有机化合物 K_{oc} 分子拓扑—极性校正模型的稳健性检验*

陶 澍 卢晓霞 曹 军 胡海瑛

(北京大学城市与环境学系,北京 100871)

摘 要 用修正的 jackknife 检验研究了估算有机化合物吸着系数(K_{oc})的分子拓扑—极性校正模型及模型参数的稳健性。检验分四种不同方式进行:随机抽除单一化合物(100次)、逐一抽除异常化合物(估算值与实测值差别在0.9个对数单位以上者,27次)、随机抽除50个化合物(30次)以及逐一抽除特定化合物类别(17类)。检验结果表明,对所研究的模型,不同检验方式得到的结果一般具有相似的趋势,即多元回归模型的可决系数表现出较强的统计稳健性,相比之下,回归系数的稳健性较低,特别是六阶链指数$^6\chi_{ch}$。在17种极性校正因子中,SO_2、O、F、NH_2和SO等的稳健性相对较低。

关键词 吸着系数(K_{oc}),分子连接性指数,极性校正因子,jackknife 检验,稳健性

中图分类号 S153

估算有机化合物土壤/沉积物吸着系数(K_{oc})的模型可简单分为两类,即根据其他环境参数推算和根据化合物分子结构估算。在各种依据分子结构估算 K_{oc} 的模型中,利用分子连接性指数并结合极性校正因子的方法具有较好的发展前景。分子连接性指数不仅成功地应用于包括亨利常数、溶解度和水/辛醇分配系数在内的许多环境参数的估算[1,2],在预测吸着系数方面也有比较成功的例子[3]。

建立吸着系数估算模型的一般方法是将已知实测值的化合物分为两组,其中一组用作模型参数识别,另一则用来对建立的统计模型进行验证。除了利用分组的方式检验模型的预测可靠性外,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)[4]。除回归模型稳健性判断外,jackknife 检验也可用来计算非线形回归模型中特定衍生参数的置信区间[5]。

本研究的目的在于以预测有机化合物吸着系数 K_{oc} 的分子拓扑-极性校正模型为例,探讨 jackknife 方法在结构-活性研究中的应用。并利用修正的 jackknife 法从若干侧面检验运用分子连接性指数-极性因子校正法建立的 K_{oc} 预测模型的稳健性。除对回归方程的

* 自然科学基金重点项目(49632060)和国家杰出青年基金(49525102)资助

收稿日期:1998-08-22,收到修改稿日期:1999-07-28

多元可决系数进行检验外，还研究了不同回归系数的稳健性特征。

1 研究方法

1.1 K_{oc} 数据收集

K_{oc}模型建立在543种有机化合物 K_{oc} 实测数据的基础上。K_{oc} 数值范围跨6个对数单位。资料来自Kenaga、Briggs、Sabljic 的研究和其他有关文献[6~20]。由于不同文献经常给出同一化合物的不同 K_{oc} 实测值，实际建立的数据库包括1,100个实测值。对于不同来源的多个数据，建立和验证模型时取中位数。凡以 K_{om} 形式发表的数据，则取1.724为换算系数。

将543个 K_{oc} 数据分为建模和检验两组。建模组由随机抽除的400个数据组成，包括85种非极性化合物和315种极性化合物。其余143个数据构成检验组，其中非极性化合物22种，极性化合物121种。建模组数据用于模型参数计算，其结果以验证组数据检验[1)]。

1.2 预测模型建立

根据前期研究结果，选择以下形式的多元表达式建立模型：

$$\log K_{oc} = \sum(a_i \chi_i) + \sum(F_i n_i) + C \qquad (1)$$

式中 χ_i 为第 i 个分子连接性指数，n_i 为第 i 种极性基团的个数，a_i、F_i 和 C 均为回归系数，其中 F_i 也是第 i 种极性基团的极性校正因子。

