

有机化合物 K_{oc} 片段常数估算模型的 误差与稳健性分析*

朴海善 陶 澍** 胡海瑛 卢晓霞 曹 军

(北京大学城市与环境学系, 北京 100871)

摘要 分析了估算有机化合物吸着系数的片段常数模型的误差. 并用 4 种不同方法检验了估算有机化合物吸着系数(K_{oc})的片段常数模型的稳健性. 重复随机抽取单一化合物、逐一抽取估算值与实测值差别较大的个别化合物、重复随机抽取 50 个化合物以及逐一抽取特定化合物类别. 检验结果表明, 模型估算的绝对误差随 $\log K_{oc}$ 实测结果增大而增大, 而误差较大的多为亲水化合物. 估算误差中的相当部分实际上来源于实测数据的误差波动. 无论是单一抽取或成组抽取化合物的 jackknife 检验结果证明模型具有较高的稳健性. 相对而言, 不同类别抽取的 jackknife 可决系数波动较大, 说明将模型推广到其他类别化合物时需特别谨慎.

关键词 吸着系数(K_{oc}), 片段常数法, jackknife 检验, 稳健性.

Robustness of A K_{oc} Prediction Model Based on Fragment Constants*

Piao Haishan Tao Shu Hu Haiying Lu Xiaoxia Cao Jun

(Department of Urban and Environmental Sciences, Peking University, Beijing 100871 China)

Abstract Modified jackknife tests were applied to examine the robustness of a prediction model for organic carbon normalized sorption coefficient (K_{oc}) based on fragment constants and structural correction factors. Similar trends were observed using various kinds of jackknife tests. Fairly high robustness of the model was demonstrated by the calculated jackknife coefficients of determination (R^2). The highest variation in the jackknife coefficient occurred when a group of chemicals within the same category was eliminated from the data set.

Keywords sorption coefficient K_{oc} , fragment constant, jackknife test, robustness.

有机化合物土壤/沉积物吸着系数(K_{oc})的估算既可以通过相关物理化学参数(亨利常数、溶解度和水/辛醇分配系数)进行^[1,2],也可以直接寻求它与某些表述分子结构的参数(如取代基常数、片段常数和分子连接性指数等)的定量关系^[3,4]. 为建立吸着系数回归模型,通常将已知实测值的化合物分为 2 组. 一组用于建模和回归参数计算,另一组用作模型验证. 通过随机安排的建模和验证分组,模型的可靠性得以检验. 除此之外,利用简单而有效的 jackknife 检验可对多元回归模型的稳健性作出定量判断. 进行 jackknife 检验时先按既定方式重复多次从数据中删除一个或一组测定值,并对每次删除后得到的数据组作回归分析. 在此基

础上比较删除前后以及多次删除之间拟合参数的差别,借以判断模型的稳健性. 常用的判断指标为多元可决系数(R^2)^[5].

本研究的目的是在已建立的预测有机化合物吸着系数 K_{oc} 的片段常数模型的基础上^[6],用不同形式的 jackknife 方法对模型稳健性进行检验.

1 研究方法

* 国家自然科学基金重点项目(Project Supported by the National Natural Science Foundation of China): 49632060 和国家杰出青年基金资助项目:(49525102)

** 通信联系人
朴海善:女,1972年生,1998年于北京大学城市与环境学系获硕士学位,现就读于美国辛辛提那大学. 主要研究方向为环境地球化学.
收稿日期:1998-12-04

1.1 K_{oc} 数据来源及模型建立

从文献中收集了 592 种有机化合物的 K_{oc} 实测数据. 对于文献中报道了 1 个以上 K_{oc} 测定值的化合物, 尽可能选取其中位数. 收集的化合物包含环境中常见的各类有机污染物, 如农药、有机染料、多环芳烃、多氯联苯以及酚类化合物等. 从 592 种化合物中随机抽取 430 种化合物用于建立模型(建模组), 剩余的 162 种化合物用于模型检验^[6]. 假定结构片段和有关校正因子对 $\log K_{oc}$ 的贡献具有简单的线性加和性, 可以建立以下多元线性回归模型:

$$\log K_{oc} = \sum_{i=1}^a n_i f_i + \sum_{j=1}^b m_j F_j \quad (1)$$

式中, a 和 b 分别是结构片段和校正因子种类总数. F_i 为第 i 种结构片段的片段常数; n_i 表示化合物中第 i 种结构片段的个数. F_j 代表第 j 种结构因子校正系数, 而 m_j 表示第 j 种结构因子的数目. 模型建立过程包括以下主要步骤:

