Relation structure/ facteur acentrique d'alcools et de phénols : approche algorithme génétique – régression linéaire multiple.

Structure / acentric factor relationship of alcohols and phenols: genetic algorithm – multiple linear regression approach

Hamza Haddag[1], Amel Bouakkadia[1, 2], Leila. Lourici[1, 3*], Nasr Eddine Chakri[1], Djelloul Messadi[1]

[1]*Laboratoire de Sécurité Environnementale et Alimentaire, Université BADJI Mokhtar, BP 12, 23000 Annaba, Algérie.*

[2]*Université Abbès Laghrour Khenchela , Algérie.*

[3]*Université Chadli Bendjedid -36000 - El Tarf, Algérie.*

Soumis le 04/09/2016 *Révisé le 05/02/2017* *Accepté le 21/02/2017*

ملخص

المعامل الغيرمركزي لـ 18 مركبا هيدروكسيليا (كحولات،فينولات)، ربطت خطيا بمواصفين جز يئيين من الصنف الهندسي تم اختيارهما بواسطة الخوارزمي الجيني من بين 1600حسبت باستعمال برنامج النمذجة الجزيئية.DRAGON. الأحصاءات المختلفة (معاملا التحديد المتعدد و التنبؤ، جذورالأخطاء المربعة المتوسطة ...) تبين جودة، متانة و قدرة التنبؤ الداخلية الجيدة للنموذج. لم نحصي أي ملاحظة نافذة أو شاذة.

الكلمات الجوهرية *: كحولات و فينولات – التمثيل الرقمي للتركيب الكيميائي – المعامل الغير مركزي– التراجع المتعدد الخطي – النموذج الهجين PSR.*

Abstract

The acentric factors of 18 hydroxy compounds (alcohols, phenols) were linearly correlated with 2 molecular descriptors of geometrical type selected by genetic algorithm, among more than 1600 derived from the molecular modeling software DRAGON. The different statistics calculated (multiple determination and prediction coefficients; roots of the mean quadratic errors; Y-scrambling) show the quality, the robustness and the good internal predictive capacity of the constructed model. No outliers or influential observation was found.

Key words*: Alcohols and phenols – Numerical representation of chemical structure – Acentric factor – Multiple linear regression – Hybrid SPR model.*

Résumé

Les facteurs acentriques de 18 composés hydroxylés (alcools, phénols), ont été corrélés linéairement avec 2 descripteurs moléculaires de type géométrique sélectionnés par algorithme génétique, parmi plus de 1600 calculés en utilisant le logiciel de modélisation moléculaire DRAGON. Les différentes statistiques établies (cœfficient de détermination multiple et de prédiction ; racines des erreurs quadratiques moyennes ; test de randomisation) montrent la qualité, la robustesse et les bonnes capacités prédictives internes du modèle construit. Aucune observation aberrante ou influente n'a été relevée.

Mots clés *: Alcools et phénols – Représentation numérique de la structure chimique – Facteur acentrique – Régression linéaire multiple – Modèle RSP hybride.*

[*]*Auteur correspondent : leilalourici@yahoo.fr*

1. INTRODUCTION

Le facteur acentrique ω est un paramètre parmi les plus courants des corps purs. Comme proposé à l'origine par Pitzer [1,2] ω représente la non sphéricité d'une molécule. De ce fait, le facteur acentrique est très utilisé pour la détermination de nombreuses propriétés thermodynamiques (facteur de compressibilité, pression de vapeur, enthalpie de vaporisation, coefficients de l'équation du viriel) et dans les études des équilibres de phases des substances [3-4].
Pour les gaz monoatomiques, ω est essentiellement nul, et pour le méthane sa valeur est encore très petite. Cependant, ω croît avec la masse moléculaire des hydrocarbures, de même qu'avec la polarité.
Si, à présent, ω est très largement utilisé pour caractériser la complexité d'une molécule du point de vue de la géométrie et de la polarité [5], les grandes valeurs du facteur acentrique de certains composés polaires (ω >0,4) ne sont pas significatives dans l'acception originelle de cette propriété.
L'objectif de ce travail vise à utiliser la méthodologie RSP (pour Relation Structure/ Propriété), dans l'approche algorithme génétique/ régression linéaire multiple (AG/RLM), pour relier les facteurs acentriques, compris entre 0,433 et 0,665, d'un ensemble hétérogène d'alcools et de phénols, à des descripteurs moléculaires reflétant certaines particularités des molécules prises en compte. L'interprétation de ces descripteurs permettrait d'avoir un aperçu sur les facteurs vraisemblablement liés aux facteurs acentriques des alcools et phénols considérés.
La qualité de l'ajustement et la robustesse du modèle ont été vérifiées.

