1-\alpha/2}^2 \widehat{W}_n^2(1/2)}{d^2} \right\}. \tag{V.8}$$

2.3 Propriétés asymptotiques

Nous commençons par décrire le comportement asymptotique (quand l'amplitude d de l'intervalle de confiance tend vers zéro) de la procédure séquentielle : consistance et convergence en moyenne quadratique de la variable d'arrêt et niveau de confiance de l'intervalle de confiance de longueur donnée du paramètre de régression obtenu.

Théorème V.1. *Sous les conditions C1 – C7 et C8 – C10, nous avons :*

1. $E_F(N_d) \to +\infty$ *quand* $d \to 0_+$,
2. $P_F\{N_d < +\infty\} = 1$ *pour tout* $d > 0$,
3. $N_d/n_d \xrightarrow{P} 1$ *quand* $d \to 0_+$, *où* $n_d = z_{1-\alpha/2}^2\,\sigma^2/d^2$,
4. $P_F\{I_{N_d} \in \beta_1\} \geq 1 - \alpha$ *pour* $\alpha \in]0,1[$ *fixé, quand* $d \to 0_+$.

Preuve du Théorème V.1.

1) Du fait que par construction la variable d'arrêt N_d croît quand d décroît et que $L_n > 0$ avec probabilité 1, nous obtenons :

$$\forall m \geq 0, \quad \lim_{d \to 0_+} P_F\{N_d < m\} \leq \lim_{d \to 0_+} P_F\{L_{N_d} < 2\,d\} = 0,$$

et donc $N_d \xrightarrow{P} \infty$ quand $d \to 0_+$. Ainsi, en utilisant le théorème de la convergence monotone nous avons :

$$\lim_{d\to 0} E_F\left(N_d\right) = E_F\left(\lim_{d\to 0} N_d\right) = \infty.$$

2) Pour tout $n \geq n_0$, nous avons :

$$P_F\left\{N_d > n\right\} \leq P_F\left\{n < d^{-2}\, z^2_{1-\alpha/2}\, \widehat{W}^2_n(1/2)\right\}.$$

D'après le Théorème (II.5), nous déduisons que $\dfrac{\widehat{W}_n(1/2)}{n} \xrightarrow{P} 0$ quand $n \to \infty$, et par conséquent

$$P_F\{N_d = \infty\} = \lim_{n\to\infty} P_F\{N_d > n\} \leq \lim_{n\to\infty} P_F\{n < d^{-2}\, z^2_{1-\alpha/2}\, \widehat{W}^2_n(1/2)\} = 0.$$

On conclut que :

$$P_F\{N_d = \infty\} = 0 \qquad \text{et} \qquad P_F\{N_d < +\infty\} = 1 \qquad \text{pour tout } d > 0.$$

3) Avec probabilité 1, nous avons

$$\frac{z^2_{1-\alpha/2}\widehat{W}^2_{N_d}(1/2)}{d^2} \leq N_d < \max\left(n_0 + 1, \frac{z^2_{1-\alpha/2}\widehat{W}^2_{N_d-1}(1/2)}{d^2} + 1\right).$$

Comme $n_d = \dfrac{z^2_{1-\alpha/2}\,\sigma^2}{d^2}$, nous avons

$$\frac{\widehat{W}^2_{N_d}(1/2)}{\sigma^2} < \frac{N_d}{n_d} < \max\left(\frac{n_0}{n_d}, \frac{\widehat{W}^2_{N_d-1}(1/2)}{\sigma^2}\right) + \frac{1}{n_d}. \tag{V.9}$$

De plus, en utilisant

$$\widehat{W}_n(1/2) \xrightarrow{P} \sigma \;\text{ quand }\; n \to \infty, \qquad N_d \xrightarrow{P} \infty \;\text{ quand }\; d \to 0_+$$

et l'uniforme continuité en probabilité de $\widehat{W}_n(1/2)$, nous obtenons

$$\frac{\widehat{W}_{N_d}(1/2)}{\sigma} \xrightarrow{P} 1 \qquad \text{et} \qquad \frac{\widehat{W}_{N_d-1}(1/2)}{\sigma} \xrightarrow{P} 1 \;\text{ quand }\; d \to 0_+. \tag{V.10}$$

D'après (V.10), nous avons quand $d \to 0_+$

$$\max\left(\frac{n_0}{n_d}, \frac{\widehat{W}^2_{N_d-1}(1/2)}{\sigma^2}\right) \longrightarrow \max(0,1) = 1.$$

Finalement, par (V.9) nous déduisons

$$\frac{N_d}{n_d} \xrightarrow{P} 1 \quad \text{quand} \quad d \to 0_+.$$

4) Puisque $\widehat{\beta}_1^n$ est asymptotiquement gaussien et uniformément continu en probabilité, nous obtenons :

$$\frac{\sqrt{N_d}\,\left(\widehat{\beta}_{N_d}(1/2) - \beta_1\right)}{\widehat{W}_{N_d}(1/2)} = \left(\frac{\sigma}{\widehat{W}_{N_d}(1/2)}\right)\left(\frac{N_d}{n_d}\right)^{1/2}\left(\frac{\sqrt{n_d}\,\left(\widehat{\beta}_{N_d}(1/2) - \beta_1\right)}{\sigma}\right)$$

$$= \left(\frac{\sigma}{\widehat{W}_{N_d}(1/2)}\right)\left(\frac{N_d}{n_d}\right)^{1/2}\left\{\frac{\sqrt{n_d}\,\left(\widehat{\beta}_{n_d}(1/2) - \beta_1\right)}{\sigma} + \frac{\sqrt{n_d}\,\left(\widehat{\beta}_{N_d}(1/2) - \widehat{\beta}_{n_d}(1/2)\right)}{\sigma}\right\}$$

$$\xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \qquad \text{quand } d \to 0_+.$$

Lorsque $d \to 0_+$, nous avons :

$$\lim_{d\to 0_+} P_F\left\{\left|\widehat{\beta}_{N_d}(1/2) - \beta_1\right| \le z_{1-\alpha/2}\,\widehat{W}_{N_d}(1/2)\,N_d^{-1/2}\right\} = 1-\alpha. \tag{V.11}$$

Par (V.11) et (V.8), nous avons finalement :

$$\lim_{d\to 0_+} P_F\left\{\widehat{\beta}_{N_d}(1/2) - d \le \beta_1 \le \widehat{\beta}_{N_d}(1/2) + d\right\} \ge 1-\alpha\ .$$

Dans ce qui suit, nous présentons des résultats intermédiaires nécessaires à la construction de la méthode séquentielle et à l'étude de son comportement asymptotique. Ces résultats couvrent essentiellement des problèmes de continuité uniforme en probabilité parfois appelés "condition d'Anscombe" [243] dont nous rappelons la définition ci-dessous.

Définition V.1. *Soit $(Y_n)_{n\ge1}$ une suite infinie de variables aléatoires. On dit que $(Y_n)_{n\ge1}$ est uniformément continue en probabilité si et seulement si :*

$\forall \epsilon > 0\ \forall \eta > 0\ \exists\, \nu,\, c > 0$ *tel que* $\forall n > \ \nu$

$$P\left\{ \max_{|n'-n|<c\,n} |Y_{n'} - Y_n| > \epsilon \right\} < \eta.$$

Ces résultats intermédiaires ont un intérêt intrinsèque en cela qu'ils précisent les vitesses de convergence de nos estimateurs.

À partir du Lemme V.1 et du Théorème V.2, il est possible de prouver la continuité uniforme en probabilité de l'estimateur L^1 et de l'estimateur de la variance asymptotique.

Lemme V.1. *Soit C un réel positif, $\gamma \in (0, a)$ et*

$$B_n = \max\left(\frac{1}{n^{2(a-\gamma)/(1+4b)}}, \frac{1}{n^{(2-\gamma)(b-a)/(1+4b)}}, \frac{1}{n^{(b-\gamma)/(1+4b)}} \right).$$

Sous les hypothèses C1 – C7, pour tout $\mu > 0$, il existe $A_\mu > 0$ tel que pour $n \to \infty$, nous avons

$$P\left\{ \sup |r_n(\boldsymbol{\beta}, \theta)| \geq A_\mu\, B_n : \|\boldsymbol{\beta}\| \leq C\sqrt{\log n}, \theta_n^* \leq \theta \leq 1 - \theta_n^* \right\} = O\left(\frac{1}{n^\mu}\right)$$

où

$$\begin{aligned} r_n(\boldsymbol{\beta}, \theta) \ &= \frac{1}{\sqrt{\theta(1-\theta)}\,\sigma_\theta} \sum_{i=1}^{n} \left[\rho_\theta\left(\varepsilon_{i\theta} - \frac{\sigma_\theta \boldsymbol{x_i}^\top \boldsymbol{\beta}}{\sqrt{n}} \right) - \rho_\theta\left(\varepsilon_{i\theta}\right) \right] \\ &+ \frac{1}{\sqrt{n\ (\theta(1-\theta))}} \sum_{i=1}^{n} \boldsymbol{x_i}^\top \boldsymbol{\beta} \psi_\theta(\varepsilon_{i\theta}) - \frac{\boldsymbol{\beta}^\top \boldsymbol{D}_n \boldsymbol{\beta}}{2}, \end{aligned}$$

et

$$\varepsilon_{i\theta} = \varepsilon_i - Q(\theta) \qquad \textit{pour } i = 1, \ldots, n.$$

Preuve du Lemme V.1. Durrieu and Briollais (2009), [178].

Ce lemme permet de démontrer le théorème suivant :

Théorème V.2. *Sous les conditions C1 – C7, pour tout $\mu > 0$, il existe $A_\mu > 0$ tel que pour $n \to \infty$, nous avons*

$$P\left\{ \left\| \frac{\sqrt{n}}{\sigma_\theta} \left(\widehat{\boldsymbol{\beta}}^n(\theta) - \widetilde{\boldsymbol{\beta}}(\theta) \right) - \frac{1}{\sqrt{n}\sqrt{\theta\,(1-\theta)}} \boldsymbol{D}_n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \psi_\theta\left(\varepsilon_{i\theta}\right) \right\| \geq A_\mu\, B_n \right\} = O\left(\frac{1}{n^\mu}\right).$$

Preuve du Théorème V.2. Durrieu and Briollais (2009), [178].

Nous déduisons alors de ce résultat les deux corollaires qui sont la clef de voûte de la construction des procédures séquentielles.