(1) 根据结构片段和校正因子的初步划分和确定原则, 对建模组 430 种化合物进行片段和因子计数. 在此基础上用多元回归分析方法建立初步模型.

(2) 分析拟合结果, 对不合理的片段和校正因子进行修正. 同时采用逐步回归方法进一步分析校正因子的合理性. 重新拟合建模组实测数据.

(3) 根据验证组 162 种化合物的实测数据对上述回归模型进行验证, 借以对模型的可推广性进行分析.

(4) 在模型的合理性得以验证的前提下, 利用全部实测数据对模型进行最终修正, 得到估算模型和所有回归系数(结构片段常数和结构校正因子).

1.2 模型误差分析和稳健性检验

在重复实测数据波动的基础上讨论了模型的估计误差. 模型稳健性用修正的 jackknife 检验进行. 该方法能够检验个别或若干数据的存在(或去除)对模型的影响^[5, 6]. 本研究按 4 种方式进行检验: 单一化合物(随机)抽取、异常(K_{oc} 预测值与实测值之差大于 1.0 个对数单位者)化合物抽取、随机化合物组抽取以及特定

化合物类型抽取.

(1) 单一化合物抽取 从 592 种化合物中重复 100 次, 每次随机抽取 1 种, 对其余 591 种进行多元回归;

(2) 异常化合物抽取 对 34 种估算误差大于 1.0 个对数单位的化合物, 分 34 次每次抽取其一, 对余下 591 种化合物进行多元回归;

(3) 50 种化合物随机抽取 从 592 种化合物中重复 40 次每次随机抽取 50 种, 对其余 542 种化合物进行回归;

(4) 分类抽取 对研究中涉及的 17 类化合物逐类抽取, 对余下 16 类化合物进行多元回归.

根据 4 种形式的 jackknife 检验结果讨论了模型的稳健性.

1.3 数据整理与统计分析

用 Excel[®]建立了 K_{oc} 数据库. 包括 jackknife 检验在内的各类回归模型计算在 Statistica[®]下实现. 用 Turbo-Pascal 进行必要的数据预处理.

2 结果与讨论

2.1 预测有机化合物吸着系数 K_{oc} 的片段常数模型

根据 592 种化合物的实测 K_{oc} 数据, 建立了估算有机化合物 K_{oc} 的片段常数模型. 模型中自变量包括 74 种结构片段和 24 个结构校正因子. 模型的回归可决系数达 0.9696, 估算结果的平均误差为 0.37 个对数单位. 模型形式如式 1 所示, 结构校正因子和模型片段常数分别列于表 1 和表 2 中.

表 1 估算 K_{oc} 的片段常数模型的结构校正因子

结构因子	系 数	结构因子	系 数
F_b	0.0884	F_{p2}	0.4291
F_{b}	0.2560	F_{p3}	0.6357
F_{CBr}	- 0.1019	F_{p4}	- 0.0854
$F_{=}$	0.1143	F_{p5}	0.1658
$F_{=}$	0.1264	F_{p6}	0.0748
$F_{=}$	- 1.0277	F_{HP1}	- 0.1332
F_{mlG1}	0.0446	F_{HP2}	- 0.1074
F_{mlG2}	0.1314	F_C	- 1.0231
F_{mlG3}	- 0.0354	F_{CH}	- 0.7309
F_{mlw}	0.0085	F_{CH2}	- 0.3764
F_{p0}	0.5604	F_{NH}	0.4302
F_{p1}	0.3731	F_{NH2}	0.7345