2. METHODOLOGIE

2.1 Ensemble de données :

Les facteurs acentriques d'un ensemble hétérogène d'alcools et de phénols ont été prélevés dans la littérature [3]. Ces données (Tab. 1) se rapportent à 13 alcanols à chaînes ouvertes (linéaires ou ramifiées) ou fermées, et 5 dérivés phénoliques ; on y relève plusieurs isomères (chaîne, position).

2.2 Descripteurs moléculaires :

La représentation numérique de la structure chimique (descripteurs moléculaires) est une étape importante de l'investigation RSP. La qualité du modèle élaboré est étroitement liée au mode de détermination de ces descripteurs.
Nous avons utilisé le logiciel de modélisation moléculaire HyperChem 6.03 [6] pour représenter les molécules, puis obtenir les géométries finales à l'aide de la méthode semi-empirique AM1. Tous les calculs ont été exécutés dans le cadre du formalisme de Hartree-Fock avec contrainte de spin (ou RHF, pour Restricted Hartree-Fock) sans interaction de configuration. Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak-Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,001 kcal.mol^{-1}. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique DRAGON [7] pour le calcul de plus de 1600 descripteurs appartenant à 20 classes différentes.
En utilisant les options correspondantes du logiciel DRAGON, nous avons d'abord éliminé les descripteurs à valeurs constantes (écarts types inférieurs à 0,0001) qui n'apportent aucune information, ensuite ceux qui sont hautement corrélés ($R \geq 0{,}9$) et qui véhiculent une information redondante. Pour chaque paire de descripteurs corrélés, est éliminé automatiquement celui qui présente les plus hautes corrélations croisées avec les autres descripteurs.

2.3 Choix d'un sous-ensemble de descripteurs (VSS, pour Variable Subset Selection) par Algorithme génétique (GA/VSS):

On dispose souvent de plus de descripteurs qu'il n'est nécessaire. Et plutôt que de chercher à expliquer la variable dépendante (facteur acentrique) par tous les régresseurs (descripteurs moléculaires) disponibles, on peut chercher seulement un ensemble réduit de régresseurs qui donne une reconstitution aussi satisfaisante de la variable à expliquer. Parmi les stratégies mises en œuvre pour la sélection d'un ensemble réduit de variables explicatives, on peut citer : les méthodes de pas-à-pas, les algorithmes évolutifs et génétiques [8,9]; la comparaison se fait souvent à l'avantage de ces derniers.
La modélisation de processus génétiques a initié le développement des algorithmes génétiques, qui peuvent être exploités dans une grande variété de problèmes d'optimisation [10].
Dans un algorithme génétique adapté à l'optimisation, une solution potentielle est considérée comme un individu dans une population. La valeur de la fonction de coût associée à une solution mesure "l'adaptation" de l'individu associé à son environnement. Un algorithme génétique simule l'évolution,

sur plusieurs générations, d'une population initiale dont les individus sont mal adaptés au moyen d'opérateurs génétiques de reproduction et de mutation. Après un certain nombre de générations, la population est constituée d'individus bien adaptés, autrement dit des solutions supposées "bonnes" au problème d'optimisation.
Dans ce travail la sélection des descripteurs a été réalisée par algorithme génétique dans le logiciel MobyDigs [11], en maximisant le coefficient de prédiction Q^2_{LOO}.