Théorème V.3.
(i) On suppose que les conditions C1 – C7 sont satisfaites. Alors la suite

$$\left(\sqrt{n}\left(\widehat{\beta}_1^n(\theta) - \beta_1 - Q(\theta)\right)\right)_{n\geq 1}$$

est uniformément continue en probabilité.
(ii) On suppose que les conditions C1 – C10 sont satisfaites. Alors la suite

$$\left(\sqrt{n\,\nu_n}\left(\widehat{W}_n(\theta) - \frac{q(\theta)}{2}\right)\right)_{n\geq 1}$$

est uniformément continue en probabilité.

Preuve du Théorème V.3. Durrieu and Briollais (2009), [178].

3 Théorème limite centrale de la variable d'arrêt

Nous étudions le comportement asymptotique de la variable d'arrêt. Dans ce qui suit nous posons $\nu_n = n^{-\delta}$ (avec $\delta = 1 - 2\,\eta = 0.42 \in]1/3, 1/2[$) ; des calculs montrent alors que pour ce choix de ν_n les conditions sur les tailles de fenêtre *C8– C9* sont vérifiées pour tout $\delta \in]1/3, 1/2[$ c'est-à-dire pour tout $\eta \in]1/4, 1/3[$. Voici un résultat intermédiaire :

Lemme V.2. *Soit $\eta \in]1/4, 1/3[$ où ν_n est tel que $\nu_n = n^{2\,\eta-1}$. Sous les conditions C1 – C7 et C10, on a lorsque $d \to 0_+$:*
(i)

$$N_d^\eta\left(\widehat{W}_{N_d}(1/2) - \sigma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\,\overline{K}\right),$$

(ii)

$$N_d^\eta\left(\widehat{W}_{N_d-1}(1/2) - \sigma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\overline{K}\right),$$

(iii) pour tout $\epsilon > 0$ *:*

$$P\left\{ N_d^{\eta} \left| \left(\frac{\sqrt{N_d} - \sqrt{N_d - 1}}{\sqrt{\overline{K}}\,\sigma} \right) \frac{d}{z_{1-\alpha/2}} \right| > \epsilon \right\} \longrightarrow 0.$$

Preuve du Lemme V.2. Durrieu and Briollais (2009), [178].

Le lemme V.2 nous permet de prouver la normalité asymptotique de la variable d'arrêt pour l'estimateur de type noyau et l'estimateur de type histogramme.

Théorème V.4. *Soit* $\eta \in]1/4, 1/3[$ *où* ν_n *est tel que* $\nu_n = n^{2\,\eta-1}$*. Sous les conditions C1 – C7 et C10, nous avons quand* $d \to 0_+$ *:*

$$d^{1-2\eta}\left(\sqrt{N_d} - \frac{z_{1-\alpha/2}\,\sigma}{d} \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^{2(1-2\eta)}\, z_{1-\alpha/2}^{2(1-2\eta)}\, \overline{K}\right).$$

Nous commençons par décrire des résultats auxiliaires. Nous étudions maintenant le comportement de la variable d'arrêt N_d. Nous supposons que $\nu_n = n^{2\,\eta-1}$ pour $\eta \in (1/4, 1/3)$.

Lemma V.1. *Soit* $\eta \in (1/4, 1/3)$ *et* ν_n *tels que* $\nu_n = n^{2\eta-1}$*. Sous des conditions de régularité, nous avons quand* $d \to 0_+$ *:*

(i)

$$N_d^{\eta}\left(\widehat{W}_{N_d}(1/2) - \sigma \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\, \overline{K}\right), \tag{V.12}$$

(ii)

$$N_d^{\eta}\left(\widehat{W}_{N_d-1}(1/2) - \sigma \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2 \overline{K}\right), \tag{V.13}$$

(iii) pour tout $\epsilon' > 0$ *:*

$$P\left\{ N_d^{\eta} \left| \left(\frac{\sqrt{N_d} - \sqrt{N_d - 1}}{\sqrt{\overline{K}}\,\sigma} \right) \frac{d}{z_{1-\alpha/2}} \right| > \epsilon' \right\} \longrightarrow 0. \tag{V.14}$$

Preuve du Lemme V.1.

(i) D'après le Théorème (II.5), nous obtenons

$$\sqrt{n\nu_n}\left(\widehat{W}_n(1/2) - \sigma \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2 \overline{K}\right). \tag{V.15}$$

En prenant $\nu_n = n^{2\,\eta-1}$, nous avons

$$\sqrt{N_d \nu_{N_d}}\left(\widehat{W}_{N_d}(1/2)-\sigma\right) = \left(\frac{N_d}{z_{1-\alpha/2}^2\sigma^2/d^2}\right)^{\eta}\left(\frac{z_{1-\alpha/2}^2\sigma^2}{d^2}\right)^{\eta}\left(\widehat{W}_{N_d}(1/2)-\sigma\right). \tag{V.16}$$

D'après le troisième résultat du Théorème V.1, nous avons quand $d \to 0_+$,

$$\frac{N_d^{\eta}}{\dfrac{z_{1-\alpha/2}^{2\eta}\,\sigma^{2\eta}}{d^{2\eta}}} \xrightarrow{P} 1 \tag{V.17}$$

et par (V.15), quand $d \to 0_+$,

$$P\left\{n^{\eta}\left(\widehat{W}_n(1/2)-\sigma\right) \le y\right\} \longrightarrow \Phi\left(\frac{y}{\sqrt{\overline{K}}\,\sigma}\right), \ \forall y \in \mathbb{R}, \tag{V.18}$$

où Φ est la fonction de répartition d'une loi normale centrée et réduite.

En complément de (V.17) et (V.18), le théorème Anscombe [243] permet de conclure que, $\forall y \in \mathbb{R}$,

$$P\left\{\frac{\sigma^{2\eta}\,z_{1-\alpha/2}^{2\eta}}{d^{2\eta}}\left(\widehat{W}_{N_d}(1/2)-\sigma\right) \le y\right\} \longrightarrow \Phi\left(\frac{y}{\sqrt{\overline{K}}\,\sigma}\right) \quad \text{quand } d \to 0_+.$$

De plus, quand $d \to 0_+$, nous avons

$$\frac{\sigma^{2\eta}\,z_{1-\alpha/2}^{2\eta}}{d^{2\eta}}\left(\widehat{W}_{N_d}(1/2)-\sigma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \overline{K}\,\sigma^2\right). \tag{V.19}$$

En utilisant (V.17), (V.19) et le Théorème de Slutsky sur (V.16), nous obtenons

$$N_d^{\eta}\left(\widehat{W}_{N_d}(1/2)-\sigma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\,\overline{K}\right) \quad \text{quand } d \to 0_+,$$

(ii) La convergence (V.13) s'obtient de la même manière.

(iii) Nous démontrons maintenant (V.14). Pour $x > 1$, nous avons $0 \le \sqrt{x} - \sqrt{x-1} \le$

$1/(2\sqrt{x-1})$. Nous avons donc, pour tout $\epsilon^\top > 0$,

$$\mathrm{P}\left\{ N_d^\eta \left(\frac{\sqrt{N_d} - \sqrt{N_d - 1}}{\sqrt{K}\,\sigma} \right) \frac{d}{z_{1-\alpha/2}} > \epsilon^\top \right\} \leq \mathrm{P}\left\{ N_d^\eta \frac{d\,(N_d-1)^{-1/2}}{2\sqrt{K}\,z_{1-\alpha/2}\,\sigma} > \epsilon^\top \right\}.$$

De plus, quand $d \to 0_+$, nous avons

$$N_d^\eta \frac{d\,(N_d-1)^{-1/2}}{\sqrt{K}\,z_{1-\alpha/2}\,\sigma\,2} \sim \frac{{N_d}^{\eta-1/2}}{2\sqrt{n_d}\sqrt{K}}.$$

Par conséquent, par Théorème V.1 (3), nous avons quand $d \to 0_+$,

$$\frac{{N_d}^{\eta-1/2}}{2\sqrt{n_d}\sqrt{K}} \xrightarrow{P} 0.$$

Preuve du Théorème V.4. D'une part, comme $N_d = \min\left\{ n \geq n_0 \mid \widehat{W}_n(1/2) \leq \dfrac{\sqrt{n}\,d}{z_{1-\alpha/2}} \right\}$,

avec $d > 0$, nous avons

$$\widehat{W}_{N_d}(1/2) \leq \frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}},$$

et, par conséquent quand $d \to 0_+$,

$$\limsup P\left\{ N_d^\eta \left(\frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}} - \sigma \right) \frac{1}{\sqrt{K}\,\sigma} \leq y \right\}$$

$$\leq \limsup P\left\{ N_d^\eta \left(\widehat{W}_{N_d}(1/2) - \sigma \right) \frac{1}{\sqrt{K}\,\sigma} \leq y \right\}.$$

D'après le Lemme V.1 (i), nous avons quand $d \to 0_+$,

$$\limsup P\left\{ N_d^\eta \left(\frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}} - \sigma \right) \frac{1}{\sqrt{K}\,\sigma} \leq y \right\} \leq \Phi(y). \tag{V.20}$$

D'autre part, selon la définition de la variable d'arrêt N_d, nous avons

$$\widehat{W}_{N_d-1}(1/2) > \frac{\sqrt{N_d-1}\,d}{z_{1-\alpha/2}},$$

et alors, quand $d \to 0_+$,

$$\liminf P\left\{ N_d^\eta \left(\widehat{W}_{N_d-1}(1/2) - \sigma\right) \frac{1}{\sqrt{\overline{K}}\,\sigma} \leq y \right\}$$

$$\leq \liminf P\left\{ N_d^\eta \left(\frac{\sqrt{N_d-1}\,d}{z_{1-\alpha/2}} - \sigma\right) \frac{1}{\sqrt{\overline{K}}\,\sigma} \leq y \right\}.$$

D'après le Lemme V.1 (ii) et le Lemme V.1 (iii), nous obtenons, quand $d \to 0_+$,

$$\liminf P\left\{ N_d^\eta \left(\frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}} - \sigma\right) \frac{1}{\sqrt{\overline{K}}\,\sigma} \leq y \right\} \geq \Phi(y). \tag{V.21}$$

Par conséquent, par (V.20) et (V.21), nous avons $\forall y \in \mathbb{R}$

$$P\left\{ N_d^\eta \left(\frac{\sqrt{N_d}\,d}{z_{1-\alpha/2}} - \sigma\right) \frac{1}{\sqrt{\overline{K}}\,\sigma} \leq y \right\} \longrightarrow \Phi(y) \text{ as } d \to 0_+.$$

Comme $d\sqrt{N_d} \xrightarrow{P} z_{1-\alpha/2}\,\sigma$, la preuve est terminée.