表 2 估算 K_{oc} 的片段常数模型的片段常数

片段常数	f	F^{Φ}	$F^{1/\Phi}$	f^{Ψ}	片段数常	f	f^{Ψ}	$F^{1/\Phi}$	f^{Ψ}
不含 C 或 H					含 H, 不含 C				
—F	1.149	0.087			—H	1.487			
—Cl	0.523	0.439			—OH	— 0.300	— 0.176		
—Br	0.558	0.404			—OP(O)(NH)O—		— 1.715		
—N=	— 0.545	— 0.626		— 0.562	含 C, 含 H				
—O—	— 0.584	— 0.723		— 0.719	—C(O)H	— 1.109			
—S—	— 0.074	— 0.307			—C(O)OH	— 0.678	— 0.425		
—NO ₂		0.168			—C(O)NH—	— 1.406	— 0.534	— 0.875	— 2.515
—SO ₂ —	— 0.970	— 1.153			—C(O)NH ₂		— 0.315		
—S(O)—	— 0.709	— 1.133			—OC(O)NH—		— 0.837	— 0.600	— 0.522
—SP(S)(O—)S—	0.002				—C(O)ONH ₂		— 0.479		
—OP(O)(O—)O—	— 0.964				—CH=N—		— 1.738		
—OP(S)(O—)O—		— 0.456			—HNC(O)NH—		— 1.158		
—OS(O)O—		— 1.455			—HNC(O)NH ₂		— 0.219		
—P(S)(O—)O—	— 0.934				—CH=NOC(O)NH—	— 0.533			
—SP(O)(O—)O—	— 1.309	— 1.204			—HNC(O)N—		— 1.476	— 1.204	
—P(O)—	— 1.906				—HNC(O)NO—		— 1.026		
—P(S)—	— 1.160				—S(O)(O)NHC(O)HN—				— 1.571
含 C, 不含 H					芳香环或类芳香环上的片段(f^{AR})				
—C	0.519	0.423		0.010	—C(H)=	0.305			
—CF ₃		0.521			—C=	0.251			
—CN		0.075			—N=	— 0.308			
—C(O)N=	— 1.767		— 1.833		—S—	0.748			
—C(O)—	— 1.355	— 0.839			—C(O)=	— 0.898			
—C(O)O—	— 0.434	— 0.427			—N—	— 0.739			
—C=N—	— 0.272				—NH—	0.413			
—OC(O)N=	— 1.831				—O—	0.533			
—SC(O)N=	— 0.365								

2.2 模型估算误差分析

为对模型的估算误差进行初步分析,比较了不同化合物的实测 K_{oc} 与估算误差(估算与实测结果之差)的关系,结果如图 1 所示。

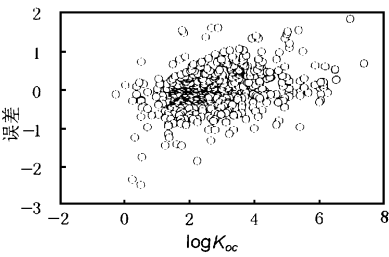


图 1 估算 K_{oc} 误差随实测 K_{oc} 取值大小变化

估算误差与 $\log K_{oc}$ 实测值之间具有某种程度的正相关关系,即 K_{oc} 较小的亲水化合物的误差相对较小,反之,憎水化合物的估算误差偏大。由于实测 K_{oc} 必然与化合物的亲水性有直接关系,某种方式的校正或许可以进一步降

低模型中的系统偏差。此外,由于土壤或沉积物样品特性的差异,实测 K_{oc} 本身带有波动。这样的误差可以根据不同文献中报道的同种化合物的重复实测 K_{oc} 值评估。模型估算误差和实测数据误差对比如图 2 所示。

图 2 中右侧为实测数据自身的波动。虽然较左侧模型估算误差小得多,但平均幅度也达 0.26 个对数单位。相比之下,模型估算误差为 0.37 个对数单位。显然,模型估算误差中的相当部分来自实测资料本身的变异。

