

有机磷酸酯类化合物定量结构-色谱保留关系及稳健性分析

赵劲松* * 于书霞

华中农业大学资源与环境学院/农业部亚热带农业资源与环境重点开放实验室,武汉 430070

摘要 以分子拓扑指数作为结构描述符,采用最佳子集回归方法,建立了 35 种有机磷酸酯类化合物在 3 种不同极性固定相上的定量结构-色谱保留关系(QSRR)模型。3 个 QSRR 模型均具有良好的拟合能力,对 QSRR 模型分别进行了交叉验证及外部数据集验证,结果表明各模型具有较强的预测能力。对 QSRR 模型的系数、标准误差及相关系数均进行了蒙特卡洛模拟,结果证实蒙特卡洛方法可用于 QSRR 模型的稳健性分析。

关键词 有机磷化合物; 定量结构-色谱保留关系; 拓扑指数; 蒙特卡洛模拟

中图分类号 O 641 文献标识码 A 文章编号 1000-2421(2010)02-0164-05

随着 20 世纪 70 年代对持久性有机氯杀虫剂的逐渐禁用,高效低残留的有机磷酸酯类化合物逐渐成为杀虫剂的首选,是迄今应用最广泛的一类农药。有机磷酸酯类化合物的广泛应用所带来的环境问题已不容忽视^[1-6]。在有机化合物的生态风险评价中,准确鉴定每一个化合物的结构有着非常重要的意义。色谱保留指数是应用色谱技术定性鉴定有机化合物的重要手段^[7],然而,有关有机磷酸酯类化合物定量结构-色谱保留关系(QSRR)的研究仍然非常少^[1,7]。

由于 QSRR 模型通常用于预测尚未投入实验或未合成的化合物的色谱保留指数,因此,模型的预测能力及稳健性非常重要。交叉验证是 QSRR 研究中广泛应用的模型验证方法,采用相关系数 Q^2 表征模型的预测能力^[8-9],然而,模型的稳健性目前还缺少有效的表征方式。蒙特卡洛(Monte Carlo)模拟基于特定概率分布进行随机抽样,模拟可能出现的随机现象,通过大量的随机抽样,可接近所模拟的随机过程的真实概率。因此,可用来评价模型的不确定性^[10-11]。随着计算机技术的快速发展,蒙特卡洛模拟在有机污染物的风险评价中已有应用^[11],但未见其用于 QSRR 模型稳健性的评价。

笔者采用广泛应用的分子拓扑指数作为结构描述符,预测 35 种有机磷酸酯类化合物在 3 种不同极

性固定相上的气相色谱 Kov á s 保留指数,并探索蒙特卡洛模拟方法在 QSRR 模型稳健性分析中的应用。

材料与方 法

数据来源及分子拓扑指数

35 种有机磷酸酯类化合物(图 1)在不同极性固定相(OV-101、DB-1701 以及 DB-WAX)上的气相色谱 Kov á s 保留指数取自参考文献[12](表 1)。35 种有机磷酸酯类化合物在结构与极性上有较大的差异,苯环上的取代基(X)的电负性变化较大,而取代基 R 的体积变化较大。

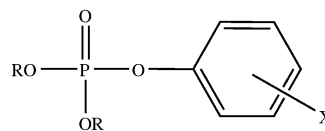


图 1 OP 的结构通式,R 为烷基,X 为苯环上的取代基

Fig. 1 The general formula of OP compounds,R is alkyl, and X is substituent group on phenyl

根据文献方法^[13-14]计算了 14 种广泛应用的分子拓扑指数,包括 Wiener 指数(W),0-3 阶 Randić 指数(n),0-3 阶 Kier & Hall 指数(n'),1-3 阶 Kier 形状指数(n''),Kier 柔性指数(n'''),以及 Balaban 指数(J)(表 1)。