4 Cas multivarié

Ce travail concerne un projet d'article à soumettre.

Nous considérons maintenant le vecteur des paramètres de régression $\boldsymbol{\beta} \in \mathbb{R}^p$, avec $p > 1$. D'après (II.3), nous obtenons :

$$\sqrt{n}\left(\hat{\beta}(\theta) - \beta\right) \xrightarrow[n\to+\infty]{D} \mathcal{N}_p\left(0, \boldsymbol{\Sigma}_\theta\right).$$

On en déduit que la forme quadratique

$$n\left(\hat{\beta}(\theta)-\beta\right)^{\top}\boldsymbol{\Sigma}_{\theta}^{-1}\left(\hat{\beta}(\theta)-\beta\right)$$

suit une loi du χ^2 à p degrés de liberté où $\hat{\beta}(\theta)$ désigne l'estimateur du quantile de régression de $\beta\in\mathbb{R}^p$ pour un échantillon de taille n.

La région de confiance de niveau de confiance $(1-\alpha)\%$ pour $\alpha\in[0,1]$ est donc :

$$RC_n(\alpha)=\left\{\beta\in\mathbb{R}^p : n\left(\hat{\beta}(\theta)-\beta\right)^{\top}\boldsymbol{\Sigma}_{\theta}^{-1}\left(\hat{\beta}(\theta)-\beta\right)\leq\chi^2_{p,1-\alpha}\right\}$$

où $\chi^2_{p,1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi du χ^2 à p degrés de liberté.

Il faut maintenant, comme dans le cas unidimensionnel, définir une variable d'arrêt. Pour cela, nous considérons l'ellipsoïde de confiance de niveau $1-\alpha$, $\alpha\in[0,1]$, défini par

$$n\left(\hat{\beta}(\theta)-\beta\right)^{\top}\boldsymbol{\Sigma}_{\theta}^{-1}\left(\hat{\beta}(\theta)-\beta\right)=\chi^2_{p,1-\alpha}. \tag{V.22}$$

Comme la matrice $\boldsymbol{\Sigma}_{\theta}$ est inconnue, nous l'estimons à partir des estimateurs de la densité du quantile $q(\theta)$ de type noyau et de ce fait la forme quadratique (V.22) suit une distribution T^2 de Hotelling de paramètres p et $n-1$, notée $T^2_{p,n-1}$ qui a la propriété d'être associée à la distribution de Fisher par

$$T^2_{p,n-1}=\frac{p(n-1)}{(n-p)}F_{p,n-p}.$$

Nous notons $\widehat{\boldsymbol{\Sigma}}_{\theta}$ l'estimateur de $\boldsymbol{\Sigma}_{\theta}$. Nous en déduisons :

$$\frac{n-p}{p(n-1)}n\left(\hat{\beta}(\theta)-\beta\right)^{\top}\widehat{\boldsymbol{\Sigma}}_{\theta}^{-1}\left(\hat{\beta}(\theta)-\beta\right)$$

suit une loi de Fisher de paramètres p et $(n-p)$ et par conséquent l'ellipsoïde de confiance de niveau $(1-\alpha)$, $\alpha \in [0,1]$ s'écrit :

$$RC_n(\alpha) = \left\{ \beta \in \mathbb{R}^p : n \left(\hat{\beta}(\theta) - \beta\right)^\top \widehat{\mathbf{\Sigma}}_\theta^{-1} \left(\hat{\beta}(\theta) - \beta\right) \leq \frac{p(n-1)}{(n-p)} F_{1-\alpha,p,n-p} \right\}$$

où $F_{1-\alpha,p,n-p}$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à p et $(n-p)$ degrés de liberté.

La longueur du plus grand axe vaut :

$$\frac{2}{\sqrt{\dfrac{n(n-p)}{p(n-1)} \dfrac{\Phi_{\min}\left(\widehat{\mathbf{\Sigma}}_\theta^{-1}\right)}{F_{1-\alpha,p,n-p}}}},$$

où $\Phi_{\min}(.)$ désigne la fonction donnant la valeur propre minimale. Afin d'imposer que chaque composante du vecteur $\boldsymbol{\beta}$ se trouve dans un intervalle de confiance de longueur au plus $2d$, il faut que :

$$\frac{n(n-p)}{p(n-1)} \geq \frac{F_{1-\alpha,p,n-p}}{d^2\, \Phi_{\min}\left(\widehat{\mathbf{\Sigma}}_\theta^{-1}\right)}.$$

Ainsi, la variable d'arrêt N_d s'écrit :

$$N_d = \min \left\{ n \geq n_0 : \frac{n(n-p)}{p(n-1)} \geq \frac{F_{1-\alpha,p,n-p}}{d^2\, \Phi^\star_{\min}\left(\widehat{\mathbf{\Sigma}}_\theta^{-1}\right)} \right\}, \tag{V.23}$$

où

$$\Phi^\star_{\min}\left(\widehat{\mathbf{\Sigma}}_\theta^{-1}\right) = \min\left\{ \Phi^\star_{\min}\left(\widehat{\mathbf{\Sigma}}_\theta^{-1}\right), \varepsilon'_n \right\},$$

avec $(\varepsilon'_n)_{n\geq 1}$, une suite de nombres réels positifs tendant vers l'infini. Ainsi,

$$N_d = \min \left\{ n \geq n_0 : \Phi^\star_{\min}\left(\widehat{\mathbf{\Sigma}}_\theta^{-1}\right) \geq \frac{F_{1-\alpha,p,n-p}}{d^2} \frac{p(n-1)}{n(n-p)} \right\}.$$

est bien une variable d'arrêt.

5 Application en chémoinformatique

La détermination des modes d'actions des composés chimiques est une problématique importante. Cette méthode sera étudiée dans le cas univarié et multivarié par simulation et pour l'étude des modes d'actions non spécifiques en chémoinformatique. En particulier, dans le cas d'un mode d'action non spécifique, cette procédure séquentielle nous permettra en utilisant différents quantiles de déterminer avec un nombre minimum de composés si

— la narcose est non polaire ou "Baseline narcosis",

— la narcose est polaire ou "Polar narcosis",

— une réactivité non spécifique ou "Reactive compounds" est présente.

.

Conclusion générale et perspectives

L'objectif principal de cette thèse est de modéliser la relation entre la structure chimique et l'activité écotoxicologique sur les algues *P. Subcapitata* pour un grand nombre de composés chimiques. Dans ce cadre, nous avons envisagé des méthodes statistiques robustes qui permettent de mieux gérer la présence de valeurs atypiques. En première application, nous avons mis en place un certain nombre de modèles permettant de prédire l'activité toxicologique afin d'aider à la prise de décision.

Dans ce but, nous avons exploré le domaine de la statistique robuste autour des méthodes quantiles qui n'est à notre connaissance pas utilisé en chémoinformatique. Nous avons montré comment ces méthodes peuvent améliorer la précision des modèles et être utilisées avantageusement sur les données traitées. Nous avons ainsi jeté les prémices des modèles robustes pour un MOA non spécifique, modèles intégrant un nombre et une diversité de composés remarquables allant des molécules chimiques simples jusqu'aux médicaments. Pour ce faire, nous avons principalement travaillé sur les modèles de régression quantile. Nous dressons ci-après un bilan des résultats obtenus dans cette thèse.

Dans beaucoup d'applications et en particulier en chimie, les données sont contaminées par des valeurs atypiques qui proviennent d'erreurs dues à la présence de composés ayant un comportement différent par rapport à la majorité des composés ou de tout autre cause, tout aussi

triviale qu'une erreur d'enregistrement ou de lecture. Face à ce type de données, nous utilisons afin de déterminer le mode d'action de composés chimiques, le rapport de toxicité et des modèles robustes de type quantile de régression linéaire et quantile de régression SVM.

Nous considérons tout d'abord 401 composés chimiques pour lesquels nous avons les valeurs de EC_{50} ainsi que des informations sur la structure des composés chimiques. Nous commençons par déterminer le mode d'action des composés chimiques afin de prédire et écarter les composés ayant un mode d'action spécifique. Ensuite, en considérant une régression médiane pour estimer les TR, un total de 336 composés chimiques est considéré comme n'ayant pas de modes d'action spécifique. On utilise ensuite une classification SVM afin d'obtenir une prédiction des modes d'action des molécules sur l'ensemble des descripteurs. Nous obtenons alors 368 composés chimiques n'ayant pas de modes d'action spécifique. Nous proposons une méthode de sélection des variables et nos résultats montrent la bonne performance du modèle QSVMR sur ces données.

À partir des résultats obtenus, il est apparu que la solubilité moléculaire, le log(P) et la polarisabilité sont les 3 descripteurs fondamentaux associés à un mode d'action de type narcose. À partir de ces descripteurs, nous avons construit un modèle permettant de discriminer des molécules atypiques par une classification SVM à une classe afin de prédire le niveau de toxicité de 36 médicaments testés au CERMN. Un intérêt de cette analyse est de pouvoir identifier les molécules se trouvant dans un espace dans lequel nous avons peu de composés. On peut souligner que l'utilisation de ce domaine de validité a conduit à la diminution de l'erreur quadratique moyenne de nos modèles.

Les modèles quantiles obtenus ont ensuite été testés sur un ensemble de médicaments. Ceux-ci ont été sélectionnés en considérant leur présence dans l'environnement (concentrations environnementales prédites). Nos modèles QSAR donnent une bonne description de l'écotoxicité

potentielle des composés de la classe 1 obtenue par la classification BDDCS qui sont considérés comme étant très solubles dans l'eau. Pour les composés de la classe 2, caractérisés par une faible solubilité dans l'eau, une légère surestimation de l'écotoxicité des composés est obtenue à partir de nos modèles. Pour la plupart des médicaments des classes 3 et 4, caractérisés par leur faible perméabilité membranaire, aucune toxicité n'est observée et une surestimation est par conséquent mesurée. Parmi l'ensemble des médicaments analysés, notre étude a mis clairement en évidence le côté préoccupant pour l'environnement d'un antibiotique la clarithromycine.

Le nombre de composés chimiques synthétisés est estimé autour de 10 millions. Les composés synthétisés à plus d'une tonne/an représentent environ 100000 composés. Il est par conséquent essentiel de pouvoir estimer l'impact des produits chimiques sur la santé humaine et l'environnement le plus rapidement possible. Les modèles QSAR ont par conséquent un intérêt majeur pour cette estimation. À ce niveau, le programme européen REACH va nous permettre de recueillir les données pour le renforcement et la définition de nos modèles QSAR.