2.3 模型稳健性的单一化合物评估

模型稳健性的单一化合物检验包括 2 种方式。前者重复 100 次每次随机抽取一个化合物,后者分 34 次每次抽取误差大于 1.0 个对数单位的化合物。在分 100 次随机抽取一个独立化合物的检验中,jackknife 可决系数大多围绕原始模型的可决系数值(0.9696)随机波动(图 3),仅一个化合物的正偏幅度较大。可见,除个

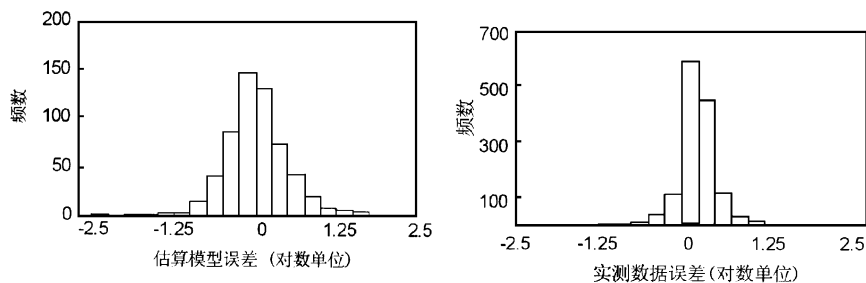


图2 模型估算及实测数据误差分布

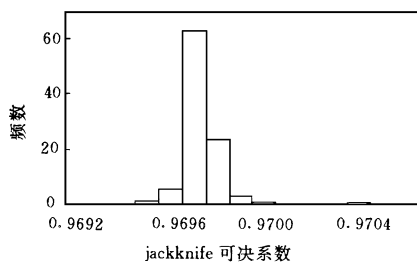


图3 单一化合物随机抽取jackknife可决系数分布

别值外,模型对所研究的化合物具有相当高的稳健性。

对研究的片段常数模型的估算结果中有34种化合物的估算误差超过1.0个对数单位。针对这些化合物的单一抽取jackknife检验结果如图4所示。

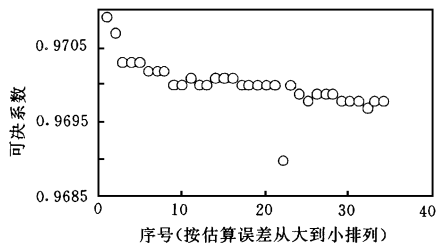


图4 单一化合物(误差超过1.0对数单位)随机抽取jackknife可决系数分布

图4中除一个化合物(fenvalerate)外,其余jackknife可决系数均在原始模型的可决系数(0.9696)之上,充分体现了这些化合物是影响模型稳健性的主要因素。即便如此,对模型影响最大的2个化合物(序号1、2)的jackknife可决系数也仅比原始模型高0.0013和0.0011,

表现出相当高的稳健性。

2.4 模型稳健性的多化合物评估

片段常数模型的稳健性也可以通过多个化合物随机抽取的方式检验。在重复了40次每次抽取50种化合物后,计算得到的jackknife可决系数的分布在图5中给出。除由于抽取量大,波动幅度较大之外,图5表现出的可决系数分布形式与单一随机抽取的结果基本一致。即大体表现为围绕原始模型可决系数(0.9696)波动的随机正态分布。

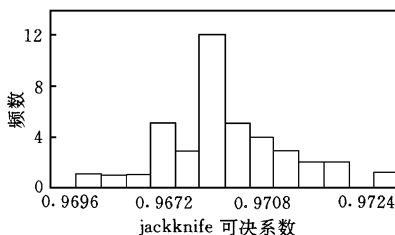


图5 40次随机抽取50种化合物jackknife可决系数分布

如果在jackknife检验中分别抽取不同类别的化合物,得到的结果显然不如成组随机抽取。研究中涉及的592种化合物可以分为表3列举的17类。

分别抽取这17类化合物,对余下16类化合物进行jackknife检验。得到的结果示于表3的最后一列。就本研究涉及的17类化合物而言,jackknife检验结果未发现任何类别化合物对模型稳健性有明显影响。多数类别的jackknife可决系数略高于原始模型。其中最高的醛(0.9727)也仅比原始可决系数(0.9696)高0.0031。由于所研究的化合物中醛类化合物数