收稿日期:2009-03-04; 修回日期:2009-11-18

* 国家自然科学基金项目(40601085)、国家科技支撑计划项目(2008BADA7B01)和华中农业大学科研启动基金(52204-05051)资助

* * 通讯作者。E-mail: jszhao@mail.hzau.edu.cn

赵劲松,男,1978 年生,副教授。研究方向:有机污染化学及生态风险评价。E-mail: jszhao@mail.hzau.edu.cn

表 1 35 种有机磷酸酯类化合物在不同固定相上的气相色谱保留及分子拓扑指数¹⁾
Table 1 The topological indices and retention index on different polar stationary phase of 35 OP compounds

No.	X	OV-101	DB-1701	DB-WAX	W	$^0\chi$	$^1\chi$	$^2\chi$	$^3\chi$	$^0\chi^v$	$^1\chi^v$	$^2\chi^v$	$^3\chi^v$	$^1\kappa$	$^2\kappa$	$^3\kappa$	Φ	J	
R=CH ₃																			
1	H	1 420	1 680	2 140	254	9.734	6.200	5.226	4.213	8.361	5.122	4.024	2.736	10.282	4.44	2.862	3.512	2.738	
2	3-CH ₃ *	1 488	1 758	2 219	310	10.604	6.593	5.860	4.534	9.284	5.533	4.527	2.983	11.276	4.62	3.111	3.721	2.825	
3	4-CH ₃	1 504	1 768	2 245	317	10.604	6.593	5.848	4.624	9.284	5.533	4.524	3.014	11.276	4.62	3.111	3.721	2.777	
4	4-OCH ₃	1 637	1 840	2 350	394	11.312	7.131	6.017	5.032	9.692	5.645	4.386	3.053	12.231	5.301	3.338	4.322	2.762	
5	3-Cl	1 560	1 816	2 300	310	10.604	6.593	5.860	4.534	9.418	5.600	4.605	3.025	11.564	4.823	3.274	3.984	2.825	
6	4-Cl	1 574	1 830	2 314	317	10.604	6.593	5.848	4.624	9.418	5.600	4.601	3.059	11.564	4.823	3.274	3.984	2.777	
7	3-Br	1 652	1 910	2 392	310	10.604	6.593	5.860	4.534	10.248	6.015	5.084	3.283	11.753	4.957	3.382	4.162	2.825	
8	4-Br	1 666	1 924	2 410	317	10.604	6.593	5.848	4.624	10.248	6.015	5.080	3.336	11.753	4.957	3.382	4.162	2.777	
9	3-CN	1 678	2 003	2 625	380	11.312	7.131	6.029	4.957	9.231	5.506	4.351	2.956	11.763	4.965	3.083	3.893	2.899	
10	4-CN*	1 706	2 041	2 662	394	11.312	7.131	6.017	5.032	9.231	5.506	4.347	2.977	11.763	4.965	3.083	3.893	2.818	
11	3-NO ₂	1 795	2 065	2 720	452	12.182	7.504	6.759	5.255	9.548	5.622	4.465	3.022	12.668	5.088	3.286	4.029	2.978	
12	4-NO ₂	1 810	2 085	2 737	473	12.182	7.504	6.747	5.323	9.548	5.622	4.462	3.041	12.668	5.088	3.286	4.029	2.873	
R=C ₂ H ₅																			
13	H	1 502	1 756	2 210	384	11.148	7.200	6.019	4.213	9.776	6.297	4.280	3.118	12.27	5.931	4.058	4.852	2.707	
14	4-CH ₃	1 617	1 847	2 310	465	12.019	7.593	6.641	4.624	10.698	6.708	4.780	3.396	13.266	6.064	4.288	5.027	2.748	
15	3-Cl	1 667	1 891	2 390	456	12.019	7.593	6.653	4.534	10.832	6.775	4.861	3.406	13.555	6.281	4.469	5.321	2.789	
16	4-Cl*	1 675	1 912	2 410	465	12.019	7.593	6.641	4.624	10.832	6.775	4.858	3.441	13.555	6.281	4.469	5.321	2.748	