用逐步回归方法从9个常用分子连接性指数(包括4项路径指数、1项路径/簇指数和4项链指数)中选出最重要的3项作为模型拓扑参数($^1\chi^v$、$^6\chi_{ch}$ 和 $^2\chi$)。对最初选定的15种极性校正基团进行逐步回归的结果表明，所有基团都对模型有显著贡献。此外，由于内在差别较大，PO_x 应当进一步细分为 PO_2、PO_3 和 PO_4 三类。最终模型包括17种极性校正因子[10]。在利用模型组数据建立回归模型后，将模型用于验证组数据。比较了验证组中143种化合物的模型预测结果与实测结果借以判断模型的优劣。结果表明，利用上述参数建立的估算模型，可在超过6个对数单位范围内进行较可靠的预测。为充分利用实测资料，在验证后用所有543个化合物实测数据建立最终模型。模型的平均绝对误差为0.346个对数单位，在543个化合物中，74.4%的绝对误差小于0.5个对数单位。模型稳健性检验即针对最终模型进行。

1.3 模型稳健性检验

Jackknife 检验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)抽除、异常(K_{oc} 预测值与实测值之差大于0.9个对数单位者)化合物抽除、随机化合物组抽除以及特定化合物类型抽除。

1.3.1 单一化合物抽除　从543种化合物中重复100次，每次随机抽除1种，对其余542种进行多元回归。

1.3.2 异常化合物抽除　对27种具预测异常值的化合物，分27次每次抽除其一，对余下542种化合物进行多元回归。

1.3.3 50种化合物随机抽除　从543种化合物中每次随机抽除50种，对其余493种化合物进行回归(共30次)。

1.3.4 分类抽除　对研究中涉及的17类化合物逐类抽除，每次对余下16类化合物进行多元回归。

1) 卢晓霞，陶澍，根据分子拓扑和基团贡献估算有机化合物的吸附系数(K_{oc})，待发表

根据四种形式的 jackknife 检验结果,分别针对模型可决系数、拟合拓扑参数和极性校正因子的波动进行分析,借以判断总体模型以及个别拟合参数的稳健性。

1.4 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 K_{oc}的分子拓扑-极性校正模型

在用逐步回归选择了必要的拓扑参数和极性校正因子后,分别利用建模组数据和验证组数据进行回归分析,最终得到如下预测模型[10]:

$$\log K_{oc} = 0.188^{1}\chi^{v} + 0.336^{2}\chi + 0.755^{6}\chi_{cb} + \sum(F_i n_i) + 0.922 \quad n = 543, R^2 = 0.8603 \quad (2)$$

式中极性校正因子 F_i的取值如表 1 所列。

表1 极性基团及极性校正因子

Table 1 Polar groups and polarity correction factors

极性基团 polar groups	极性因子,F_1 polarity factors	极性基团 polar groups	极性因子,F_1 polarity factors	极性基团 polar groups	极性因子,F_1 polarity factors
—COOH	−1.411	—PO_2—	−0.955	—NO_2	−0.481
—CON—	−1.501	—PO_3—	−1.017	—O—	−0.449
—NCON—	−1.443	—PO_4—	−1.839	—N═	−0.379
—NCOO—	−1.315	—COO—	−0.833	—F	−0.304
—SO_2—	−1.578	—CO—	−0.771	—NH_2(NH, N)	−0.284
—SO—	−1.183	—OH	−0.788		

在样本量为 543 的条件下,多元回归的可决系数达 0.8603,即 86% 的 K_{oc}波动可由模型说明。可见模型可以较好地预测大多数有机化合物的 K_{oc}值[10]。

2.2 回归模型的稳健性

在以不同方式进行 jackknife 抽样后,可以根据 jackknife 可决系数对模型的总体稳健性进行检验。当模型不稳健时,jackknife 可决系数的正偏幅度较大。虽然还没有有效的定量检验方法,但仍可根据其变幅作出判断。如图 1 所示,在分 100 次随机抽除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(0.8603)随机波动,两个正偏幅度较大的化合物恰好是两个异常值