Un autre aspect qui nous a motivé est le nombre et le mode d'acquisition des données qui est en général réalisé de manière séquentielle par les chimistes. Les méthodes de statistiques séquentielles sont donc tout à fait adaptées à ce type de situation. La définition d'une règle d'arrêt nous indique si nous pouvons arrêter l'expérience après les n premières mesures ou si nous devons continuer avec une nouvelle observation ou un nouveau lot d'observations. Nous proposons une procédure séquentielle pour un paramètre réel du vecteur de régression QSAR et ensuite, nous généralisons les résultats au cas multidimensionnel. Pour le MOA non spécifique, cette procédure séquentielle nous permettra de déterminer avec un nombre minimum de composés à partir de différents niveaux de quantiles si

— la narcose est non polaire ou "Baseline narcosis",

— la narcose est polaire ou "Polar narcosis",

— une réactivité non spécifique ou “Reactive compounds” est présente.

Références bibliographiques

[1] Reach. http ://ec.europa.eu/environment/chemicals/reach/reach_intro.htm. 1, 8, 47

[2] Ecvam. https ://eurl-ecvam.jrc.ec.europa.eu/. 2

[3] Qsartoolbox. www.qsartoolbox.org. 2

[4] Feixiong Cheng, Weihua Li, Yadi Zhou, Jie Li, Jie Shen, Philip W. Lee, and Yun Tang. *Prediction of human genes and diseases targeted by xenobiotics using predictive toxicogenomic-derived models (ptdms).* Mol. BioSyst. **9**, 1316–1325 (2013). 2

[5] Stefan Trapp and Stefan Schwartz. *Proposals to overcome limitations in the eu chemical risk assessment scheme.* Chemosphere **41**(7), 965–971 (2000). 2

[6] aniline germany. http ://echa.europa.eu/documents/10162/462b7066-c639-4883-b384-3daf4ec88ded. 2

[7] Oecd_ligne_directrice. http ://www.oecd.org/chemicalsafety/testing/ oecdguidelinesforthetestingofchemicals.htm. 2

[8] Stefan Scholz, Erika Sela, Ludek Blaha, Thomas Braunbeck, Malyka Galay-Burgos, Mauricio Garcia-Franco, Joaquin Guinea, Nils Kluever,

Kristin Schirmer, Katrin Tanneberger, et al. *A european perspective on alternatives to animal testing for environmental hazard identification and risk assessment.* Regulatory Toxicology and Pharmacology **67**(3), 506–530 (2013). 3

[9] Anr pharmecotox. http ://www.agence-nationale-recherche.fr/ ?Projet=ANR-10-CESA-0013. 4

[10] James Blake. *On the connection between chemical constitution, and physiological action.* Nature **34**, 594–595 (1886). 7

[11] Spencer M Free and James W Wilson. *A mathematical contribution to structure-activity studies.* Journal of Medicinal Chemistry **7**(4), 395–399 (1964). 7

[12] Corwin Hansch, Peyton P Maloney, Toshio Fujita, and Robert M Muir. *Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients.* Nature Publishing Group (1962). 7

[13] Steven P Bradbury, Christine L Russom, Patricia K Schmieder, Terry W Schultz, Robert Diderich, and Charles M Auer. *Advancing computational toxicology in a regulatory setting : A selected review of the accomplishments of gilman d. veith (1944–2013).* Applied In Vitro Toxicology **1**(1), 16–25 (2015). 8

[14] ministere. http ://www.developpement-durable.gouv.fr/-Gestion-des-produits-chimiques-.html. 8

[15] Echa. http ://echa.europa.eu/. 8, 9

[16] Organisation for Economic Co-operation and Development. *Test No. 201 : Freshwater Alga and Cyanobacteria, Growth Inhibition Test.* OECD Publishing (2011). 8

[17] Ecotox japan. https ://www.env.go.jp/chemi/sesaku/02e.pdf. 9, 51, 52

[18] Ecb. http ://esis.jrc.ec.europa.eu/index.php ?PGM=hpv. 9, 51, 52

[19] CL Russom, EB Anderson, BE Greenwood, and A Pilli. *Aster : an integration of the aquire data base and the qsar system for use in ecological risk assessments.* Sci Total Environ **109-110**, 667–670 (1991). 10, 51, 52

[20] JC Faucon, R Bureau, J Faisant, F Briens, and S Rault. *Prediction of the fish acute toxicity from heterogeneous data coming from notification files.* Chemosphere **38**, 3261–3276 (1999). 10, 47, 51, 52

[21] Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics.* John Wiley & Sons (2009). 10

[22] Corwin Hansch, Albert Leo, David Hoekman, and Albert Leo. *Exploring Qsar.* American Chemical Society Washington, DC (1995). 11

[23] Accelrys. *Pipeline pilot ; 10188 telesis court.* San Diego : SciTegic **Suite 100**, Inc. . 7.5 edn., (2009). 11, 13, 53

[24] John P Perdew and Yue Wang. *High precision sampling for brillouin-zone integration in metals.* Phys. Rev. B **45**(23), 13 (1992). 12

[25] IV Tetko, VY Tanchuk, TN Kasheva, and AE Villa. *Estimation of aqueous solubility of chemical compounds using e-state indices.* J Chem Inf Comput Sci **41**, 1488–1493 (2001). 12, 53, 79

[26] RH Rohrbaugh and PC Jurs. *Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships.* Analytica Chimica Acta **199**, 99–109 (1987). 13, 53

[27] B Ehresmann, MJ de Groot, A Alex, and T Clark. *New molecular descriptors based on local properties at the molecular surface and a boiling-point model derived from them.* J Chem Inf Comput Sci pages 658–668 (2004). 13, 53

[28] L Mavridis, BD Hudson, and DW Ritchie. *Toward high throughput 3d virtual screening using spherical harmonic surface representations.* J Chem Inf Model **47**, 1787–1796 (2007). 13, 68, 79

[29] Jean-Jacques Droesbeke and Gilbert Saporta. *Approches non paramétriques en régression.* Editions Technip (2011). 14

[30] Elizbar A Nadaraya. *On estimating regression.* Theory of Probability & Its Applications **9**(1), 141–142 (1964). 15, 21

[31] Geoffrey S Watson. *Smooth regression analysis.* Sankhyā : The Indian Journal of Statistics, Series A pages 359–372 (1964). 15, 21

[32] Hongying Du, Jie Wang, Zhide Hu, and Xiaojun Yao. *Quantitative structure-retention relationship study of the constituents of saffron aroma in spme-gc–ms based on the projection pursuit regression method.* Talanta **77**(1), 360–365 (2008). 16

[33] Hongying Du, June Watzl, Jie Wang, Xiaoyun Zhang, Xiaojun Yao, and Zhide Hu. *Prediction of retention indices of drugs based on immobilized artificial membrane chromatography using projection pursuit regression and local lazy regression.* Journal of separation science **31**(12), 2325–2333 (2008). 16

[34] Hongying Du, Xiaoyun Zhang, Jie Wang, Xiaojun Yao, and Zhide Hu. *Novel approaches to predict the retention of histidine-containing peptides in immobilized metal-affinity chromatography.* Proteomics **8**(11), 2185–2195 (2008). 16, 21

[35] Alan R Katritzky, Liliana Pacureanu, Dimitar Dobchev, and Mati Karelson. *Qspr modeling of hyperpolarizabilities.* Journal of molecular modeling **13**(9), 951–963 (2007). 16

[36] Cristina Ventura, Diogo ARS Latino, and Filomena Martins. *Comparison of multiple linear regressions and neural networks based qsar models for the design of new antitubercular compounds.* European journal of medicinal chemistry **70**, 831–845 (2013). 16

[37] Anwar Rafique Shaikh, Steffi Ignatius Gonsalves, Amruta Nikam, Sanjay J Kshirsagar, and Yogita Thombare. *Predicting pyrazinecarboxamides derivatives as an herbicidal agent : 3d qsar by knn-mfa and multiple linear regression approach.* World Applied Sciences Journal **33**(6), 980–989 (2015). 16

[38] Ashwini H Pagare, Rani S Kankate, and Anwar R Shaikh. *2d and 3d qsar using knn-mfa method of the novel 3, 4-dihydropyrimidin-2 (1h)-one urea derivatives of n-aryl urea as an antifungal agents.* Current Pharma Research **5**(2), 1473 (2015). 16

[39] Mansour Arab Chamjangali, Ghadamali Bagherian, Motahhareh Ashrafi, and Amir Hossein Amin. *Journal of applied chemistry prediction of the anti-hiv activities of pett analogs as non-nucleoside hiv-1 reverse transcriptase inhibitors by linear and non-linear qsar models.* Journal of Applied Chemistry Vol **9**(32) (2015). 16

[40] Eslam Pourbasheer, Abolghasem Beheshti, Saadat Vahdani, Mehdi Nekoei, Mohammad Danandeh, Maryam Abbasghorbani, and Mohammad Reza Ganjali. *Simple qspr modeling for prediction of the gc retention indices of essential oil compounds.* Journal of Essential Oil Bearing Plants **18**(6), 1298–1309 (2015). 16

[41] Ardeshir Khazaei, Negin Sarmasti, Jaber Yousefi Seyf, Zahra Rostami, and Mohammad Ali Zolfigol. *Qsar study of the non-peptidic inhibitors of procollagen c-proteinase based on multiple linear regression, principle component regression, and partial least squares.* Arabian Journal of Chemistry (2015). 16

[42] Binbin Xia, Kunping Liu, Zhiguo Gong, Bo Zheng, Xiaoyun Zhang, and Botao Fan. *Rapid toxicity prediction of organic chemicals to chlorella vulgaris using quantitative structure–activity relationships methods.* Ecotoxicology and environmental safety **72**(3), 787–794 (2009). 16, 26, 47

[43] Yuxi Lu, Rongchao Li, Lili Tang, and Feng Luan. *Prediction of muti-anticancer activity of curcumin-related compounds by qsar approach.* Journal of Computational Science & Engineering **18**, 613–621 (2015). 16

[44] Jenny Balfer and Jürgen Bajorath. *Systematic artifacts in support vector regression-based compound potency prediction revealed by statistical and activity landscape analysis.* PloS one **10**(3), e0119301 (2015). 16

[45] Joseph Rebehmed, Florent Barbault, Cátia Teixeira, and François Maurel. *2d and 3d qsar studies of diarylpyrimidine hiv-1 reverse transcriptase inhibitors.* Journal of computer-aided molecular design **22**(11), 831–841 (2008). 16