17	3-Br	1 777	1 996	2 490	456	12.019	7.593	6.653	4.534	11.662	7.190	5.340	3.664	13.744	6.424	4.589	5.518	2.789	
18	4-Br*	1 790	2 010	2 510	465	12.019	7.593	6.641	4.624	11.662	7.190	5.337	3.717	13.744	6.424	4.589	5.518	2.748	
19	3-CN	1 780	2 096	2 725	544	12.726	8.132	6.822	4.957	10.646	6.682	4.607	3.338	13.754	6.431	4.215	5.203	2.860	
20	4-CN	1 800	2 140	2 762	562	12.726	8.132	6.810	5.032	10.646	6.682	4.604	3.358	13.754	6.431	4.215	5.203	2.790	
21	3-NO ₂	1 880	2 264	2 820	634	13.596	8.504	7.551	5.255	10.962	6.797	4.722	3.403	14.660	6.511	4.403	5.303	2.934	
22	4-NO ₂	1 895	2 284	2 838	661	13.596	8.504	7.540	5.323	10.962	6.797	4.718	3.422	14.660	6.511	4.403	5.303	2.841	
23	2,4-Cl	1 837	2 031	2 490	532	12.889	8.004	7.170	5.071	11.889	7.258	5.386	3.778	14.840	6.643	4.529	5.799	2.893	
24	2,5-Cl	1 844	2 044	2 502	524	12.889	8.004	7.170	5.087	11.889	7.258	5.386	3.794	14.840	6.643	4.529	5.799	2.927	
R=C ₄ H ₉																			
25	H	1 889	2 103	2 596	784	13.977	9.200	7.433	5.274	12.604	8.297	5.819	3.887	16.255	9.159	6.785	7.836	2.591	
26	3-CH ₃ *	2 025	2 202	2 614	894	14.847	9.594	8.067	5.594	13.527	8.708	6.322	4.133	17.252	9.194	6.958	7.931	2.672	
27	4-CH ₃	2 047	2 224	2 630	907	14.847	9.594	8.055	5.685	13.527	8.708	6.319	4.165	17.252	9.194	6.958	7.931	2.644	
28	4-OCH ₃	2 183	2 239	2 861	1050	15.554	10.132	8.224	6.093	13.935	8.820	6.181	4.203	18.210	9.982	7.115	8.655	2.665	
29	3-Cl*	2 065	2 270	2 730	894	14.847	9.594	8.067	5.594	13.661	8.775	6.399	4.175	17.541	9.431	7.165	8.271	2.672	
30	4-Cl	2 085	2 295	2 774	907	14.847	9.594	8.055	5.685	13.661	8.775	6.396	4.209	17.541	9.431	7.165	8.271	2.644	
31	4-Br	2 190	2 417	2 810	907	14.847	9.594	8.055	5.685	14.491	9.190	6.875	4.486	17.731	9.586	7.301	8.499	2.644	
32	3-CN	2 215	2 515	2 989	1 024	15.554	10.132	8.236	6.018	13.474	8.682	6.145	4.107	17.741	9.595	6.790	8.105	2.742	
33	4-CN	2 228	2 552	3 020	1 050	15.554	10.132	8.224	6.093	13.474	8.682	6.142	4.127	17.741	9.595	6.790	8.105	2.693	
34	3-NO ₂ *	2 326	2 637	3 112	1 156	16.424	10.504	8.966	6.315	13.790	8.797	6.260	4.172	18.648	9.583	6.929	8.123	2.814	
35	4-NO ₂	2 345	2 674	3 180	1 195	16.424	10.504	8.954	6.383	13.790	8.797	6.256	4.191	18.648	9.583	6.929	8.123	2.745	

1) * 用于外部验证的数据集。Data set for external validation.