2.4 Modèle de régression multiple :

Par souci de simplicité on utilise la régression linéaire multiple (RLM) qui impose des transformations linéaires dans les relations entre descripteurs et propriétés étudiées.
Un modèle de régression multiple entre une variable expliquée Y et l variables explicatives $X_1, X_2, ..., X_l$, s'écrit pour tout $i = 1,2,...,n$:

$$y_i = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} + \varepsilon_i \qquad (1)$$

où les $y_i, x_{i1}, ..., x_{il}$ sont des données respectivement relatives aux variables $Y, X_1, ..., X_l$.
Les estimateurs des coefficients β_j sont calculés en utilisant la méthode des moindres carrés ordinaires. Les variables aléatoires ε_i représentent les termes d'erreur non observables du modèle. On peut estimer ces erreurs par les résidus ordinaires e_i, différences entre les valeurs observées y_i et les valeurs estimées $\hat{y}_i$.
L'estimation par les moindres carrés des coefficients de régression suppose que les données suivent la loi normale, ce qui sera vérifié systématiquement.
Deux paramètres statistiques sont couramment utilisés pour l'évaluation de la qualité du modèle :

- Le coefficient de détermination multiple :

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

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

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

$$EQMP = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2} = \sqrt{\frac{PRESS}{n}} \qquad (3)$$

Il est intéressant de considérer, également, la racine de l'écart quadratique moyen calculé sur l'ensemble de calibrage (EQMC), c'est-à-dire l'ensemble qui a servi à la construction du modèle.

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

La validation croisée par "Leave –One –Out" (LOO) [12] consiste à calculer le modèle sur (n-1) composés, et à utiliser le modèle ainsi obtenu pour calculer le facteur acentrique du composé écarté, noté $\hat{y}_{(i)}$. On répète le procédé pour chacun des n composés. La somme des carrés des erreurs de prédiction, désignée par le symbole PRESS dans l'équation (3), est une mesure de la dispersion des estimations. On l'utilise pour définir le coefficient de prédiction :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \quad (5)$$

Contrairement à R^2, qui augmente avec le nombre de paramètres du modèle, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO} .Une valeur de Q^2_{LOO} > 0,5 est considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [13].

3. RESULTATS ET DISCUSSION

3.1 Sélection des descripteurs moléculaires :

L'optimisation par algorithme génétique (GA-VSS) conduit à de nombreux modèles de différentes dimensions. Parmi les modèles sélectionnés nous avons retenu le plus simple à deux variables explicatives (de coefficient de corrélation r= 0,092 pour une valeur de p= 0,716) qui sont des descripteurs moléculaires géométriques : l'autocorrélation à levier pondéré de distance topologique 3/ pondérée par les volumes atomiques de van der Waals v (HATS3v), et la seconde composante de l'indice de taille WHIM dirigé/ pondérée par les polarisabilités p (L2p).

Les descripteurs moléculaires à invariant holistique pondéré (WHIM) [14,15], permettent de saisir dans le détail les informations relatives à la taille, la forme, la symétrie et la distribution des atomes d'une molécule par rapport à des cadres de références fixes. Le calcul des descripteurs WHIM repose sur l'analyse en composantes principales de la matrice de covariance des coordonnées atomiques pondérées, dont les éléments sont définis par :

$$s_{jk} = \frac{\sum_{i=1}^{n} w_i \left(q_{ij} - \overline{q}_j\right)\left(q_{ij} - \overline{q}_k\right)}{\sum_{i=1}^{n} w_i} \quad (6)$$

où n représente le nombre d'atomes de la molécule, w_i , le poid du i^{ème} atome, q_{ij} la j^{ème} coordonnée cartésienne de l'atome i (j=1,2,3) alors que $\overline{q}_j$ est la moyenne de cette j^{ème} coordonnée.