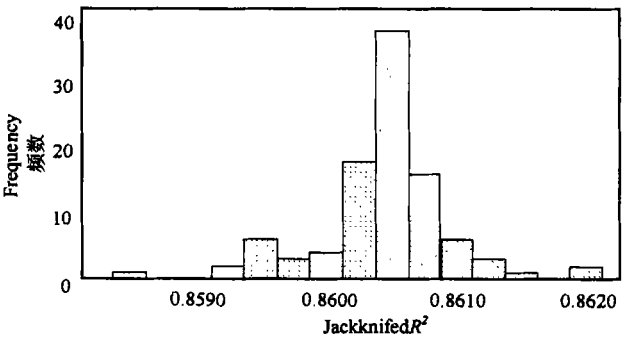


图 1　100次随机抽除单一化合物的jackknifed R^2分布

Fig.1 Distribution of the jackknifed R^2 derived from deletion of 100 random single chemical

(jackknife 可决系数分别为 0.8618 和 0.8688)。可见,除个别异常值外,模型对所研究的化合物具有相当高的稳健性。

如果直接抽除异常值(计算结果与实测结果差别在 0.9 个对数单位之上者)进行检验,所得结果如图 2 所示。

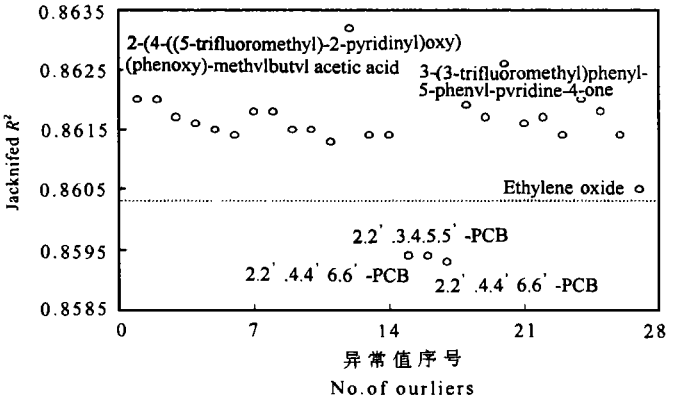
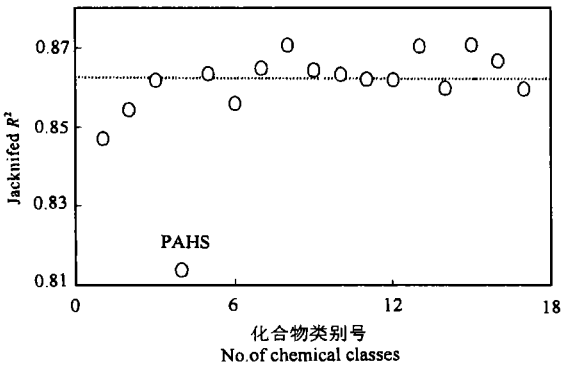


图 2 对 27 个异常值进行的 jackknife 检验结果(可决系数)

Fig.2 The jackknifed R^2 derived from deletion of 27 outliers

除 3 个 PCB 外,大多数异常化合物的 jackknife 可决系数略高于模型可决系数(R^2 = 0.8603,图中虚线),反映了异常化合物存在对模型的影响。尽管如此,其中影响最大的两个化合物的 jackknife 可决系数也只比模型可决系数高 0.0023 和 0.0029。在总样本量高达 543 的模型中,这些异常值的存在对模型稳健性的影响可以忽略不计。3 个 PCB 之所以产生负效应是因为其 $\log K_{oc}$ 均在 6 左右。它们的相对误差并不太高,并非严格意义上的异常化合物。



1. 多氯联苯, 2. 卤代烃, 3. 单环芳烃, 4. 多环芳烃, 5. 酸, 6. 醇、酚, 7. 醛、酮, 8. 胺、苯胺, 9. 酰胺, 10. 硝基苯、腈, 11. (硫代)氨基甲酸酯, 12. 酯, 13. 唑、吡啶、嘧啶、三嗪, 14. 醚, 15. 有机磷, 16. 脲, 17. 其他

图 3 化合物类别 jackknife 检验结果(jackknife R^2)