统计方法

采用最佳子集线性回归方法建立 QSRR 模型。对所有描述符的组合进行遍历搜索,并去除方差膨胀因子大于 10 的描述符组合,以消除共线性对 QSRR 模型质量的影响^[15]。为了更加准确地评价模型的预测能力,将全部数据随机分为训练集与检验集。训练集有 28 个样本,用于建立 QSRR 模型和内部验证。模型拟合能力用相关系数 R^2 、拟合标准误差 SE 表征;模型内部预测能力由逐一剔除交叉验证的相关系数 Q_{cv}^2 ,预测标准误差 SE_{cv} 表征。检验集为 7 个样本,用于检验模型对外部数据的预测能力,用 Q_{ext}^2 和 SE_{ext} 表征。

蒙特卡洛模拟

本研究中, QSRR 模型可由 $y = Xb + e$ 表示,其中 y 为 RI 向量, X 为描述符矩阵, b 为回归系数向量, e 为残差向量且 $e \sim N(0, \sigma^2)$, 其中 σ 为模型残差的标准偏差, 样本容量为 n 。对 QSRR 模型的蒙特卡洛模拟过程如下^[16]: (1) 从 $N(0, \sigma^2)$ 中生成 n 个随机数, 记作 $e_{(b)}$; (2) 通过式 $y_{(b)} = Xb + e_{(b)}$ 得到 $y_{(b)}$; (3) 对 $y_{(b)}$ 和 X 进行线性回归, 求得 QSRR 模型相关参数的蒙特卡洛估计; (4) 重复步骤 (1) 至步骤 (3) B 次; (5) 根据模拟参数蒙特卡洛估计的分布, 判断模型的稳健性。通常模拟次数越大, 其结果越接近于真实值, 在这里 B 取 9 999。所有的统计分析均采用统计计算的语言与环境 R^[17] 完成。

结果与分析

模型的建立

对 3 种不同极性固定相上的气相色谱 Kováts 保留指数 RI 分别与 14 种拓扑指数描述符进行最

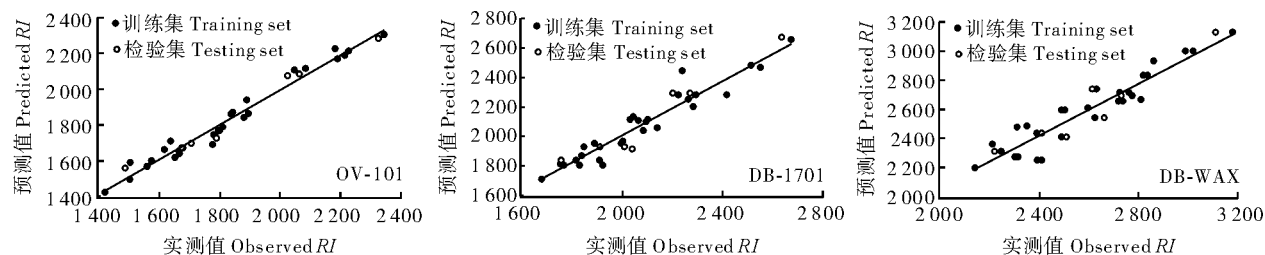


图 2 不同固定相上色谱保留指数实测值与预测值的散点图

Fig. 2 Scatter plot of observed versus predicted retention index on different stationary phase

在本研究中, OV-101 是弱极性固定相, 因此, 有机磷酸酯类化合物在该固定相中的保留行为主要是由分子的立体效应所决定^[7], 3 和 3 在一定程度上可以较好地描述这种空间效应^[14], 因此给出了较

佳子集线性回归, 结果表明, 所有 QSRR 模型中均只含有 2 个描述符。在不同固定相上的 QSRR 模型分别如下所示(括号中为回归系数的标准误差):

OV-101 (固定相为 100% 甲基硅酮):

$$RI = -277.703 (\pm 70.904) + 278.394 (\pm 21.702)^3 + 192.421 (\pm 26.286)^3 \quad (1)$$

$n = 28, R^2 = 0.973, SE = 39.870, Q_{cv}^2 = 0.966,$
 $SE_{cv} = 44.914, Q_{ext}^2 = 0.971, SE_{ext} = 44.697$

