Modélisation de la rétention de phénols séparés par CLHP – PI avec une phase mobile méthanol – eau

Modelling the retention of phenols separated by RP – HPLC with a methanol – water mobile phase.

Djelloul Messadi *, Hamza Boufenaya, Leila Lourici, & Mohamed Lotfi Souici

Laboratoire de Sécurité Environnementale et Alimentaire (LASEA),
Université Badji Mokhtar Annaba, BP 12, 23 000, Annaba, Algérie.

Soumis le : 02.07.2014 *Révisé le 09.02.2015* *Accepté le : 16.02.2015*

ملخص

الإستبقاء (log k) لمجموعة غير متجانسة من الفينولات مفصولة بالطريقة الإيزوكراتيكية (isochratique) للكروماتوغرافيا السائلة عالية الأداء (HPLC-PI) ذات الطور المعكوس بإستعمال عمود بارتيزيل(partisil ODS) ، مع طور متحرك متكون من ميثانول – ماء، تم ربط هذا الإستبقاء مع الشروط التجريبية (درجة الحرارة، التجزئة الحجمية، طور مساعد-مذيب عضوي) و معامل التقسيم أكتانول –ماء لموريقوشي (Moriguchi) MlogP محسوب بمساعدة البرنامج الألي DRAGON.
مجموعة تشكيل النموذج (40 عنصر) تحصلنا عليها بإستعمال خوارزمية DUPLEX، فتمكنا من إيجاد نموذج يوفر الفرضيات لنموذج إحصائي خطي ذو تأثير ثابت،قوي،وله قابلية تصديق داخلية ليست بعيدة عن القابلية التعديلية .
التصديق الخارجي لمجموعة تحتوي على 26 عنصر تؤكد القابلية التنبؤية الجيدة ل لوغاريتم معامل التقسيم log k التي لم تستعمل في تشكيل النموذج.

الكلمات الدالة: فينولات – كروماتوغرافيا سائلة عالية الأداء – طور معكوس – نموذج *QSRR*

Résumé

La rétention (log k) d'un mélange hétérogène de phénols séparés en régime isochratique par CLHP – PI, sur une colonne Partisil ODS, avec une phase mobile méthanol – eau a été reliée aux conditions d'analyse (température T ; fraction volumique, φ, du co-solvant organique) et au coefficient de partage n-octanol / eau de Moriguchi (M log P) calculé à l'aide du logiciel DRAGON.
L'ensemble de calibrage (40 éléments), obtenu en appliquant l'algorithme DUPLEX, permet de calculer un modèle vérifiant les hypothèses d'un modèle statistique linéaire à effets fixes, robuste, et dont la capacité de prédiction interne n'est pas trop dissemblable de son pouvoir d'ajustement. La validation statistique externe, sur un ensemble test de 26 éléments, atteste de la bonne capacité prédictive des log k n'ayant pas servi au calcul du modèle.

Mots-clés: *Phénols – CLHP / PI – Rétention –Modèle QSRR.*

Abstract

The retention (log k) of an heterogeneous set of phenols separated by reversed phase chromatography (RP – HPLC) on a Partisil ODS column by using isochratic elution with methanol – water has been correlated to analytical condition (temperature T; volume fraction, φ, of organic co-solvent) and Moriguchi n-octanol / water partition coefficient calculated with DRAGON software for molecular modeling.
The DUPLEX algorithm was used to split the original data set into a training set (40 objects) and a validation set (26 objects). The proposed model, which fulfils the assumptions of a linear statistical model with fixed effects, is robust, and has internal predictivity not too dissimilar from fitting power. The model is also predictive for objects not used in the model development (statistical external validation on validation set objects).

Key-words: *Phenols – RP / HPLC – Retention – QSRR model*

Auteur correspondant : d_messadi@yahoo.fr

1. INTRODUCTION

Les phases mobiles hydro-organiques binaires sont couramment utilisées en chromatographie liquide haute performance à polarité de phase inversée (CLHP – PI), aussi la disponibilité d'un modèle de rétention du soluté est importante pour l'optimisation chromatographique.

La littérature en fait ressortir plusieurs [1,2] basés sur différentes relations fonctionnelles. Les modèles sont souvent évalués selon le critère statistique d'ajustement de la rétention d'un seul soluté pour différentes compositions de la phase mobile.