[46] Tao Wang, Hongzong Si, Pingping Chen, Kejun Zhang, and Xiaojun Yao. *Qsar models for the dermal penetration of polycyclic aromatic hydrocarbons based on gene expression programming.* QSAR & Combinatorial Science **27**(7), 913–921 (2008). 16

[47] Luana Janaína de Campos and Eduardo Borges de Melo. *Modeling structure–activity relationships of prodiginines with antimalarial activity using ga/mlr and ops/pls.* Journal of Molecular Graphics and Modelling **54**, 19–31 (2014). 16

[48] Farhad Gharagheizi. *Qspr studies for solubility parameter by means of genetic algorithm-based multivariate linear regression and generalized regression neural network*. QSAR & Combinatorial Science **27**(2), 165–170 (2008). 16

[49] AK Saxena and P Prathipati. *Comparison of mlr, pls and ga-mlr in qsar analysis**. SAR and QSAR in Environmental Research **14**(5-6), 433–445 (2003). 16

[50] N Sukumar, Ganesh Prabhu, and Pinaki Saha. Applications of genetic algorithms in qsar/qspr modeling. In *Applications of Metaheuristics in Process Engineering*, pages 315–324. Springer (2014). 16

[51] Jacek J Fisz. *Combined genetic algorithm and multiple linear regression (ga-mlr) optimizer : Application to multi-exponential fluorescence decay surface*. The Journal of Physical Chemistry A **110**(48), 12977–12985 (2006). 16

[52] Herman Wold et al. *Estimation of principal components and related models by iterative least squares*. Multivariate analysis **1**, 391–420 (1966). 17

[53] Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, III. *The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses*. SIAM Journal on Scientific and Statistical Computing **5**(3), 735–743 (1984). 17

[54] Svante Wold, Henrik Antti, Fredrik Lindgren, and Jerker Öhman. *Orthogonal signal correction of near-infrared spectra*. Chemometrics and Intelligent Laboratory Systems **44**(1), 175–185 (1998). 17

[55] Nora Priolo, Cecilia M Arribére, Néstor Caffini, Sonia Barberis, Rodolfo Nieto Vázquez, and Juan M Luco. *Isolation and purification of cysteine peptidases from the latex of araujia hortorum fruits : Study of their esterase activities*

using partial least-squares (pls) modeling. Journal of Molecular Catalysis B : Enzymatic **15**(4), 177–189 (2001). 18

[56] Yang Shan-Bin, Xia Zhi-Ning, Shu Mao, Mei Hu, Lue Feng-Lin, Zhang Mei, Wu Yu-Qian, and Li Zhi-Liang. *Vhseh descriptors for the development of qsams of peptides*. CHEMICAL JOURNAL OF CHINESE UNIVERSITIES-CHINESE **29**(11), 2213–2217 (2008). 18

[57] Liang Gui-Zhao, Mei Hu, Zhou Yuan, Yang Shan-Bin, Wu Shi-Rong, and Li Zhi-Liang. *Using szott descriptors for the development of qsams of peptides*. CHEMICAL JOURNAL OF CHINESE UNIVERSITIES-CHINESE **27**(10), 1900–1902 (2006). 18

[58] Hai-Xia Long, Yuan-Qiang Wang, Yong Lin, and Zhi-Hua Lin. *Qsar study on ace inhibitors by using osc-pls algorithm*. Journal of the Chinese Chemical Society **57**(3A), 417–422 (2010). 18

[59] Reyhaneh Jahangiri, Somaieh Soltani, and Abolfazl Barzegar. *A review of qsar studies to predict activity of ace peptide inhibitors*. Pharmaceutical Sciences **20**(3), 122 (2014). 18

[60] Fatemeh Bagheban Shahri, Ali Niazi, and Ahmad Akrami. Application of wavelet and genetic algorithms for qsar study on 5-lipoxygenase inhibitors and design new compounds, (2015). 18

[61] WJ Dunn III and D Rogers. Genetic partial least squares in qsar, (1996). 18

[62] Tarnvir Sammi, Om Silakari, and Muttineni Ravikumar. *Three-dimensional quantitative structure-activity relationship (3d-qsar) studies of various benzodiazepine*

analogues of γ-secretase inhibitors. Journal of molecular modeling **15**(4), 343–348 (2009). 18

[63] Zu-Guang Li, Ke-Xian Chen, Hai-Ying Xie, and Jian-Rong Gao. *Quantitative structure–activity relationship analysis of some thiourea derivatives with activities against hiv-1 (iiib).* QSAR & Combinatorial Science **28**(1), 89–97 (2009). 18

[64] Matthew N Davies, Channa K Hattotuwagama, David S Moss, Michael GB Drew, and Darren R Flower. *Statistical deconvolution of enthalpic energetic contributions to mhc-peptide binding affinity.* BMC structural biology **6**(1), 5 (2006). 18

[65] Aditi Singh, Sukriti Goyal, Salma Jamal, Bala Subramani, Mriganko Das, Nikita Admane, and Abhinav Grover. *Computational identification of novel piperidine derivatives as potential hdm2 inhibitors designed by fragment-based qsar, molecular docking and molecular dynamics simulations.* Structural Chemistry pages 1–11 (2015). 18

[66] Kunal Roy and J Thomas Leonard. *Topological qsar modeling of cytotoxicity data of anti-hiv 5-phenyl-1-phenylamino-imidazole derivatives using gfa, g/pls, fa and pcra techniques.* Indian J. Chem. Sect. A-Inorg. Bio-Inorg. Phys. Theor. Anal. Chem **45**, 126–137 (2006). 18

[67] J Thomas Leonard and Kunal Roy. *Comparative qsar modeling of ccr5 receptor binding affinity of substituted 1-(3, 3-diphenylpropyl)-piperidinyl amides and ureas.* Bioorganic & medicinal chemistry letters **16**(17), 4467–4474 (2006). 18

[68] Asim Sattwa Mandal and Kunal Roy. *Predictive qsar modeling of hiv reverse transcriptase inhibitor tibo derivatives.* European journal of medicinal chemistry **44**(4), 1509–1524 (2009). 18

[69] Kunal Roy and Gopinath Ghosh. *Qstr with extended topochemical atom (eta) indices 8. a qsar for the inhibition of substituted phenols on germination rate of cucumis sativus using chemometric tools.* QSAR & Combinatorial Science **25**(10), 846–859 (2006). 18

[70] Nilanjan Adhikari, Milan Kumar Maiti, and Tarun Jha. *Predictive comparative qsar modeling of 4-pyridones as potent antimalarials.* Internet Electron J Mol Des **9**, 1–19 (2010). 18

[71] Nasser Goudarzi, Mohammad Goodarzi, and Tao Chen. *Qsar prediction of hiv inhibition activity of styrylquinoline derivatives by genetic algorithm coupled with multiple linear regressions.* Medicinal Chemistry Research **21**(4), 437–443 (2012). 18

[72] Ker-Chau Li. *Sliced inverse regression for dimension reduction.* Journal of the American Statistical Association **86**(414), 316–327 (1991). 18

[73] Y Aragon and Jerome Saracco. Sliced inverse regression (sir) : an appraisal of small sample alternatives to slicing, (1996). 19

[74] Jérôme Saracco. *Asymptotics for pooled marginal slicing estimator based on sirα approach.* Journal of multivariate Analysis **96**(1), 117–135 (2005). 19

[75] Raphaël Coudret, Stephane Girard, and Jerome Saracco. *A new sliced inverse regression method for multivariate response.* Computational Statistics & Data Analysis **77**, 285–299 (2014). 19

[76] Hong Yin, Yizeng Liang, and Qinnan Hu. *Boiling points predictions study via dimension reduction methods : Sir, pcr and plsr.* Journal of Data Science **1**(4), 461–480 (2003). 19

[77] Kai-Tai Fang, Hong Yin, and Yi-Zeng Liang. *New approach by kriging models to problems in qsar.* Journal of chemical information and computer sciences **44**(6), 2106–2113 (2004). 19

[78] Jessica J Kraker, Douglas M Hawkins, Subhash C Basak, Ramanathan Natarajan, and Denise Mills. *Quantitative structure–activity relationship (qsar) modeling of juvenile hormone activity : Comparison of validation procedures.* Chemometrics and intelligent laboratory systems **87**(1), 33–42 (2007). 20

[79] Abdo M Al-Fakih, Madzlan Aziz, Hassan H Abdallah, Zakariya Y Algamal, Muhammad H Lee, and Hasmerya Maarof. *High dimensional qsar study of mild steel corrosion inhibition in acidic medium by furan derivatives.* Int. J. Electrochem. Sci **10**, 3568–3583 (2015). 20

[80] Zakariya Yahya Algamal, Muhammad Hisyam Lee, Abdo M Al-Fakih, and Madzlan Aziz. *High-dimensional qsar prediction of anticancer potency of imidazo [4, 5-b] pyridine derivatives using adjusted adaptive lasso.* Journal of Chemometrics **29**(10), 547–556 (2015). 20

[81] Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. *Choosing feature selection and learning algorithms in qsar.* Journal of chemical information and modeling **54**(3), 837–843 (2014). 20

[82] Veli-Matti Taavitsainen. *Rational function ridge regression in kinetic modeling : A case study.* Chemometrics and Intelligent Laboratory Systems **120**, 136–141 (2013). 20

[83] Subhash C Basak and Subhabrata Majumdar. *Editorial : The importance of rigorous statistical practice in the current landscape of qsar modelling.* Current computer-aided drug design **11**(1), 2–4 (2015). 20

[84] Jerome H Friedman and Werner Stuetzle. *Projection pursuit regression.* Journal of the American statistical Association **76**(376), 817–823 (1981). 20, 21

[85] Yongna Yuan, Ruisheng Zhang, Rongjing Hu, and Xiaofang Ruan. *Prediction of ccr5 receptor binding affinity of substituted 1-(3, 3-diphenylpropyl)-piperidinyl amides and ureas based on the heuristic method, support vector machine and projection pursuit regression.* European journal of medicinal chemistry **44**(1), 25–34 (2009). 21

[86] Hongying Du, Jie Wang, Zhide Hu, Xiaojun Yao, and Xiaoyun Zhang. *Prediction of fungicidal activities of rice blast disease based on least-squares support vector machines and project pursuit regression.* Journal of agricultural and food chemistry **56**(22), 10785–10792 (2008). 21

[87] Yongna Yuan, Ruisheng Zhang, and Rongjing Hu. *Prediction of photolysis of pcdd/fs adsorbed to spruce [picea abies (l.) karst.] needle surfaces under sunlight irradiation based on projection pursuit regression.* QSAR and Combinatorial Science **28**(2), 155–162 (2009). 21