DB-1701 (固定相为 86% 二甲苯 + 14% 丁氰基-苯基取代聚氧硅烷):

$$RI = -818.366 (\pm 499.707) + 0.999 (\pm 0.061) W + 830.208 (\pm 172.780) J \quad (2)$$

$n = 28, R^2 = 0.917, SE = 72.077, Q_{cv}^2 = 0.895,$
 $SE_{cv} = 81.194, Q_{ext}^2 = 0.920, SE_{ext} = 74.882$

DB-WAX (固定相为 100% 聚氧乙烯):

$$RI = 216.810 (\pm 194.175) + 418.956 (\pm 42.429)^1 - 215.205 (\pm 35.495)^3 \quad (3)$$

$n = 28, R^2 = 0.887, SE = 88.619, Q_{cv}^2 = 0.860,$
 $SE_{cv} = 98.661, Q_{ext}^2 = 0.892, SE_{ext} = 84.968$

统计结果表明, RI 与进入模型描述符之间存在良好的线性关系, 表现在模型具有较高的 R^2 。模型的内部交叉验证以及对外部数据集的预测结果均表明, 模型具有较强的预测能力, 表现在交叉验证的 Q_{cv}^2 和外部验证的 Q_{ext}^2 与模型的 R^2 相近, SE_{cv} 和 SE_{ext} 与 SE 相近。

模型的解释

从 QSRR 模型的 R^2 、 Q_{cv}^2 、 Q_{ext}^2 的变化趋势可以看出, 当固定相极性由弱变强时, QSRR 模型的相关性及预测能力均呈下降趋势。从保留指数的实测值与 QSRR 模型的预测结果的相关图(图 2)上也可以清晰地看到这种趋势。

好的模拟结果。

对于中等极性的 DB-1701 和极性的 DB-WAX 固定相而言, 有机磷酸酯类化合物在固定相中的保留机制不仅受化合物立体效应的影响, 同时也与化

合物与固定相之间的相互作用,如偶极效应等有着密切关系^[7]。进入 QSRR(2)的 W 和 QSRR(3)的 1 主要表征有机磷酸酯类分子的立体效应^[14];而 J 和 3 由于考虑了分子中非碳原子的种类、环的存在及环上取代基的位置等^[14],可以间接描述有机磷酸酯类化合物与固定相之间的非立体效应的相互作用,但效果有限。因此,所获得的 QSRR 模型相关性与预测能力均较弱极性固定相上的 QSRR 模型稍差。

模型稳健性分析

蒙特卡洛模拟是不确定分析中最常用的一种概率分析方法,即从特定概率分布中随机生成大量伪样本来模拟给定问题的概率统计模型,给出问题数值解的估计^[10,16]。由于不受样本容量的限制,因此,适合类似 QSRR 研究中小样本回归分析的稳健性分析^[16]。表 2 给出了对上述 3 个 QSRR 模型进

行蒙特卡洛模拟的结果。从表中可以看出,蒙特卡洛估计与模型参数非常接近,表明有机磷酸酯类化合物在不同固定相上的气相色谱保留指数的 QSRR 模型是稳健的。

为了更清晰地说明蒙特卡洛模拟对 QSRR 模型稳健性的评价,以 QSRR(1)为例,将 3 系数的蒙特卡洛模拟结果在图 3 中给出。其中,左侧图给出了模拟获得的 3 系数的估计分布及概率密度。从图中可以看出模型的 3 的系数(278.394,实线)与模拟获得的 3 的系数(278.479,虚线)之间几乎完全重叠。从图 3 右侧的正态分布 Q-Q 图可以看出,模拟获得的 3 的系数呈正态分布。由于 B 足够大,模拟获得的 3 的系数分布可以作为模型真实情况的估计^[16]。对模型各个统计量的蒙特卡洛模拟均可获得相似的结果。因此,各个固定相上的 QSRR 模型均具有良好的稳健性。