Six modèles de pondération, rapportés à l'échelle de l'atome de carbone, sont proposés, et selon le mode adopté on obtient différentes matrices de covariance et différents axes principaux (c'est-à-dire des composantes $\mathbf{t_m}$, m=1, 2,3) pour la molécule.

On distingue les descripteurs WHIM dirigés, calculés individuellement selon les directions des composantes principales, et les descripteurs WHIM non dirigés, ou globaux, calculés pour la molécule entière à partir des combinaisons des premiers.

Les indices de taille WHIM dirigés, Lkw , sont définis par les valeurs propres $\lambda_k \left(k = 1,2,3\right)$ de la matrice de covariance des coordonnées atomiques pondérées de la molécule. Chaque vecteur propre mesure la dispersion (variance pondérée) des atomes projetés sur l'axe principal considéré, renseignant ainsi sur la dimension de la molécule selon cette direction principale.

Le descripteur moléculaire L2p qui est lié à la dimension des molécules sur le deuxième axe principal, met également en évidence le rôle de la polarisabilité.

Les descripteurs "Assemblage de géométrie, topologie et poids atomiques" GETAWAY (pour GEometry, Topology, and Atom Weights AssempblY) [16,17] sont basés sur les formules d'autocorrélation spatiales, en pondérant les atomes dans les molécules par des propriétés physico-chimiques, et par les informations 3D contenues dans les éléments des matrices influence moléculaires **H** et influence/distances **R** qui en est déduite (par minimisation des interactions entre paires d'atomes trop éloignés). La matrice **H,** elle-même, est définie à partir de la matrice moléculaire **M** des coordonnées cartésiennes x, y, z des atomes (y compris les hydrogènes) prises par rapport au barycentre de la molécule, considérée dans la conformation choisie. Les éléments diagonaux h_{ii} de **H** (ou leviers) renseignent sur "l'influence" de chaque atome de la molécule quant à déterminer la forme

globale de celle-ci; en fait, les atomes périphériques possèdent toujours de plus grands h_{ii} que les atomes voisins du barycentre de la molécule. De plus, l'ampleur du levier maximal d'une molécule dépend de sa grosseur et de sa forme. Notons enfin, ce qui peut être déduit de la géométrie moléculaire, que les valeurs des leviers sont sensibles à des changements conformationnels significatifs, et aux longueurs de liaison qui tiennent compte des types d'atomes et de la multiplicité des liaisons.

Les descripteurs d'autocorrélation à levier pondéré de distance topologique k (=3, dans notre cas), sont calculés à partir de l'équation :

$$HATSkw = \sum_{n=1}^{n_{AT}\text{-}1} \sum_{j>i} \left(w_i h_{ii}\right)\left(w_j h_{jj}\right)\delta\left(k; d_{ij}\right) \qquad (7)$$

$$k = 0,1,2,...,8$$

n_{AT} est le nombre d'atomes de la molécule ; d_{ij} est la distance topologique entre les atomes i et j, c'est-à-dire le nombre de liaisons du chemin le plus court reliant ces deux atomes; w_i est une pondération atomique physico-chimique (volume de van der Waals dans le cas présent);

$\delta\left(k; d_{ij}\right)$ est une fonction delta de Dirac ($\delta = 1$ si $d_{ij} = k$, sinon zéro).

Ils apportent une information sur la position effective, dans l'espace moléculaire, des substituants et des fragments de la molécule. De plus, ils renseignent, jusqu'à un certain point, sur la dimension et la forme moléculaire, ainsi que sur les propriétés atomiques spécifiques.

3.2 Modèle AG/MLR :

Avant de procéder au développement effectif des équations de régression, la qualité statistique des variables dépendante et explicatives a été vérifiée.