Bien que le mécanisme de rétention des silices greffées n-alkyles ne soit pas établi de manière définitive, une hypothèse simple plausible est que le solvant le moins polaire de la phase mobile (solvant organique) s'adsorbe préférentiellement à la surface des greffons apolaires : il y a alors partage des solutés entre la phase mobile et la phase liquide adsorbée. Ainsi, il peut être possible de relier les facteurs de rétention observés aux cœfficients de partage entre l'eau et un solvant organique [3], comme par exemple le coefficient de partage P (ou son logarithme) dans le système n-octanol/eau, développé par Hansch et ses collaborateurs [4,5] pour caractériser les propriétés hydrophobes des médicaments et autres substances biologiques actives.

De nombreuses méthodes ont été développées pour le calcul de log P à partir de la structure moléculaire, en prenant en compte différents critères [6].

Des propriétés de la phase mobile comme la viscosité, la constante diélectrique ou la tension superficielle peuvent jouer un rôle important dans la détermination de la rétention absolue et de la sélectivité en CLHP – PI. Cependant, si une viscosité faible diminue la rétention, au contraire, une constante diélectrique élevée favorisera les interactions entre les groupements polaires du soluté et la phase mobile et diminuera par conséquent le facteur de capacité [3].

Dans le mélange méthanol – eau, la tension superficielle et la constante diélectrique varient uniformément avec la fraction volumique, φ, du co-solvant organique, contrairement à la viscosité qui croît avec φ jusqu'à un maximum puis diminue de façon monotone [3,7]. L'influence de la température portera tout à la fois sur la fixation de la pression en tête de colonne, la durée de l'analyse, et la résolution. La perte de charge le long de la colonne est reliée au débit par la relation connue de Darcy. Elle montre que, sous certaines conditions, le débit peut être augmenté sans changement de la pression de travail si la viscosité diminue, ce qui peut être obtenu par élévation de la température. Une plus grande diminution du temps d'analyse peut être obtenue à la suite de la réduction du facteur de rétention due à l'effet de l'augmentation de la température, phénomène qui est largement déterminé par l'enthalpie d'interaction du soluté avec la phase stationnaire [8].

Dans ce travail nous nous intéresserons à quelques phénols analysés par CLHP – PI avec une phase mobile méthanol-eau, pour différentes fractions volumiques du co-solvant organique et pour différentes températures, T, de la colonne.

Nous présenterons un modèle de la rétention des phénols sélectionnés, qui tient compte des conditions d'analyse (φ, T), et qui fait intervenir un coefficient de partage n-octanol/eau, MlogP, calculé selon l'approche de Moriguchi [9,10], à l'aide du logiciel DRAGON [11].

Les logarithmes des facteurs de rétention, log k, mesurés seront éclatés en deux sous ensembles : l'un de calibrage pour le calcul du modèle, et l'autre de test uniquement utilisé pour la validation statistique externe. La qualité du modèle pouvant dépendre du choix de ces 2 ensembles, deux approches différentes seront comparées : le choix aléatoire et le choix basé sur l'algorithme DUPLEX [12]. Les hypothèses d'un modèle linéaire statistique à effets fixes, de même que la qualité de l'ajustement, ainsi que la robustesse du modèle, et ses capacités prédictives (interne et externe) seront examinées.

2. MATERIEL ET METHODE

Le chromatographe utilisé est un système Philips (Pye Unicam) constitué d'une pompe PU 4010, d'un injecteur Rhéodyne 7125, d'un détecteur UV / Visible à longueur d'onde variable PU 4200 et d'un enregistreur PU 8251. Une boucle de 20 µL

est remplie avec une seringue Hamilton de 25 µL.

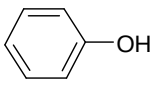
Une colonne de remplissage en inox (L : 25 cm ; diam. int. = 4,6 mm) contenant des radicaux octadécyl-silyle (Partisil ODS), dont le diamètre des particules support est 10 µm, a été utilisée pour les analyses. La colonne est placée dans une enveloppe en verre où circule de l'eau thermostatée, ce qui permet de fixer la température à 1 °C près.