表 2 3 个 QSRR 模型的系数的蒙特卡洛估计¹⁾

Table 2 Monte Carlo simulated coefficient of three QSRR models

	QSRR(1)		QSRR(2)		QSRR(3)
Intercept	- 277.406 (±70.279)	Intercept	- 824.622 (±500.167)	Intercept	214.417 (±192.684)
3	278.479 (±21.488)	W	0.999 (±0.061)	1	419.655 (±42.220)
$^3 v$	195.249 (±26.205)	J	832.475 (±172.969)	3	- 215.841 (±35.442)
R^2	0.973 (±0.008)	R^2	0.918 (±0.023)	R^2	0.887 (±0.031)
SE	39.442 (±5.592)	SE	71.288 (±10.134)	SE	87.663 (±12.425)

1) 括号中为统计量的标准偏差。Standard deviation of statistic in parentheses.

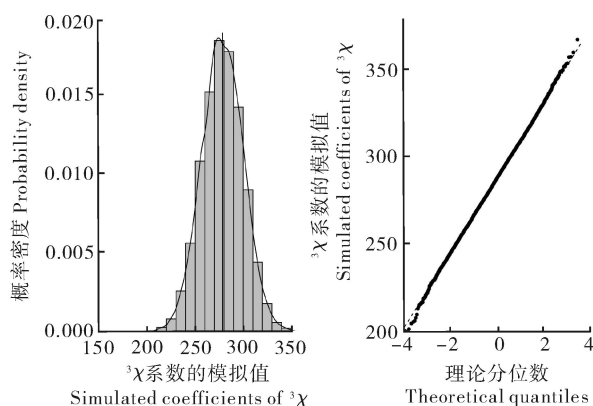


图 3 QSRR (1) 模型 3 系数的蒙特卡洛模拟

Fig.3 Monte Carlo simulated coefficients of 3 in QSRR (1) model

讨论

以分子拓扑指数为分子结构描述符,建立了 35 种有机磷酸酯类化合物在 3 种不同极性固定相上的气相色谱 Kovás 保留指数的 QSRR 模型。所得 QSRR 模型在弱极性固定相上具有良好的相关性及预测能力,在极性固定相上则相对较弱。其主要原

因是拓扑指数描述符通常不具有表征化合物电子结构的能力,而化合物在固定相中的分离不仅仅与立体效应相关^[7]。基于拓扑指数的非极性化合物的 QSRR 模型通常具有较高的拟合和预测能力,而对于极性化合物,加入表征化合物电子结构或表征与固定相相互作用的描述符则可以改善模型的质量^[7,18]。

虽然交叉验证 Q_{cv}^2 通常用于表征 QSRR 的预测能力,但它只是表征 QSRR 模型预测能力的必要而非充分条件^[8],因此,加入模型对外部检验数据集的预测是非常必要的^[8-9]。本研究通过随机抽样的方式获得 20% 的样本用于 QSRR 模型的外部验证。由于外部数据集的选择是随机的,因此其验证能力仍存在不确定性^[9]。但对于本数据集而言,内部与外部的验证结果均表明 QSRR 模型具有良好的预测能力。

基于多元线性回归的建模方法受数据集本身的特征影响显著,其参数估计的稳健性是关系模型质量的重要因素^[15]。蒙特卡洛方法通过对残差的抽样构造新数据集^[16],对 QSRR 模型的系数、标准误