Tableau 1 : Valeurs des facteurs acentriques et des descripteurs moléculaires sélectionnés

N° (i)	Nom	ω_i	L2p	HATS3v	h_{ii}	e_{istd}
1	o-Cresol	0,433	1,726	0,118	0,265	0,2882
2	Phenol	0,438	1,481	0,115	0,186	1,3342
3	m-Cresol	0,454	1,516	0,123	0,178	0,5987
4	Pentafluorophenol	0,502	1,539	0,148	0,150	-0,9329
5	p-Cresol	0,505	1,297	0,137	0,101	-0,2194
6	Cyclohexanol	0,528	1,479	0,218	0,230	1,8875
7	Methanol	0,556	0,239	0,092	0,267	1,0597
8	Heptan-1-ol	0,560	0,565	0,107	0,146	-0,5199
9	Hexan-1-ol	0,560	0,541	0,119	0,123	0,2334
10	Butan-2-ol	0,577	0,703	0,177	0,086	1,2230
11	Pentan-1-ol	0,579	0,54	0,134	0,102	-0,0220
12	Octan-1-ol	0,587	0,563	0,100	0,166	-2,4464
13	2-Methylpropan-1-ol	0,592	1,220	0,192	0,112	-1,6401
14	Butan-1-ol	0,593	0,496	0,160	0,105	0,7535
15	2-Methylpropan-2-ol	0,612	1,218	0,237	0,262	-0,6252
16	Propan-1-ol	0,623	0,487	0,189	0,147	0,7080
17	Ethan-1-ol	0,644	0,377	0,203	0,214	1,0411
18	Propan-2-ol	0,665	0,792	0,212	0,161	-2,2746

Les diagrammes de probabilités établis à partir des données du tableau 1 montrent que les variables considérées se distribuent selon la loi normale, puisque les R obtenus sont systématiquement supérieurs aux R critiques (R_C) donnés par les tables pour les niveaux α=1% et α=5% , pour n=18 individus (Tab. 2).

Tableau 2 : Vérification de la loi de Laplace-Gauss pour n=18 individus.

	ω	HATS3v	L2p
R (%)	97,84	97,42	94,99
R_C (%)	94,55 (pour $\alpha = 5\%$) et 92,18 (pour $\alpha = 1\%$)		

Le modèle basé sur les descripteurs sélectionnés a pour équation :

$$\hat{\omega} = 0{,}510(\pm 0{,}022) + 0{,}938(\pm 0{,}124)\,\mathrm{HATS3v} - 0{,}107(\pm 0{,}011)\,\mathrm{L2p} \quad (8)$$

Il vérifie les hypothèses d'un modèle statistique linéaire à effets fixes. En effet la figure 1 reproduit la distribution des résidus normalisés RESN (Rapport : résidus ordinaires/ racine du carré moyen des écarts) en fonction des valeurs ajustées AJUST, qui semble aléatoire (sans tendance particulière), ce qui montre la constance des variances σ^2, c'est-à-dire leur indépendance des régresseurs et de la variable dépendante ajustée.

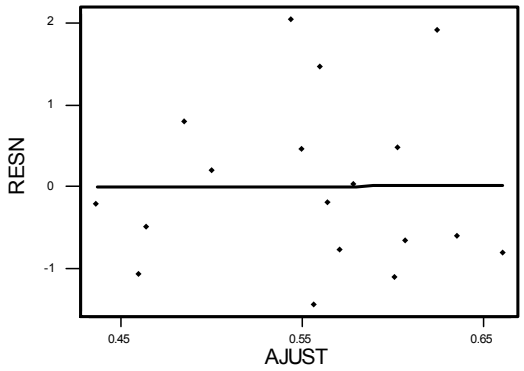

Figure 1: Graphe des résidus normalisés en fonction des facteurs acentriques ajustés.

La quasi-linéarité ($R = 0{,}9675; R_C = 0{,}9455$) du diagramme des scores normaux (Fig. 2) est un indice de normalité. La statistique de Durbin-Watson [18], $d=1{,}85$, est plus grande que la valeur supérieure donnée par les tables pour 2 régresseurs, et pour tout risque raisonnable α, ce qui établit l'indépendance des résidus.