Les analyses ont été réalisées en régime isochratique avec des phases mobiles constituées de mélanges (méthanol + eau) dans les rapports (V : V) 15 : 85, 25 : 75, 50 : 50, 70 : 30 et 85 : 15, pour un débit réglé à 2 ml / min. L'effet de la température sur la rétention a été examiné pour l'intervalle 12 – 52 °C (285 – 325 K) en faisant varier la température par bonds de 10 °, la phase mobile étant un mélange volume à volume méthanol – eau. Nous avons pris pour temps de rétention nulle, t_0, le temps de rétention de ^2H_2O qui absorbe à 190 nm. Pour faciliter le traitement des données nous avons codé les observations en désignant chacun des 8 phénols considérés par une lettre majuscule, suivie d'une lettre minuscule qui indique la température de la colonne (en K), et d'un nombre qui renseigne sur la fraction volumique φ (en %) du co-solvant organique (Tableau 1, figure 1).

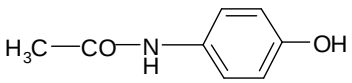
Tableau 1. Codage des observations. La valeur de M log P calculée est donnée, entre parenthèses, à la suite de chaque phénol.

A :	Phénol	(1.506)	**B :**	Paracétamol	(1.06)	**C :**	Résorcinol	(0.893)
D :	Gaïacol	(1.246)	**E :**	3-chlorophénol	(2.127)	**F :**	m-Xylenol	(2.193)
G :	Thymol	(2.813)	**H :**	m-Crésol	(1.859)			
p = 285 K ; q = 295 K ; r = 305 K ; s = 315 K ; t = 325 K								

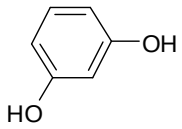
Ainsi, par exemple, le code Cq 25 correspond au résorcinol analysé à 295 K, lorsque la fraction volumique du méthanol dans la phase mobile est 25 %.



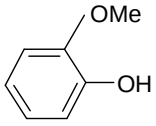
A/ Phénol



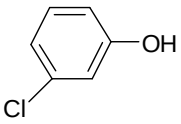
B/ Paracétamol



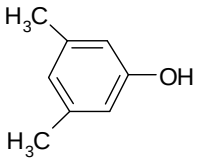
C/ Résorcinol



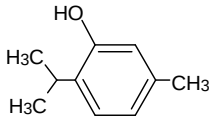
D/ Gaïacol



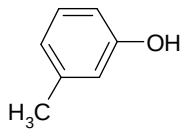
E/ 3-Chlorophénol



F/ m-Xylénol



G/ Thymol



H/ m-Crésol

Figure 1 : Noms et structures des phénols étudiés.

2.1. Choix des ensembles de calibrage et de test

Les 66 données x_{ij} (i = 1, 2, ……, 66 ; j = 1, 2, 3) du (tableau 2) ont été séparées en un ensemble de calibrage de 40 éléments, et un ensemble de test de 26 éléments. En plus du choix aléatoire (commande SAMPLE du logiciel MINITAB) [13], qui conduit en général à des compositions différentes de ces ensembles lorsqu'on répète l'opération, nous avons appliqué l'algorithme DUPLEX [12], que nous présentons succinctement.

Cet algorithme commence avec la liste des n (= 66) observations, les ℓ (= 3) régresseurs étant standardisés à l'unité selon :

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

Où : s_j : Ecart-type du jème régresseur.

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

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

n : Nombre d'observations.

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

$$\mathbf{Z'.Z} = \mathbf{T'.T} \qquad (2)$$

$\mathbf{T}$ est une matrice (ℓ × ℓ) triangulaire supérieure unique, dont les éléments peuvent être obtenus par la méthode de Cholesky [14]. On opère alors la transformation :

$$\mathbf{W} = \mathbf{Z.T^{-1}} \qquad (3)$$

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

2.2. Calcul et validation du modèle

L'analyse de régression multilinéaire est effectuée avec le logiciel MobyDigs [15] en utilisant la méthode des moindres carrés ordinaires.

La qualité de l'ajustement est évaluée par le coefficient de détermination, R^2, et l'écart quadratique moyen calculé sur l'ensemble de calibrage :

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

y_i et $\hat{y}_i$ étant les valeurs observées et calculées de la variable dépendante.

Les techniques de validation croisée ont été exploitées pour l'évaluation de la prédiction interne (bootstrap) et de la robustesse $\left(Q_{LOO}^2\right)$ du modèle.