表 3 研究化合物类别及各类平均估算误差

序 号	类 别	化合物个数	log K_{oc} 均值	平均绝对误差	平均相对误差	R^2_f
1	酸	30	1. 63	0. 522	0. 320	0. 9722
2	醚	15	2. 47	0. 680	0. 275	0. 9727
3	醇、酚	55	2. 06	0. 486	0. 236	0. 9718
4	磷酸酯	57	2. 67	0. 441	0. 165	0. 9710
5	(硫代)氨基甲酸酯	45	2. 23	0. 346	0. 155	0. 9705
6	单环芳烃	16	2. 99	0. 438	0. 147	0. 9696
7	硝基苯	13	2. 85	0. 418	0. 147	0. 9696
8	酰胺	51	1. 94	0. 283	0. 146	0. 9707
9	胺、苯胺	50	2. 84	0. 382	0. 135	0. 9697
10	酯	28	3. 17	0. 409	0. 129	0. 9701
11	唑、吡啶、嘧啶、三嗪	40	2. 65	0. 340	0. 128	0. 9703
12	脲	56	2. 01	0. 248	0. 123	0. 9699
13	醛、酮	22	2. 38	0. 291	0. 122	0. 9702
14	卤代烃	45	2. 69	0. 243	0. 090	0. 9691
15	多氯联苯	20	5. 11	0. 343	0. 067	0. 9669
16	多环芳烃	42	4. 70	0. 246	0. 052	0. 9652
17	其他	7	3. 22	0. 123	0. 038	0. 9693
	合计	592				

量相对较少(15), 0. 0031 的差异也比随机抽取检验结果大. 可见, 模型类别之间的外推性较差. 这些结果与用分子连接性指数和建立 K_{oc} 预测模型的检验结果基本一致^[7].

从表 3 中还可以看出, 在 17 类化合物中, 酸类化合物的绝对误差最大. 一般而言, 对 log K_{oc} 较小的亲水化合物的估算绝对误差大于对憎水化合物的估算结果. 其原因之一可能是这些化合物 K_{oc} 测定误差较大.

3 结论

对建立在 592 种化合物实测 K_{oc} 资料基础上的片段常数估算模型而言, 其绝对估算误差与 K_{oc} 值正相关. 而误差则随化合物亲水性增强而增大. 无论采用哪种抽取方法, jackknife 可决系数都表现出较高的稳健性. 相对而言, 估算模型在化合物类别之间的推广可能造成的误差较大.

参 考 文 献

1 Gawlik. B M N Sotiriou et al. Alternatives for the determination of the soil adsorption coefficient, K_{oc} , of non-ionic organic compounds—— a review. Chemosphere, 1997, **34**: 2525~2551

2 Sabljic H, Gusten H, Verhaar H, Hermens J. QSAR modeling of soil sorption, improvements and systematics of log K_{oc} correlations. Chemosphere, 1995, **31**: 4489~4514

3 Gerstl Z. Estimation of organic chemical sorption by soils. J. Contam. Hydrol., 1990, **6**: 357~375

4 Muller M and Kordel W. Comparison of screening methods for the estimation of adsorption coefficients on soil. Chemosphere, 1996, **32**: 2493~2504

5 Nirmalakhandan N N and Speece R E. Prediction of aqueous solubility of organic chemicals based on molecular structure. Environ. Sci. Technol., 1988, **22**: 328~338

6 Dietrich S W, Nicholas D D, Dreyer D and Hansch C. Confidence interval estimators for parameters associated with quantitative structure-activity relationship. J. Med. Chem., 1980, **23**: 1201~1205

7 陶澍, 卢晓霞, 胡海瑛. 有机化合物 K_{oc} 分子拓扑-极性校正模型的稳健性检验. 土壤学报, 1999, **36**(3)