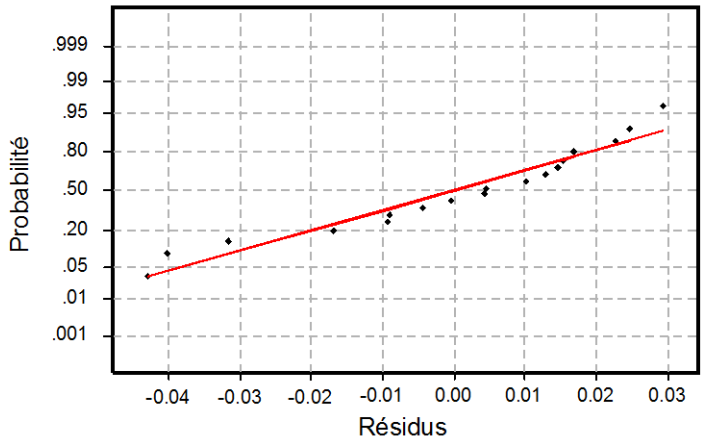

Figure 2 : Diagramme des scores normaux

Les diagnostiques statistiques du modèle sont rapportés ci-après :

$$R^2(\%) = 89{,}83\ ;\ Q^2(\%) = 85{,}35\ ;\ R^2_{adj} = 88{,}47\ ;\ EQMC = 0{,}020\ ;\ EQMP = 0{,}025\ ;$$
$$F = 66{,}24\ ;\ SE = 0{,}023$$

Les valeurs de R^2 et de R^2_{adj} montrent la qualité de l'ajustement, alors que la petite différence entre R^2 et Q^2 renseigne sur la robustesse du modèle qui, en outre, est hautement significatif (grande valeur du paramètre de Fisher F). De plus, la similitude de $EQMC$ et $EQMP$ signifie que la capacité de prédiction interne du modèle n'est pas trop dissemblable de son pouvoir d'ajustement

Les modèles RSP, à cause (souvent) de leur complexité et de la sophistication des outils de chimiométrie employés, peuvent constituer une source de corrélation fortuite. Dans le but d'établir que le modèle obtenu n'est pas dû au hasard, nous avons appliqué le test de randomisation de y . Ce test consiste à générer un vecteur "facteur acentrique" par permutation aléatoire des composantes du vecteur réel. On calcule alors sur le vecteur obtenu (considéré comme vecteur expérimental réel) un modèle RSP, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 dans notre cas).

La figure 3 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (carrés pleins) au modèle de départ (astérisque). Il est clair que les statistiques obtenues pour les vecteurs modifiés des facteurs acentriques sont plus petites (la majorité des valeurs de Q^2 sont même négatives) que celles du modèle RSP réel, ce qui permet d'assurer qu'une relation structure/facteur acentrique réelle a été établie.

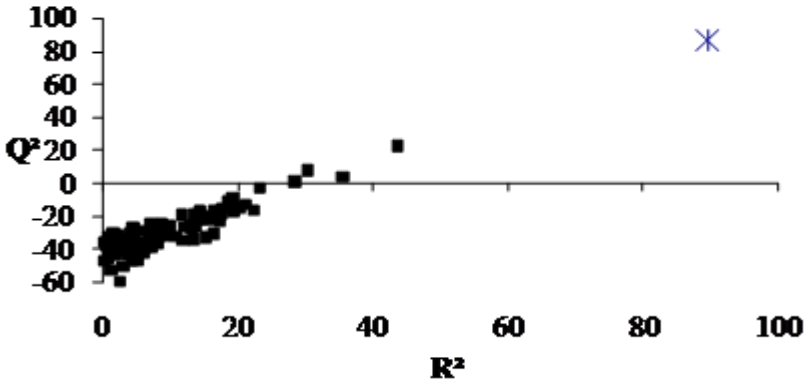

Figure 3 : Test de randomisation associé au modèle RSP

Pour détecter les observations aberrantes nous avons utilisé les résidus de prédiction standardisés [12] :