La somme des carrés des erreurs de prédiction (PRESS) est une mesure de la dispersion des estimations, on l'utilise pour définir le coefficient de prédiction $\left(Q_{LOO}^2\right)$ et l'écart quadratique moyen de prédiction (ou EQMP) :

$$Q_{LOO}^2 = 1 - \frac{PRESS}{SCT} = 1 - \frac{\sum_1^n\left(y_i - \hat{y}_{(i)}\right)^2}{\sum_1^n\left(y_i - \bar{y}\right)^2} \qquad (5)$$

$$EQMP = \sqrt{PRESS/n} \qquad (6)$$

Tableau 2**.** Compositions des ensembles de calibrage (cal) et de test (val) obtenues selon un choix aléatoire, ou à l'aide de l'algorithme DUPLEX.

			Choix aléatoire			**Choix par DUPLEX**		
i	**Code**	Y_{Exp}	**Statut**	Y_{Pred}	e_{istd}	**Statut**	Y_{Pred}	e_{istd}
1	Ap50	0.1443	cal	0.1972	0.5869	val	0.1815	0.3688
2	Aq15	0.4853	val	0.6273	1.6263	cal	0.6014	1.1815
3	Aq25	0.3732	cal	0.4972	1.3856	cal	0.4652	0.9158
4	Aq50	0.0449	val	0.1247	0.8711	cal	0.1243	0.7735
5	Aq70	- 0.1313	cal	- 0,1414	- 0.1132	val	- 0,1260	0.0525
6	Aq85	- 0.2994	cal	- 0.3590	- 0.6942	val	- 0.3209	- 0.2205
7	Ar50	- 0.0013	cal	0.0574	0.6408	val	0.0623	0.6194
8	As50	- 0.0578	val	- 0.0135	0.4909	val	0.0027	0.5984
9	At50	- 0.0829	cal	- 0.0827	0.0030	cal	- 0.0533	0.3038
10	Bp50	- 0.0714	cal	0,0221	1.0493	val	0.0140	0.8548
11	Bq15	0.4216	val	0.4477	0.3044	cal	0.4210	- 0.0061
12	Bq25	0.2065	cal	0.3198	1.2853	cal	0.2995	0.9360
13	Bq50	- 0.1171	val	- 0.0548	0.6877	cal	- 0.0416	0.7438
14	Bq70	- 0.2484	cal	- 0.3278	- 0.8970	val	- 0.2935	- 0.4534
15	Bq85	- 0.2935	cal	- 0.5773	- 3.3431	cal	- 0.5238	- 2.4030
16	Br50	- 0.1694	cal	- 0.1212	0.5338	cal	- 0.1015	0.6702
17	Bs50	- 0.2121	val	0.1931	0.2144	cal	- 0.1603	0.5210
18	Bt50	- 0.2370	val	- 0.2622	- 0.2948	val	- 0.2245	0.1307
19	Cp50	- 0.1542	cal	- 0.0417	1.2732	cal	- 0.0373	1.1812
20	Cq15	0.2273	cal	0.4116	2.1723	cal	0.3801	1.5843
21	Cq25	0.1012	val	0.2409	1.5982	val	0.2284	1.2890
22	Cq50	- 0.1442	cal	- 0.1203	0.2667	cal	- 0.1056	0.3838
23	Cq70	- 0.2952	val	- 0.3874	- 1.0494	cal	- 0.3636	- 0.6947
24	Cq85	- 0.3188	val	- 0.5969	- 3.2964	val	- 0.5511	- 2.4433
25	Cr50	- 0.1957	cal	- 0.1908	0.0542	val	- 0.1679	0.2760
26	Cs50	- 0.2319	cal	- 0.2639	- 0.3647	val	- 0.2276	0.0439
27	Ct50	- 0.2648	cal	- 0.3436	- 0.9346	cal	- 0.2918	- 0.2847
28	Dp50	0.1291	cal	0.0892	- 0.4784	cal	0.0805	- 0.4840
29	Dq15	0.7485	cal	0.4873	- 3.0165	val	0.4909	- 2.6274
30	Dq25	0.5324	val	0.3830	- 1.6794	cal	0.3463	- 1.8597
31	Dq50	0.0946	val	0.0201	0.0400	val	0.0243	- 0.6883
32	Dq70	- 0.1316	val	- 0.2453	- 1.2751	cal	- 0.2313	- 0.9961
33	Dq85	- 0.2285	val	- 0.4548	- 2.6412	cal	- 0.4492	- 2.2821
34	Dr50	0.0432	cal	- 0.0532	- 1.0584	cal	- 0.0387	- 0.8027
35	Ds50	- 0.0067	cal	- 0.1274	- 1.3498	cal	- 0.1020	- 0.9495
36	Dt50	- 0.0392	val	- 0.1874	- 1.7156	cal	- 0.1730	- 1.3818
37	Ep50	0.3853	cal	0.4493	0.7186	cal	0.4174	0.3221
38	Dr50	0.0432	cal	- 0.0532	- 1.0584	cal	- 0.0387	- 0.8027
39	Eq25	0.8164	cal	0.7296	- 0.9774	cal	0.6785	- 1.3920
40	Eq50	0.3211	cal	0.3775	0.6213	cal	0.3569	0.3521
41	Eq70	0.0053	val	0.1093	1.1758	cal	0.1165	1.1145
42	Eq85	- 0.2650	val	- 0.1001	1.9460	val	- 0.0877	1.8336
43	Er50	0.2527	val	0.3056	0.5812	cal	0.2976	0.4412