Fig.3 The jackknifed R^2 derived from deletion of compound class

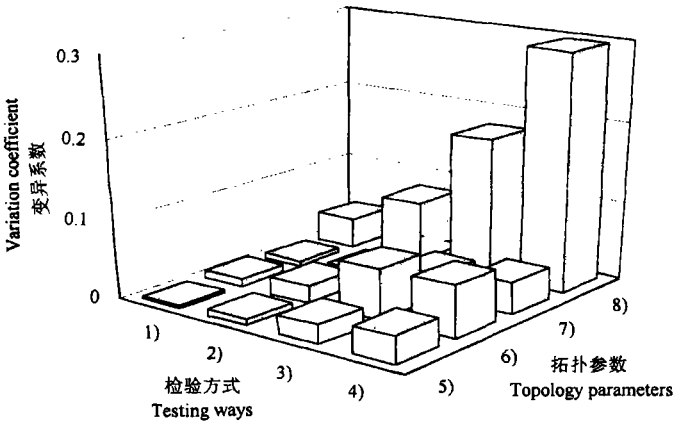
若将所有 543 种化合物分成 17 类,然后分别根据类型进行 jackknife 检验,计算可决系数在图 3 中列举。

就本研究涉及的化合物类别而言,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。多环芳烃表现出来的 jackknife R^2 负偏作用则说明了此类化合物在分子拓扑模型中的核心作用。

根据随机方式抽除 50 种化合物(共 30 次)进行的 jackknife 检验结果未发现任何非稳健因素。

2.3　拟合拓扑参数的相对稳健性

为检验并比较回归模型得到的拓扑参数和模型截距(常数项)的稳健性,计算了根据不同检验方式得到的三种拓扑参数($^{1}\chi^{v}$、$^{6}\chi_{ch}$和$^{2}\chi$)和模型截距的变异系数。四种检验方式包括随机抽除单一化合物(100 次)、逐一抽除异常化合物(27 次)、随机抽除 50 个化合物(30 次)以及逐一抽除特定类别化合物(17 类)。计算结果见图 4。



1) 单一随机　2) 单一异常　3) 50随机　4) 类别　5) 截距　6) $^{1}\chi^{v}$　7) $^{2}\chi$　8) $^{6}\chi_{ch}$

图 4　不同方式 jackknife 检验得到的拓扑参数和截距的变异系数比较

Fig.4　Comparison of variation coefficient of topological parameters and interception derived from various jackknife tests

虽然不同回归系数在模型中的取值有较大差别,如$^{1}\chi^{v}$和$^{6}\chi_{ch}$的回归系数分别为 0.188 和 0.755,但由于在计算变异系数时实际上已根据它们各自均值进行了校正,图中变异系数的大小如实反映了不同参数稳健性的差异。从图中可见,无论采用何种检验方式,变异最大的拓扑参数是$^{6}\chi_{ch}$的回归系数,其余三种回归系数的变异显著低于前者。图中不同检验方式的差别与模型无关。每次抽除数较少的方法(每次抽除一种化合物的单一随机抽除和单一异常化合物抽除)得到的变异系数较低在情理之中。此外,随机抽除的结果一般比相当抽除量的非随机抽除(对应于单一随机抽除的单一异常抽除以及对应于 50 个化合物随机抽除的按类别抽除)波动大。

2.4　拟合极性校正因子的相对稳健性

对于所选择的 17 种极性校正因子,分别用四种方式的 jackknife 检验研究了拟合模型中不同因子的相对稳健性。由于不同系数的大小甚至量级有较大差别,对 jackknife 检验中计算得到的 17 个极性校正因子的回归系数进行了归一化处理,结果以玫瑰图形式在图 5 中显示。

从图中可见,虽然大部分极性因子包括 $-N=$、OH、COOH、CO、SO_2、CON、NCON、NCOO 和 COO 等在所有检验结果中均表现出很小的变异,但其余极性因子在不同检验方法结果中表现出较大差别,因此四幅玫瑰图形态各不相同。例如,在单一随机抽除检验和异常抽除检验中 O 和 SO 等官能团表现出较强的波动,但在多化合物抽除的后两类检验中