$$e_{istd} = \frac{e_{(i)}}{\sqrt{s^2_{(i)}\left(1 - h_{ii}\right)}} \tag{9}$$

pour lesquels l'estimation $S^2_{(i)}$ de σ^2 est calculée selon :

$$S^2_{(i)} = \frac{\left[n - \left(l+1\right)\right] CME - e_i^2 / \left(1 - h_{ii}\right)}{n - l - 2} \tag{10}$$

pour $\left(n - 1\right)$ observations, la i[eme] étant exclue ; e_i est le résidu ordinaire ; CME est le carré moyen des écarts, et $\left(l + 1\right)$ le nombre de paramètres du modèle.

Les valeurs absolues des résidus de prédiction standardisés (tableau 1, dernière colonne) étant toutes inférieures en valeur absolue à 3 unités d'écart type $\left(\left|e_{istd}\right| < 3\sigma\right)$ aucune donnée aberrante n'est ainsi détectée pour le modèle.

Les leviers h_{ii}, éléments diagonaux de la matrice **H** de passage du vecteur y au vecteur $\hat{y}$, permettent de juger de l'influence d'une observation i dans la détermination de l'équation de régression (lorsque h_{ii} est supérieur à la valeur critique $3\left(l+1\right)/n$). Toutes les valeurs reproduites dans la colonne h_{ii} (Tab. 1) étant inférieures à $3 \times 3/18 = 0,5$, aucune observation n'est influente.

Du fait de la différence entre les bases de données modélisées (du point de vue source et nombre de données) d'une part et des complexités des méthodes utilisées d'autre part, la comparaison des erreurs de calcul a été privilégiée. Les erreurs de calcul du facteur acentrique par les méthodes de contribution de groupes (MCG) [3] se distribuent entre 0,04 et 0,07 en unité log. Elles sont supérieures à celles des travaux cités dans le tableau 3 où AAD est la moyenne des valeurs absolues des déviations, et AAD% la valeur relative.

Tableau 3 : Comparaison avec les travaux antérieurs

	n	Méthode	AAD (AAD%)	SE	EQMC
Carande et al. [19]	614	RQSP (RVS)[b]	0,0310 (6,9)	0,023	0,048
Wang et al. [20]	477(48)[a]	MCG	0,0613 (10,39)	-	-
Mokshina et al. [21]	331	RQSP (RF)[c]	0,0140 (-)	0,027	-
Notre travail	18	RSP (AG/RLM)	0,0172 (3,05)	0,023	0,020

[a] 48 alcools pour lesquels les résultats sont rapportés.
[b] Régression par vecteurs supports.
[c] Random Forest pour la sélection des descripteurs.

Si le modèle de Mokshina et al. est le meilleur, il ne reste pas moins difficile à mettre en œuvre.la méthode de calcul des descripteurs n'étant pas automatisée contrairement au modèle bilinéaire que nous présentons dont les variables explicatives sont calculables rapidement par les logiciels disponibles.

4. CONCLUSION

Des logiciels informatiques (DRAGON, HyperChem …) permettent le calcul de nombreux descripteurs moléculaires utilisés pour modéliser une grande variété de propriétés. On trouve dans d'autres (MobyDigs …), du moins partiellement, une série d'outils créés pour la validation des modèles de régression, dont l'utilisation permet de mettre en évidence des situations particulières.

Ainsi, les facteurs acentriques d'un mélange hétérogène de composés hydroxylés (alcools, phénols), dont plusieurs isomères, ont pu être corrélés avec 2 indices structuraux de type géométrique. Le modèle hautement significatif obtenu, dont nous avons pu vérifier les hypothèses de départ, permet de reproduire les facteurs acentriques observés avec une précision moyenne inférieure à 3 % ; il possède une robustesse et une capacité prédictive satisfaisantes. Nous n'avons pas relevé d'observation présentant des valeurs extrêmes des caractéristiques établies, et qui puisse être considérée comme aberrante ou influente.