44	Es50	0.1658	Cal	0.2422	0.8529	val	0.2359	0.6991
45	Et50	0.1203	Cal	0.1746	0.6276	cal	0.1850	0.6680
46	Fp50	0.4076	Cal	0.4769	0.7793	val	0.4395	0.3212
47	Fq25	0.8910	Cal	0.7505	- 1.5878	cal	0.6966	- 1.9694
48	Fq70	0.0154	Val	0.1359	1.3665	val	0.1320	1.1712
49	Fq85	- 0.2765	Val	- 0.0736	2.4042	cal	- 0.0262	2.5997
50	Fr50	0.2876	Val	0.3322	0.4915	val	0.3203	0.3222
51	Fs50	0.2028	Cal	0.2682	0.7329	cal	0.2657	0.6304
52	Ft50	0.1617	Cal	0.1991	0.4334	cal	0.2075	0.4740
53	Gp50	0.7860	Cal	0.7055	- 0.9555	cal	0.6474	- 1.4682
54	Gq50	0.7049	Val	0.6509	- 0.6249	val	0.6127	- 0.9534
55	Gr50	0.6309	Cal	0.5742	- 0.6541	cal	0.5408	- 0.9293
56	Gs50	0.5310	Cal	0.5092	- 0.2558	cal	0.4863	- 0.4690
57	Gt50	0.4811	Val	0.4435	- 0.4553	val	0.4339	- 0.5116
58	Hp50	0.2341	Val	0.3360	1.1333	cal	0.3198	0.8523
59	Hq15	0.8393	Cal	0.7596	- 0.9140	cal	0.7052	- 1.3708
60	Hq25	0.6669	Val	0.6297	- 0.4157	val	0.5912	- 0.7566
61	Hq50	0.1831	Cal	0.2698	0.9474	val	0.2545	0.6967
62	Hq70	- 0.0994	Cal	0.0103	1.2285	cal	0.0144	1.1307
63	Hq85	- 0.2294	Cal	- 0.2040	0.2972	cal	- 0.1823	0.4845
64	Hr50	0.1308	Cal	0.1999	0.7550	val	0.1948	0.6248
65	Hs50	0.0630	Cal	0.1329	0.7752	cal	0.1400	0.7628
66	Ht50	0.0265	Val	0.0594	0.3774	val	0.0756	0.5026

Y_{Exp} et Y_{Pred} désignent respectivement les valeurs expérimentales et prédites de log k ; e_{istd} est l'erreur standard de prédiction

$\hat{y}_{(i)}$ désignant la réponse de l'i-ème objet estimée en utilisant un modèle obtenu sans faire intervenir cet i-ème objet, et $\bar{y}$ la valeur moyenne des n observations ; la somme porte sur l'ensemble des composés de calibrage.

Une valeur $Q^2_{LOO} > 0,5$ est, en général, considérée comme satisfaisante, et une valeur $Q^2_{LOO} > 0,9$ est excellente [16].