差、相关系数进行模拟估计,很好地检验了 QSRR 模型。因此,在 QSRR 模型的稳健性评价中应用蒙特卡洛方法是可行的。

参 考 文 献

- [1] KNAACK J B, DARY C C, POWER F, et al. Physicochemical and biological data for the development of predictive organophosphorus pesticide QSARs and PBPK/ PD models for human risk assessment [J]. *Crit Rev Toxicol*, 2004, 34(2) :143-207.
- [2] ZHAO J S, WANG B, DAI Z X, et al. 3D-quantitative structure-activity relationships study of organophosphate compounds [J]. *Chin Sci Bull*, 2004, 49(3) :240-245.
- [3] CHELME A YALA P, LI X, NOUR M, et al. Pesticides and herbicides [J]. *Water Environ Res*, 2007, 79(10) :1766-1850.
- [4] COSTA L G. Current issues in organophosphate toxicology [J]. *Clin Chim Acta*, 2006, 366(1/2) :1-13.
- [5] POPE C N. Organophosphorus pesticides: do they all have the same mechanism of toxicity [J]. *Toxicol Environ Health*, B: *Crit Rev*, 1999, 2(2) :161-181.
- [6] 陶宏亮, 关燕萍, 苏晓峰, 等. SPE-HPLC 用于蔬菜中甲基对硫磷和对硫磷同时测定 [J]. *华中农业大学学报*, 2006, 25(1) :46-49.
- [7] KALISZAN R. QSRR: Quantitative structure-(chromatographic) retention relationships [J]. *Chem Rev*, 2007, 107(7) :3212-3246.
- [8] GOLBRAIKH A, TROPSHA A. Beware of q^2 ! [J]. *J Mol Graph Model*, 2002, 20(4) :269-276.
- [9] GRAMATICA P. Principles of QSAR models validation: internal and external [J]. *QSAR Comb Sci*, 2007, 26(5) :694-701.
- [10] MUN J. Modeling risk: applying monte carlo simulation, real options analysis, forecasting, and optimization techniques [M]. New York: John Wiley & Sons, 2006.
- [11] BOGEN K T, CULLEN A C, FREY H C, et al. Probabilistic exposure analysis for chemical risk characterization [J]. *Toxicol Sci*, 2009, 109(1) :4-17.
- [12] GANDHE B R, PURNANAND, PRASAD R, et al. Use of gas chromatographic retention indices for quantitative structure activity relationship studies of dialkyl phenyl phosphates [J]. *Pestic Sci*, 1990, 29(4) :379-385.
- [13] KATRITZKY A R, IGNATICHENKO E S, BARCOCK R A, et al. Prediction of gas chromatographic retention times and response factors using a general quantitative structure property relationship treatment [J]. *Anal Chem*, 1994, 66(11) :1799-1807.
- [14] ESTRADA E, URIARTE E. Recent advances on the role of topological indices in drug discovery research [J]. *Curr Med Chem*, 2001, 8(13) :1573-1588.
- [15] KLEINBAUM D G, KUPPER L L, Muller K E, et al. Applied regression analysis and other multivariable methods [M]. New York: Duxbury Press, 1998.
- [16] DAVISON A C, HINKLEY D V. Bootstrap methods and their application [M]. New York: Cambridge University Press, 1997.
- [17] R Development Core Team 2008. R: A language and environment for statistical computing [P]. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [18] KRAWCZUK A, VOELKEL A, LULEKJ, et al. Use of topological indices of polychlorinated biphenyls in structure retention relationships [J]. *J Chromatogr A*, 2003, 1018(1) :63-71.

Quantitative Structure-Retention Relationships of Organophosphate Compounds and Robust Analysis

ZHAO Jin-song YU Shu-xia

Key Laboratory of Subtropical Agriculture and Environment, Ministry of Agriculture/ College of Resources and Environment, Huazhong Agricultural University, Wuhan 430070, China

Abstract Using the best subset regression technique, the quantitative structure-retention relationships (QSRR) of 35 organophosphate compounds on three different polar stationaries were established with topological indices as structural descriptors. All three QSRR models had good calibration ability. The cross-validation and external-validation showed that the corresponding QSRR model had strong predictive power. Monte Carlo simulation on R^2 , SE and coefficients of different QSRR models demonstrated that Monte Carlo method can be used to analyze the robustness of QSRR models.

Key words organophosphate compounds; quantitative structure-retention relationships; topological indices; Monte Carlo simulation

(责任编辑:陆文昌)