[88] Joachim Altschuh, Dieter Lenoir, Florian Rehfeldt, and Rainer Bruggemann. *Applicability domain of nonlinear property-property relationships–example : Estimation of vapour pressure.* MATCH-COMMUNICATIONS IN MATHEMATICAL AND IN COMPUTER CHEMISTRY **73**(2), 303–326 (2015). 21

[89] Anqiang Huang, Kinkeung Lai, Yinhua Li, and Shouyang Wang. *Forecasting container throughput of qingdao port with a hybrid model.* Journal of Systems Science and Complexity **28**(1), 105–121 (2015). 21

[90] Bernard W Silverman. *Density estimation for statistics and data analysis.* CRC press (1986). 22

[91] M Duflo. *Random iterative methods*. Appl. Math **34** (1997). 22

[92] Wolfgang Hardle and James Stephen Marron. *Optimal bandwidth selection in nonparametric regression function estimation*. The Annals of Statistics pages 1465–1481 (1985). 24

[93] Wolfgang Hardle, JS Marron, and MP Wand. *Bandwidth choice for density derivatives*. Journal of the Royal Statistical Society. Series B (Methodological) pages 223–232 (1990). 24

[94] T John McNeany and Jonathan D Hirst. *Inhibition of the tyrosine kinase, syk, analyzed by stepwise nonparametric regression*. Journal of chemical information and modeling **45**(3), 768–776 (2005). 24

[95] Cleo Tebby and Enrico Mombelli. *A kernel-based method for assessing uncertainty on individual qsar predictions*. Molecular Informatics **31**(10), 741–751 (2012). 24

[96] Peter P Mager. *Subset selection and docking of human p2x7 inhibitors*. Current Computer-Aided Drug Design **3**(4), 248–253 (2007). 24

[97] Michael Reutlinger, Wolfgang Guba, Rainer E Martin, Alexander I Alanine, Torsten Hoffmann, Alexander Klenner, Jan A Hiss, Petra Schneider, and Gisbert Schneider. *Neighborhood-preserving visualization of adaptive structure–activity landscapes : Application to drug discovery*. Angewandte Chemie International Edition **50**(49), 11633–11636 (2011). 24

[98] Richard A Lewis and David Wood. *Modern 2d qsar for drug discovery*. Wiley Interdisciplinary Reviews : Computational Molecular Science **4**(6), 505–522 (2014). 24

[99] Matthias Rupp, Matthias R Bauer, Rainer Wilcken, Andreas Lange, Michael Reutlinger, Frank M Boeckler, and Gisbert Schneider. *Machine learning estimates of natural product conformational energies.* PLoS Comput Biol **10**(1), e1003400 (2014). 24

[100] Pere Constans and Jonathan D Hirst. *Nonparametric regression applied to quantitative structure-activity relationships.* Journal of chemical information and computer sciences **40**(2), 452–459 (2000). 24

[101] Jure Zupan and Johann Gasteiger. *Neural networks in chemistry and drug design.* John Wiley & Sons, Inc. (1999). 25

[102] John C Dearden and Philip H Rowe. *Use of artificial neural networks in the qsar prediction of physicochemical properties and toxicities for reach legislation.* Artificial Neural Networks pages 65–88 (2015). 25

[103] WJ Lü, YL Chen, WP Ma, XY Zhang, F Luan, MC Liu, XG Chen, and ZD Hu. *Qsar study of neuraminidase inhibitors based on heuristic method and radial basis function network.* European journal of medicinal chemistry **43**(3), 569–576 (2008). 26

[104] Manish K Gupta, Swati Gupta, and Ravindra K Rawal. *Impact of artificial neural networks in qsar and computational modeling.* Artificial Neural Network for Drug Design, Delivery and Disposition page 153 (2015). 26

[105] Andrea Gissi, Anna Lombardo, Alessandra Roncaglioni, Domenico Gadaleta, Giuseppe Felice Mangiatordi, Orazio Nicolotti, and Emilio Benfenati. *Evaluation and comparison of benchmark qsar models to predict a relevant reach*

endpoint : The bioconcentration factor (bcf). Environmental research **137**, 398–409 (2015). 26

[106] Feng Luan, Weiping Ma, Xiaoyun Zhang, Haixia Zhang, Mancan Liu, Zhide Hu, and BT Fan. *Quantitative structure-activity relationship models for prediction of sensory irritants (logrd 50) of volatile organic chemicals.* Chemosphere **63**(7), 1142–1153 (2006). 26

[107] Weiping Ma, Feng Luan, Haixia Zhang, Xiaoyun Zhang, Mancang Liu, Zhide Hu, and Botao Fan. *Accurate quantitative structure–property relationship model of mobilities of peptides in capillary zone electrophoresis.* Analyst **131**(11), 1254–1260 (2006). 26

[108] Donald F Specht. *A general regression neural network.* Neural Networks, IEEE Transactions on **2**(6), 568–576 (1991). 26

[109] Maciej Szaleniec, Ryszard Tadeusiewicz, and Małgorzata Witko. *How to select an optimal neural model of chemical reactivity ?* Neurocomputing **72**(1), 241–256 (2008). 26

[110] Farhad Gharagheizi, Behnam Tirandazi, and Reza Barzin. *Estimation of aniline point temperature of pure hydrocarbons : A quantitative structure- property relationship approach.* Industrial & Engineering Chemistry Research **48**(3), 1678–1682 (2008). 26

[111] Hong-ze Li, Sen Guo, Chun-jie Li, and Jing-qi Sun. *A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm.* Knowledge-Based Systems **37**, 378–387 (2013). 26

[112] V Vapnik. *Statistical learning theory*. Wiley-Interscience, New York, (1998). 26, 32, 63, 95

[113] B Boser, I Guyon, and V Vapnik. *A training algorithm for optimal margin classifiers*. Fifth Annual Workshop on Computational Learning Theory pages 144–152 (1992). 30

[114] Antreas Afantitis, Georgia Melagraki, Haralambos Sarimveis, Panayiotis A Koutentis, Olga Igglessi-Markopoulou, and George Kollias. *A combined ls-svm and mlr qsar workflow for predicting the inhibition of cxcr3 receptor by quinazolinone analogs*. Molecular diversity **14**(2), 225–235 (2010). 33

[115] Mohsen Shahlaei, Razieh Sabet, Maryam Bahman Ziari, Behzad Moeinifard, Afshin Fassihi, and Reza Karbakhsh. *Qsar study of anthranilic acid sulfonamides as inhibitors of methionine aminopeptidase-2 using ls-svm and grnn based on principal components*. European journal of medicinal chemistry **45**(10), 4499–4508 (2010). 33

[116] Noslen Hernández, Rudolf Kiralj, Márcia MC Ferreira, and Isneri Talavera. *Critical comparative analysis, validation and interpretation of svm and pls regression models in a qsar study on hiv-1 protease inhibitors*. Chemometrics and Intelligent Laboratory Systems **98**(1), 65–77 (2009). 33

[117] F Bagheban-Shahri, A Niazi, and AHMAD Akrami. *Quantitative structure activity relationship study of inhibitory activities of 5-lipoxygenase and design new compounds by different chemometrics methods*. Iranian Journal of Mathematical Chemistry **7**(1), 47–59 (2016). 33

[118] A Chamkalani. *Application of ls-svm classifier to determine stability state of asphaltene*

in oilfields by utilizing sara fractions. Petroleum Science and Technology **33**(1), 31–38 (2015). 33

[119] Liangliang Qiao and Qijuan Chen. *Forecasting models for hydropower unit stability using ls-svm*. Mathematical Problems in Engineering **2015** (2015). 33

[120] Johan AK Suykens and Joos Vandewalle. *Least squares support vector machine classifiers*. Neural processing letters **9**(3), 293–300 (1999). 33

[121] Eslam Pourbasheer, Reza Aalizadeh, and Mohammad Reza Ganjali. *Qsar study of ck2 inhibitors by ga-mlr and ga-svm methods*. Arabian Journal of Chemistry (2015). 33

[122] Omar Deeb and Mohammad Goodarzi. *Exploring qsars for inhibitory activity of non-peptide hiv-1 protease inhibitors by ga-pls and ga-svm*. Chemical biology & drug design **75**(5), 506–514 (2010). 33

[123] Eslam Pourbasheer, Reza Aalizadeh, Mohammad Reza Ganjali, and Parviz Norouzi. *Qsar study of $\alpha 1\beta 4$ integrin inhibitors by ga-mlr and ga-svm methods*. Structural Chemistry **25**(1), 355–370 (2014). 33

[124] Eslam Pourbasheer, Siavash Riahi, Mohammad Reza Ganjali, and Parviz Norouzi. *Application of genetic algorithm-support vector machine (ga-svm) for prediction of bk-channels activity*. European journal of medicinal chemistry **44**(12), 5023–5028 (2009). 33

[125] Leo Breiman. *Random forests*. Machine learning **45**(1), 5–32 (2001). 34

[126] Pavel G Polishchuk, Eugene N Muratov, Anatoly G Artemenko, Oleg G Kolumbin, Nail N Muratov, and Victor E Kuz'min. *Application of random*

forest approach to qsar prediction of aquatic toxicity. Journal of chemical information and modeling **49**(11), 2481–2488 (2009). 34

[127] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. *Random forest : a classification and regression tool for compound classification and qsar modeling.* Journal of chemical information and computer sciences **43**(6), 1947–1958 (2003). 34

[128] Vladimir Svetnik, Andy Liaw, Christopher Tong, and Ting Wang. Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In *Multiple Classifier Systems*, pages 334–343. Springer (2004). 34

[129] Robert P Sheridan. *Three useful dimensions for domain applicability in qsar models using random forest.* Journal of chemical information and modeling **52**(3), 814–823 (2012). 34

[130] Bin Chen, Robert P Sheridan, Viktor Hornak, and Johannes H Voigt. *Comparison of random forest and pipeline pilot naive bayes in prospective qsar predictions.* Journal of chemical information and modeling **52**(3), 792–803 (2012). 34

[131] Søren H Welling, Line KH Clemmensen, Stephen T Buckley, Lars Hovgaard, Per B Brockhoff, and Hanne HF Refsgaard. *In silico modelling of permeation enhancement potency in caco-2 monolayers based on molecular descriptors and random forest.* European Journal of Pharmaceutics and Biopharmaceutics **94**, 152–159 (2015). 34

[132] Harinder Singh, Sandeep Singh, Deepak Singla, Subhash M Agarwal, and

Gajendra PS Raghava. *Qsar based model for discriminating egfr inhibitors and non-inhibitors using random forest.* Biology direct **10**(10) (2015). 34