Dans la technique de validation par bootstrap [17] on simule de nouveaux échantillons, de taille n, par tirages aléatoires avec remise. De cette façon, l'ensemble de calibrage, qui conserve sa taille initiale n, se compose en général d'objets répétés, l'ensemble de test rassemblant les objets exclus. Le modèle est calculé sur l'ensemble de calibrage et les réponses prédites pour l'ensemble d'évaluation. Tous les carrés des différences entre valeurs prédites et réelles des éléments de l'ensemble d'évaluation sont collectés dans le PRESS. Cette procédure est répétée 3000 fois dans cette étude, les PRESS sont additionnés, et une capacité de prédiction moyenne calculée.

L'application du modèle, calculé sur l'ensemble de calibrage, aux composés de l'ensemble de test, permet de vérifier de manière fiable la capacité prédictive du modèle obtenu.

L'équation (7) permet le calcul de Q^2_{ext} :

$$Q^2_{ext} = 1 - \frac{\sum_{1}^{n_{ext}} \left(\hat{y}_{(i)} - y_i\right)^2 / n_{ext}}{\sum_{1}^{n_{tr}} \left(y_i - \bar{y}_{tr}\right)^2 / n_{tr}} = 1 - \frac{PRESS/n_{ext}}{SCT/n_{tr}} \quad (7)$$

L'indice (ext) se rapporte aux objets de l'ensemble de validation externe (ou à ceux de l'ensemble d'évaluation obtenu par bootstrap), et l'indice (tr) à ceux de l'ensemble de calibrage (training set).

Avec R^2, le paramètre $EQMP_{ext}$ est également utile. On le calcule selon l'équation

$$EQMP_{ext} = \sqrt{\frac{1}{n_{ext}} \sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_i\right)^2} \quad (8)$$

la somme portant sur les objets de l'ensemble de test (n_{ext}).
La détection des observations aberrantes a été basée sur la valeur du résidu de prédiction standardisé :

$$e_{i_{std}} = \frac{e_{(i)}}{\sqrt{S^2_{(i)}\left(1 - h_{ii}\right)}} \quad (9)$$

pour lequel l'estimation $S^2_{(i)}$ de σ^2 est calculée selon [18]:

$$S^2_{(i)} = \frac{(n-p)\, CME - e_i^2 / (1 - h_{ii})}{n - p - 1} \quad (10)$$

pour (n – 1) observations, la i[ème] étant exclue ; $CME = \sum_1^n \left(y_i - \hat{y}_{(i)}\right)^2 / (n - p)$, est le carré moyen des écarts ; p est le nombre de régresseurs (3, dans le cas présent).

3. RESULTATS ET DISCUSSION

Les compositions des ensembles de calibrage (cal) et de test (val) obtenues soit par choix aléatoire, soit par exploitation du programme DUPLEX écrit dans notre laboratoire à l'aide de MATLAB 7 [19], sont présentées dans le (tableau. 2).

Les équations des modèles en (φ, T, M log P), calculés sur chacun des ensembles de calibrage ainsi obtenus, à l'aide des valeurs centrées – réduites des régresseurs, sont :

3.1. Choix aléatoire :

$$\log k = 0{,}172\ (\pm\ 0{,}015) - 0{,}088\ (\pm\ 0{,}015)\ T - 0{,}246\ (\pm\ 0{,}015)\ \ddot{} + 0{,}231\ (\pm\ 0{,}015)\ M\ \log P \quad (11)$$

3.2. Choix par algorithme Duplex :

$$\log k = 0{,}149\ (\pm\ 0{,}016) - 0{,}076\ (\pm\ 0{,}017)\ T - 0{,}250\ (\pm\ 0{,}017)\ \ddot{} + 0{,}217\ (\pm\ 0{,}017)\ M\ \log P \quad (12)$$

Les régresseurs φ et M log P qui sont les plus corrélés avec log k, contrôlent notablement celui-ci, comme le montrent les valeurs de leurs coefficients dans les 2 équations précédentes, et du t Student associé (tableau 3).

Tableau 3 : Importance des régresseurs (X) dans le modèle ; r (X, log k) est le coefficient de corrélation régresseur – log k.