5. REFERENCES

[1] Pitzer K.S., 1955. The volumetric and thermodynamic properties of fluids. I. Theoretical basis and virial coefficients, Journal of the American Chemical Society, Vol. 77 (13), 3427–3433.

[2] Pitzer K.S., Lippmann D.Z., Curl R.F, Huggins C.M. & Petersen D.E., 1955. The volumetric and thermodynamic properties of fluids. II. Compressibility factor, vapor pressure and entropy of vaporization, Journal of the American Chemical Society, Vol. 77 (13), 3433–3440.

[3] Poling B.E., Prausnitz J.M. &. O'Connell J.P., 2001. The properties of gases & liquids, Fifth Ed., Mc Graw-Hill. 803p.

[4] Gharagheizi F., Eslamimanesh A., Sattari M., Mohammadi A.H. & Richon D., 2015. Computation of the second virial coefficient of chemical compounds using a corresponding states based method. In Advances in Chemistry Research. J. C. Taylor (Eds), Nova Science Publishers, Inc., Vol. 24, 91-112.

[5] Todeschini R., Consonni V., 2008. Handbook of molecular descriptors. R. Mannhold, H. Kubinyi & H. Timmermann, eds, Wiley, VCH. 688p

[6] Hyperchem ^{TM} Release 6.03 for windows, Molecular Modeling System (2000).

[7] Todeschini R., Consonni V. & Pavan M., 2005. DRAGON, Software for the Calculation of Molecular Descriptors. Release 5.3 for windows, Milano.

[8] Leach A.R., 2001. Molecular modelling: principles and applications. Second Ed., Prentice Hall. 744p

[9] Leach A.R & Gillet V.L., 2007. An introduction to chemoinformatics: Revised Ed.; Springer. 274p

[10] Chambers L., 1995. Practical handbook of genetic algorithms: Applications Volume I; CRC Press. 568p.

[11] Todeschini R., Ballabio D., Consonni V., Mauri A., Pavan M., MobyDigs Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 1.0 for Windows, Milano (2004).

[12] Draper N.R. & Smith H., 1998. Applied regression analysis, Third Ed., Wiley series in Probability and Statistics. 736p.

[13] Eriksson L., Jaworska J., Worth A.P., Cronin M.T.D., Mc Dowell R.M. & Gramatica P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs, Environmental Health Perspectives, Vol. 111 (10), 1361–1375.

[14] Todeschini R., Lasagni M. & Marengo E., 1994. New molecular descriptors for 2D and 3D structures. Theory, Journal of Chemometrics, Vol. 8 (4), 263–272.

[15] Todeschini R. & Gramatica P., 1997. 3D-Modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors, Quantitative Structure-Activity Relationships, Vol. 16, 113-119.

[16] Consonni V., Todeschini R. & Pavan M., 2002. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors, Journal of Chemical Information and Computer Sciences, Vol. 42 (3), 682–692.

[17] Consonni V., Todeschini R., Pavan M. & Gramatica P., 2002. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies, Journal of Chemical Information and Computer Sciences, Vol. 42 (3), 693–705.

[18] Durbin J. & Watson G.S., 1971, Testing for serial correlation in least squares regression III, Biometrika, Vol. 58 (1), 1-19.

[19] Carande W.H., Kazakov A., Muzny C. & Frenkel M., 2015. Quantitative Structure-Property Relationship Predictions of Critical Properties and Acentric Factors for Pure Compounds, Journal of Chemical & Engineering Data, Vol. 60, 1377–1387.

[20] Wang Q.,Jia Q. & Ma P., 2012. Prediction of the Acentric Factor of Organic Compounds with the Positional Distributive Contribution Method, Journal of Chemical & Engineering Data, Vol. 57, 169–189.

[21] Mokshina E.G., Kuz'min V.E. & Nedostup V.I., 2014. QSPR Modeling of Critical Parameters of Organic Compounds Belonging to Different Classes in Terms of the Simplex Representation of Molecular Structure, Russian Journal of Organic Chemistry, Vol. 50 (3), 314–321.