[133] Andy Liaw and Vladimir Svetnik. *Qsar modeling : prediction of biological activity from chemical structure.* Statistical Methods for Evaluating Safety in Medical Product Development pages 66–83 (2015). 34

[134] Peter J Huber. *The 1972 wald lecture robust statistics : A review.* The Annals of Mathematical Statistics pages 1041–1067 (1972). 35

[135] Peter J Huber et al. *Robust estimation of a location parameter.* The Annals of Mathematical Statistics **35**(1), 73–101 (1964). 35

[136] David F Andrews. *A robust method for multiple linear regression.* Technometrics **16**(4), 523–531 (1974). 36

[137] Peter J Bickel. *One-step huber estimates in the linear model.* Journal of the American Statistical Association **70**(350), 428–434 (1975). 36

[138] Peter J Huber. *Robust regression : asymptotics, conjectures and monte carlo.* The Annals of Statistics pages 799–821 (1973). 36

[139] Steven C Huber. *Interspecific variation in activity and regulation of leaf sucrose phosphate synthetase.* Zeitschrift für Pflanzenphysiologie **102**(5), 443–450 (1981). 36

[140] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Linear models : Robust estimation.* Robust Statistics : The Approach Based on Influence Functions pages 307–341 (1986). 36

[141] Jana Jurečková and Pranab Kumar Sen. *Robust statistical procedures : asymptotics and interrelations.* John Wiley & Sons (1996). 36, 100

[142] Stephen Portnoy, Roger Koenker, et al. *The gaussian hare and the laplacian tortoise : computability of squared-error versus absolute-error estimators.* Statistical Science **12**(4), 279–300 (1997). 39

[143] AH Welsh. *One-step l-estimators for the linear model.* The Annals of Statistics pages 626–641 (1987). 40

[144] R Koenker and G Basset. *Regression quantiles.* Econometrica **46**, 33–50 (1978). 40, 50, 54, 55

[145] R Koenker. *Quantile regression.* Cambridge University Press, New York (2005). 41, 55, 79

[146] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees.* CRC press (1984). 44

[147] Bradley Efron. *Computers and the theory of statistics : thinking the unthinkable.* SIAM review **21**(4), 460–480 (1979). 44

[148] Leo Breiman. *Bagging predictors.* Machine learning **24**(2), 123–140 (1996). 44

[149] Nicolai Meinshausen. *Quantile regression forests.* The Journal of Machine Learning Research **7**, 983–999 (2006). 45, 46

[150] V Aruoja, M Sihtmae, H-C Dubourguier, and A Kahru. *Toxicity of 58 substituted anilines and phenols to algae pseudokirchneriella subcapitata and bacteria vibrio fischeri : Comparison with published data and qsars.* Chemosphere **84**, 1310–1320 (2011). 47, 50

[151] T Netzeva, M Pavan, and A Worth. *Review of data sources, qsars and integrated testing strategies for aquatic toxicity.* EU Book Shop, EU (2007). 47, 50

[152] S-H Hsieh, C-H Hsu, D-Y Tsai, and C-Y Chen. *Quantitative structure-activity relationships for toxicity of nonpolar narcotic chemicals to pseudokirchneriella subcapitata.* Environ Toxicol Chem **25**, 2920–2926 (2006). 47, 50

[153] CJ Van Leeuwen, PTJ Van Der Zandt, T Aldenberg, HJM Verhaar, and JLM Hermens. *Application of qsars, extrapolation and equilibrium partitioning in aquatic effects assessment. i. narcotic industrial pollutants.* Environ Toxicol Chem **11**, 267–282 (1992). 47, 50

[154] Henk JM Verhaar, Cees J Van Leeuwen, and Joop LM Hermens. *Classifying environmental pollutants.* Chemosphere **25**(4), 471–491 (1992). 47

[155] Cornelis Johannes van Leeuwen and Theodorus Gabriel Vermeire. *Risk assessment of chemicals : an introduction.* Springer Science & Business Media (2007). 47, 48

[156] BI Escher and JL Hermens. *Modes of action in ecotoxicology : their role in body burdens, species sensitivity, qsars, and mixture effects.* Environ Sci Technol **36**, 4201–4217 (2002). 48

[157] J Vogelgesang. *The ec white paper on a strategy for a future chemicals policy.* Altern Lab Anim **30**, 211–212 (2002). 50

[158] TI Netzeva, M Pavan, and AP Worth. *Review of (quantitative) structure–activity relationships for acute aquatic toxicity.* QSAR Combinatorial Science **27**, 77–90 (2008). 50

[159] J Chen, X Li, H Yu, Y Wang, and X Qiao. *Progress and perspectives of quantitative structure-activity relationships used for ecological risk assessment of toxic organic compounds.* Science in China Series B : Chemistry **51**, 593–606 (2007). 50

[160] E Furusjo, A Svenson, M Rahmberg, and M Andersson. *The importance of outlier detection and training set selection for reliable environmental qsar predictions.* Chemosphere **63**, 99–108 (2006). 50

[161] L Fu, JJ Li, Y Wang, XH Wang, Y Wen, WC Qin, LM Su, and YH Zhao. *Evaluation of toxicity data to green algae and relationship with hydrophobicity.* Chemosphere **120**, 16–22 (2015). 50

[162] JA McGrath, TF Parkerton, and DM Di Toro. *Application of the narcosis target lipid model to algal toxicity and deriving predicted-no-effect concentrations.* Environ Toxicol Chem **23**, 2503–2517 (2004). 50, 70

[163] SP Bradbury. *Quantitative structure-activity relationships and ecological risk assessment : an overview of predictive aquatic toxicology research.* Toxicol Lett **79**, 229–237 (1995). 51

[164] H Verhaar, CV Leeuwen, and J Hermens. *Classifying environmental pollutants. 1. structure-activity relationships for prediction of aquatic toxicity.* Chemosphere **25**, 471–491 (1992). 51, 62, 79

[165] L Wang, K Chen, Y Ong, C Hwang, and J Shim. *A simple quantile regression via support vector machine. in : Advances in natural computation.* Lecture Notes in Computer Science. Springer Berlin Heidelberg **3610**, 512–520 (2005). 51

[166] WM Meylan and PH Howard. *Atom/fragment contribution method for estimating octanol-water partition coefficients.* J Pharm Sci **84**, 83–92 (1995). 53, 79

[167] AK Ghose. *Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods : an analysis of alogp and clogp methods.* J Phys Chem **102**, 3762–3772 (1998). 53

[168] M Hahn. *Receptor surface models. 1. definition and construction.* J Med Chem **38**, 2080–2090 (1995). 53

[169] JK Labanowski and JW Andzelm. *Density functional methods in chemistry.* Springer-Verlag ; New York, Inc. (1991). 53

[170] W Kohn and LJ Sham. *Self-consistent equations including exchange and correlation effects.* Physical Review **140**, A1133 (1965). 53

[171] JP Perdew and Y Wang. *Accurate and simple analytic representation of the electron-gas correlation energy.* Physical Review **B 45**, 13244 (1992). 53

[172] DT Stanton and PC Jurs. *Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies.* Anal Chem **62**, 2323–2329 (1990). 53

[173] Parasurf. http ://www.ceposinsilico.de/products/parasurf.htm. 54

[174] T Clark, A Alex, B Beck, J Chandrasekhar, P Gedeck, AHC Horn, M Hutter, B Martin, G Rauhut, W Sauer, T Schindler, and T Steinke. *Vamp. 10.0, computer-chemie-centrum.* Universität Erlangen-Nürnberg, Erlangen (2008). 54

[175] CJF Bottcher, A Rip, OC Van Belle, and P Bordewijk. *Theory of electric polarization.* Elsevier Scientific Pub. Co, Amsterdam, New York (1952). 54, 79

[176] AJ Hopfinger. *Conformational properties of macromolecules.* Molecular biology, Academic Press, New York (1973). 54, 79

[177] X He and Q Shao. *A general bahadur representation of m-estimators and its application to linear regression with non stochastic designs.* Ann Stat **24**, 2608–2630 (1996). 56

[178] Gilles Durrieu and Laurent Briollais. *Sequential determination of sample size for robust linear regression : application to microarray experimental designs.* J Am Stat Assoc **104**, 650–660 (2009). 56, 79, 99, 104, 107, 108, 109

[179] Y Dodge and J Jurečková. *Estimation of quantile density function based on regression quantiles.* Stat Probab Lett **23**, 73–78 (1995). 56

[180] R Koenker. *Confidence intervals for regression quantiles.* Springer, New York (1994). 56

[181] R Koenker. *Rank tests for linear models.* Springer, New york (1996). 56

[182] C Gutenbrunner, J Jurečková, R Koenker, and S Portnoy. *Tests of linear hypotheses based on regression rank scores.* J Non Parametr Stat **2**, 307–333 (1993). 56

[183] M Kocherginsky, X He, and Y Mu. *Practical confidence intervals for regression quantiles.* J Comput Graph Stat **14**, 41–55 (2005). 56

[184] M Parzen, L Wei, and Z Ying. *A resampling method based on pivotal estimating functions.* Biometrika **81**, 341–350 (1994). 56

[185] Y Bilias, S Chen, and Z Ying. *Simple resampling methods for censored regression quantiles.* J Econ **99**, 373–386 (2000). 56

[186] E Khmaladze. *Martingale approach in the theory of goodness-of-fit tests.* Theory Probab Appl **26**, 240–257 (1981). 56

[187] R Koenker and Z Xiao. *Inference on the quantile regression process.* Econometrica **81**, 1583–1612 (2002). 56

[188] L Briollais and G Durrieu. *Application of quantile regression to recent genetic and -omic studies.* Hum Genet **133**, 951–966 (2014). 56, 79

[189] Y Liu, C Zou, and R Zhang. *Empirical likelihood ratio test for a change-point in linear regression model.* Communication in statistics - theory and methods **37**, 2551–2563 (2008). 56

[190] H Kubinyi. *Quantitative structure–activity relationships. 7. the bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character.* J Med Chem **20**, 625–629 (1977). 57

[191] B Caputo, K Sim, F Furesjo, and A Smola. *Appearance-based object recognition using svms : Which kernel should i use.* Proceeding of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision (2002). 59, 84

[192] Cherkassky V and Y Ma. *Practical selection of svm parameters and noise estimation for svm regression.* Neural Netw **17**, 113–126 (2004). 59

[193] D Mattera and S Haykin. *Advances in kernel methods.* MIT Press Cambridge pages 211–241 (1999). 59