Choix de l'ensemble de calibrage	X	r (X, log k)	p	Cœfficient	*t*	p
Aléatoire	T	- 0.229	0.156	- 0.088	- 5.88	0.000
	φ	- 0.717	0.000	- 0.246	- 16.47	0.000
	MLOGP	0.569	0.000	0.231	15.41	0.000
Algorithme DUPLEX	T	- 0.204	0.206	- 0.076	- 4.57	0.000
	φ	- 0.697	0.000	- 0.250	- 14.96	0.000
	MLOGP	0.569	0.000	0.217	12.92	0.000

Nous avons pu, pour chacun des 2 modèles, vérifier de façon empirique (distribution des résidus standardisés en fonction des valeurs ajustées) la constance des variances σ^2, c'est-à-dire leur indépendance des régresseurs et de la variable dépendante ajustée. Le diagramme des scores normaux fait ressortir un petit défaut de normalité (pour un niveau de signification α = 0,05) dans le cas du choix aléatoire de l'ensemble de calibrage : R = 0,9679 < R_c = 0,9715, pour R = 0,9753 dans le cas du choix par algorithme DUPLEX. La valeur de la statistique de Durbin – Watson (d = 1,80), dans le cas du choix par algorithme DUPLEX, est plus grande que la valeur supérieure donnée par les tables pour 3 régresseurs, et pour tout risque raisonnable α, ce qui établit l'indépendance des résidus.
Pour le choix aléatoire, la valeur d = 1,43 qui est comprise entre les valeurs inférieure et

supérieure données par les tables, ne permet pas de conclure à l'indépendance des résidus. Ainsi, le modèle calculé sur l'ensemble de calibrage choisi aléatoirement, contrairement à celui correspondant au choix par algorithme DUPLEX, ne vérifie pas l'hypothèse d'un modèle statistique linéaire à effets fixes.

Les diagnostiques statistiques réunis dans le tableau 4 permettent de faire des comparaisons et de tirer plusieurs conclusions.

Tableau 4 : Diagnostiques statistiques pour les modèles calculés.

Choix	R^2	Q^2_{LOO}	Q^2_{boot}	Q^2_{ext}	$R^2_{ajusté}$
Aléatoire	94.06	92.31	91.23	89.19	93.56
DUPLEX	91.45	89.08	87.68	90.85	90.74
Choix	*EQMC*	*EQMP*	$EQMP_{ext}$	*F*	*SE*
Aléatoire	0.088	0.1	0.119	189.97	0.0931
DUPLEX	0.099	0.112	0.102	128.36	0.1042

Les valeurs de R^2et de $R^2_{ajusté}$ montrent, à chaque fois, la qualité de l'ajustement, alors que les faibles différences entre R^2et Q^2_{LOO} renseignent sur la robustesse des modèles qui sont, en outre, très hautement significatifs (valeurs élevées de la statistique F de Fisher). De plus, la similitude de $EQMC$ et $EQMP$ signifie que les capacités de prédiction internes des modèles ne sont pas trop dissemblables de leurs pouvoirs d'ajustement.

La validation par bootstrap $\left(Q^2_{boot}\right)$ confirme tout à la fois la capacité de prédiction interne et la stabilité des modèles.

La validation statistique externe $\left(Q^2_{ext}\ ;\ EQMP_{ext}\right)$ atteste de la bonne capacité prédictive des log k n'ayant pas servi au calcul des modèles. Notons que le modèle calculé sur l'ensemble de calibrage obtenu par application de l'algorithme DUPLEX est à accréditer des meilleures performances $\left(Q^2_{ext} > Q^2_{LOO}\ ;\ EQMP_{ext} < EQMP\right)$.

Les données caractérisées par des résidus de prédiction standardisés (colonnes e_{istd} du tableau 2) supérieurs, en valeur absolue, à 3 unités d'écart type (3 σ) sont aberrantes. Il en est ainsi des données Bq85, Dq15, de l'ensemble de calibrage, et Cq85, de l'ensemble de test, choisis aléatoirement. Par contre, pour le choix des ensembles par algorithme DUPLEX aucune donnée aberrante n'est détectée pour le modèle.

4. CONCLUSION

Nous avons relié simplement les logarithmes des paramètres de rétention (log k) de phénols différemment substitués, séparés en régime isochratique par CLHP – PI sur une colonne Partisil ODS, à : la température (T) de la colonne, la fraction volumique (φ) du méthanol dans la phase mobile hydro-organique binaire, et le coefficient de partage n-octanol / eau. Différentes approches pour le calcul de ce descripteur moléculaire caractéristique de la lipophilie étant disponibles, nous avons adopté celle de Moriguchi (M log P), dans la version DRAGON de Todeschini, qui conduit aux meilleurs résultats. Les régresseurs M log P et, notablement, φ contrôlent log k.

Contrairement au choix aléatoire, l'utilisation de l'algorithme DUPLEX conduit à un ensemble de calibrage permettant le calcul d'un modèle multilinéaire en (T, φ, M log P) qui vérifie les hypothèses d'un modèle statistique linéaire à effets fixes, et pour lequel aucun point aberrant n'est observé. Les différentes statistiques calculées confirment la validité du modèle ainsi obtenu, qui pourra être avantageusement utilisé comme alternative à la méthode par approximations successives appliquée pour l'optimisation des conditions chromatographiques de séparation.

REFERENCES

[1] Zapala W., Kaczmarski K., Kowalska T., 2002. Comparison of different retention models in normal – and reversed - phase liquid

chromatography with binary mobile phase. Journal of Chromatographic Science. Sci. 40. Pp. 575 – 580.

[2] Ko V., Ford J.C., 2001.comparison of selected retention models in reversed-phase liquid chromatography, Journal of Chromatography A, 913. Pp. 3-13.

[3] Melander W.R., Horváth C., 1980. Reversed - Phase Chromatography. In High Performance Liquid Chromatography - Advances and Perspectives. C. Horváth, ed. Academic Press. Vol. 2.

[4] Hansch C., 1969. Quantitative approach to biochemical structure – Activity relationships. *Accounts of chemical research,* 2.pp. 232 – 239.

[5] Leo A., Hansch C., Elkins D. 1971, Partition Coefficients and their Uses. *Chemical Reviews*, 71. Pp. 525 – 616.

[6] Todechini R., Consonni V., Pavan M., 2006. DRAGON, Software for the calculation of molecular Descriptors. Release 5.4 for Windows, Milano.

[7] Timmermans J. 1960, The Physico-Chemical Constant of Binary Systems Solution. Vol. 4. Wiley Interscience, New York.

[8] Kaliszan R., 1987. Quantitative Structure – Chromatographic Retention Relationships. Wiley Interscience. New York.

[9] Moriguchi I., Hirono S., Liu Q., Nakagome I., Matsushita Y, 1992. Simple Method of Calculating Octanol / water Partition Coefficient. Chemical and pharmaceutical bulletin, 40. pp. 127 – 130.

[10] Moriguchi I., Hirono S., Nakagome I., Hirono H. 1994, Comparison of Reliability of Log P Values for Drugs Claculated by Several methods. Chemical and pharmaceutical bulletin, 42. Pp. 976 – 978.

[11] Todeschini R., Ballabio D., Consonni V., Mauri A. ; Pavan M, , 2009. Mobydigs Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 1.1 for Windows, Milano.

[12] Snee R.D, 1977. Validation of regression models: Methods and examples. Technometrics, 19. pp. 415-428.

[13] Minitab, 2003.Release 14.1, statistical software.

[14] Graybill F.A, 1976. Theory and Application of the Linear Model, Duxbury, North Scituate, Mass. pp. 231 – 236.

[15] Todeschini R., Consonni, V. 2009. Molecular descriptor for the chemoinformatics, Vol. 1. WILEY-V c H. Verlog. Gmb, H et Co-KGaA, Weinheine.

[16] Eriksson L., Jaworska J., Worth A., Cronin M., Mc Dowell R.M., Gramatica P., 2003. Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs. Environ. Health Persp., 111 (10). Pp. 1361 – 1375.

[17] Efron P., 1982. The Jacknife, The Bootstrap and Other Resampling Planes. Society for Industrial and Applied Mathematics, Philadelphia (PA), 92p.

[18] Montgomery D.C., Peck E.A., 1992, Introduction to linear regression analysis, 2nd edition, Wiley Interscience, New York.

[19] Matlab, 2004. Version 7.0.0.19920 (Release 14). The language of technical computing. The Math Works, Inc. May 06.