[194] J Thioulouse and S Dray. *Interactive multivariate data analysis in r with the ade4 and ade4tkgui packages.* Journal of Statistical Software **22**, 1–14 (2007). 60

[195] A Karatzoglou, A Smola, K Hornik, and A Zeileis. *kernlab - an s4 package for kernel methods in r.* J Stat Softw **11**, 1–20 (2004). 60

[196] LHMLM Santos, AN Araujo, A Fachini, A Pena, C Delerue-Matos, and MCBSM Montenegro. *Ecotoxicological aspects related to the presence of pharmaceuticals in the aquatic environment.* J Hazard Mater **175**, 45–95 (2010). 62

[197] WH Vaes, EU Ramos, C Hamwijk, I van Holsteijn, BJ Blaauboer, W Seinen, HJ Verhaar, and JL Hermens. *Solid phase microextraction as a tool to determine membrane/water partition coefficients and bioavailable concentrations in in vitro systems.* Chem Res Toxicol **10**, 1067–1072 (1997). 62

[198] L Michielan, L Pireddu, M Floris, and S Moro. *Support vector machine (svm) as alternative tool to assign acute aquatic toxicity warning labels to chemicals.* Molecular Informatics **29**, 51–64 (2010). 63

[199] Dropdata a guide to pesticides grouped by mode of action, http ://www.dropdata.org /rpu/pesticides_moa.htm, (2009). 66

[200] R Munday. *Toxicity of thiols and disulphides : involvement of free-radical species.* Free Radical Biol Med **7**, 659–673 (1989). 66

[201] K Fent, AA Weston, and D Caminada. *Ecotoxicology of human pharmaceuticals.* Aquat Toxicol **76**, 122–159 (2006). 77

[202] SD Richardson and TA Ternes. *Water analysis : Emerging contaminants and current issues.* Anal Chem **86**, 2813–2848 (2014). 77

[203] CG Daughton and TA Ternes. *Pharmaceuticals and personal care products in the environment : agents of subtle change ?* Environ Health Perspect **107**, 907–938 (1999). 77

[204] Markus Liebig, Johann Moltmann, and Thomas Knacker. *Evaluation of measured and predicted environmental concentrations of selected human pharmaceuticals and personal care products (10 pp).* Environmental Science and Pollution Research **13**(2), 110–119 (2006). 78

[205] L Minguez, J Pedelucq, E Farcy, C Ballandonne, H Budzinski, and MP Halm-Lemeille. *Toxicities of 48 pharmaceuticals and their freshwater and marine environmental assessment in northwestern france.* Environ Sci Pollut Res (2014). 78, 80

[206] WA Warr. *Scientific workflow systems : Pipeline pilot and knime.* J Comput Aided Mol Des **26**, 801–804 (2012). 78

[207] J Villain, S Lozano, M-P Halm-Lemeille, G Durrieu, and R Bureau. *Quantile regression model for a diverse set of chemicals : application to acute toxicity for green algae.* J Mol Mod **20**, 2508–2521 (2014). 79, 82, 84, 87

[208] Gunter Ritter and María Teresa Gallegos. *Outliers in statistical pattern recognition and an application to automatic chromosome classification.* Pattern Recognition Letters **18**(6), 525–539 (1997). 84

[209] V Bamnett and T Lewis. Outliers in statistical data, (1994). 84

[210] B Schölkopf, RC Williamson, JS Alex, S-T John, and CP John. *Support vector method for novelty detection.* NIPS proceeding pages 582–588 (2000). 84, 95

[211] C-Y Wu and LZ Benet. *Predicting drug disposition via application of bcs : Transport/absorption/ elimination interplay and development of a biopharmaceutics drug disposition classification system.* Pharm Res **22**, 11–23 (2005). 86, 91

[212] H Hoffmann. *Kernel pca for novelty detection.* Pattern Recogn **40**, 863–874 (2007). 87

[213] A Varnek and I Baskin. *Machine learning methods for property prediction in chemoinformatics : Quo vadis ?* J Chem Inf Model **52**, 1413–1437 (2012). 87

[214] M Gonzalez-Pleiter, S Gonzalo, Rodea-Palomares, F I, Leganes, R Rosal, K Boltes, E Marco, and F Fernandez-Pinas. *Toxicity of five antibiotics and their mixtures towards photosynthetic aquatic organisms : implications for environmental risk assessment.* Water Res **47**, 2050–2064 (2013). 90

[215] V Law, C Knox, Y Djoumbou, T Jewison, AC Guo, Y Liu, A Maciejewski, D Arndt, M Wilson, V Neveu, A Tang, G Gabriel, C Ly, S Adamjee, ZT Dame, B Han, Y Zhou, and DS Wishart. *Drugbank 4.0 : shedding new light on drug metabolism.* Nucleic Acids Res **42** (2014). 90, 94

[216] AL Manuell, K Yamaguchi, PA Haynes, RA Milligan, and SP Mayfield. *Composition and structure of the 80s ribosome from the green alga chlamydomonas reinhardtii : 80s ribosomes are conserved in plants and animals.* J Mol Biol **351**, 266–279 (2005). 90

[217] MB Miller, BA Haubrich, Q Wang, WJ Snell, and WD Nes. *Evolutionarily conserved delta(25(27))-olefin ergosterol biosynthesis pathway in the alga chlamydomonas reinhardtii.* J Lipid Res **53**, 1636–1645 (2012). 90

[218] WH Beggs and CE Hughes. *Exploitation of the direct cell damaging action of antifungal azoles.* Diagn Microbiol Infect Dis **6**, 1–3 (1987). 91

[219] I Knutter, G Kottra, W Fischer, H Daniel, and M Brandsch. *High-affinity interaction of sartans with h+/peptide transporters.* Drug Metab Dispos **37**, 143–149 (2009). 91

[220] RB Kim. *Transporters and xenobiotic disposition.* Toxicology **181–182**, 291–297 (2002). 91

[221] LZ Benet, F Broccatelli, and TI Oprea. *Bddcs applied to over 900 drugs.* The AAPS Journal **13**, 519–547 (2011). 91, 92

[222] S Takeo, H Yamada, K Tanonaka, M Hayashi, and N Sunagawa. *Possible involvement of membrane-stabilizing action in beneficial effect of beta adrenoceptor blocking agents on hypoxic and posthypoxic myocardium.* J Pharmacol Exp Ther **254**, 847–856 (1990). 93

[223] AN Dohadwalla, AS Freedberg, and EM Vaughan Williams. *The relevance of β-receptor blockade to ouabain-induced cardiac arrhythmias.* British journal of pharmacology **36**(2), 257–267 (1969). 93

[224] HIDEAKI Sada and TAKASHI Ban. *Effects of acebutolol and other structurally related beta adrenergic blockers on transmembrane action potential in guinea-pig papillary muscles.* Journal of Pharmacology and Experimental Therapeutics **215**(2), 507–514 (1980). 93

[225] Bramah N Singh, Heather D Nisbet, Edward A Harris, and Robert ML Whitlock. *A comparison of the actions of ici66082 and propranolol on cardiac and peripheral β-adrenoceptors.* European journal of pharmacology **34**(1), 75–86 (1975). 93

[226] Hideaki Sada and Takashi Ban. *Effects of various structurally related beta-adrenoceptor blocking agents on maximum upstroke velocity of action potential in guinea-pig papillary muscles.* Naunyn-Schmiedeberg's archives of pharmacology **317**(3), 245–251 (1981). 93

[227] J Neuwoehner, K Fenner, and BI Escher. *Physiological modes of action of fluoxetine and its human metabolites in algae.* Environ Sci Technol **43**, 6830–6837 (2009). 93

[228] Lúcia HMLM Santos, Alberto N Araújo, Adriano Fachini, Angelina Pena, Cristina Delerue-Matos, and MCBSM Montenegro. *Ecotoxicological aspects related to the presence of pharmaceuticals in the aquatic environment.* Journal of hazardous materials **175**(1), 45–95 (2010). 94

[229] M Markou and S Singh. *Novelty detection : a review-part 1 : statistical approaches.* Signal Processing **83**, 2481–2497 (2003). 95

[230] Bertrand Clarke, Ernest Fokoue, and Hao Helen Zhang. *Principles and theory for data mining and machine learning.* Springer Science & Business Media (2009). 95

[231] Emanuel Parzen. *On estimation of a probability density function and mode.* The annals of mathematical statistics **33**(3), 1065–1076 (1962). 97

[232] Murray Rosenblatt et al. *Remarks on some nonparametric estimates of a density function.* The Annals of Mathematical Statistics **27**(3), 832–837 (1956). 97

[233] Raphaël Coudret, Gilles Durrieu, and Jerôme Saracco. *Comparison of kernel density estimators with assumption on number of modes.* Communications in Statistics-Simulation and Computation **44**(1), 196–216 (2015). 97

[234] Charles Stein. *A two-sample test for a linear hypothesis whose power is independent of the variance.* The Annals of Mathematical Statistics **16**(3), 243–258 (1945). 100

[235] Yuan S Chow and Herbert Robbins. *On the asymptotic theory of fixed-width sequential confidence intervals for the mean.* The Annals of Mathematical Statistics **36**(2), 457–462 (1965). 100

[236] Malay Ghosh and Pranab Kumar Sen. *On bounded length confidence interval for the regression coefficient based on a class of rank statistics*. Sankhyā : The Indian Journal of Statistics, Series A pages 33–52 (1972). 100

[237] Roger Joseph Boscovich. *De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa*. Bononiensi Scientarum et Artum Instituto Atque Academia Commentarii **4**, 353–396 (1757). 101

[238] PS Laplace. Mémoire sur les solutions particulières des équations différentielles et sur les inégalites séculaies des planàtes, oeuvres completes ix 325, (1895). 101

[239] Francis Y Edgeworth. *On observations relating to several quantities*. Hermathena **6**(13), 279–285 (1887). 101

[240] RW Farebrother. *Algorithm as 231 : The distribution of a noncentral chi2 variable with nonnegative degrees of freedom*. Journal of the Royal Statistical Society. Series C (Applied Statistics) **36**(3), 402–405 (1987). 101

[241] James E Gentle. *Least absolute values estimation : An introduction*. Communications in Statistics-Simulation and Computation **6**(4), 313–328 (1977). 102

[242] P Bloomfield and WL Steiger. *Least absolute deviations, theory, applications*. Algorithms (1983). 102

[243] Francis J Anscombe. Large-sample theory of sequential estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, pages 600–607. Cambridge Univ Press (1952). 106, 110

N°404