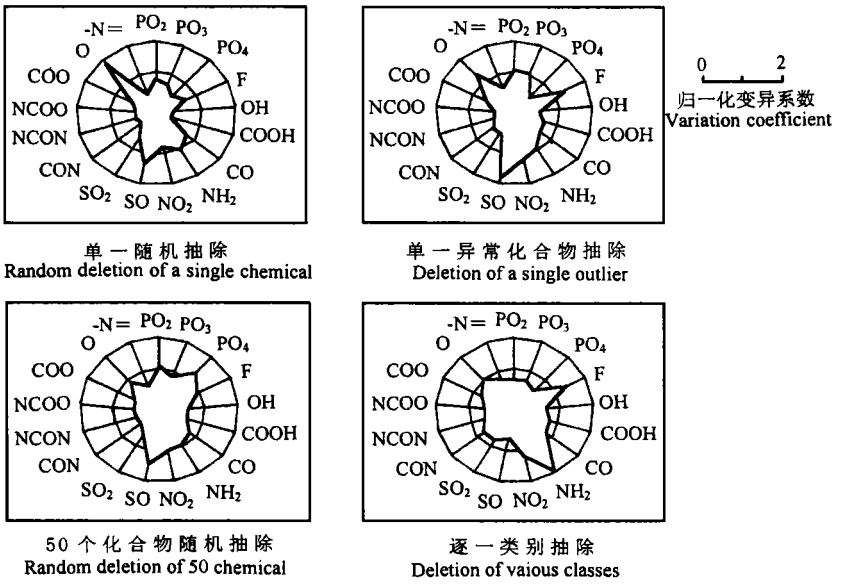


图 5 不同方式 jackknife 检验得到的极性因子变异系数(经归一化)比较(内、外环刻度分别为 1 和 2)

Fig.5 Comparison of variation coefficient of polarity correction factors by various jackknife tests (scale of the inner and the outer circle are 1 and 2, respectively)

则不然。逐一类别抽除检验结果中波动较大者往往与某些类别特有的官能团(如 F、NO_2 和 NH_2)有关。由此可见,各种极性校正因子在回归模型中的稳健性有很大差别,其中稳健性较差的官能团包括 SO_2、O、F、NH_2和 SO 等。

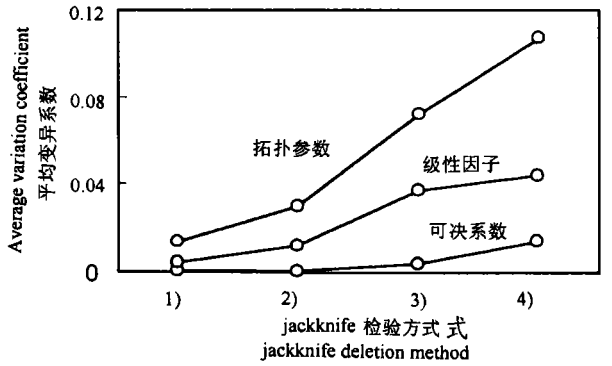


图 6 不同类型参数jackknife变异系数大小比较

Fig.6 Comparison of variation coefficient of various types of parameters by tests

2.5 不同类别统计量稳健性比较

为比较回归模型可决系数、极性校正因子和拓扑参数(分子连接性指数)的回归系数这三类统计量在本研究所建立的回归模型中的稳健性,分别计算了不同 jackknife 检验方式下得到的参数变异系数的算术均值。计算结果列于图 6 中。图中三类统计量的稳健性差别一目了然。无论采用哪种检验方式,回归可决系数的变异系数都最小,其次为极性校正因子,拓扑参数的回归系数的变异系数显著高于其余两者。其中导致较大波动的参数显然是$^6\chi_{ch}$(见图 4)。当大多数研究者仅仅针对可决系数进行 jackknife 检验时,很容易忽略多元回归模型回归系数本身的稳健性。

3 结　论

Jackknife 检验可用于判断估算有机化合物吸着系数 K_{oc}的拓扑—极性校正因子模型的稳健性。对利用 543 种化合物实测数据建立的多元回归模型,无论采用何种检验方式,模型可决系数的波动都很小,表现出较高的总体稳健性。相比之下,回归系数的稳健性较低。一般而言,拓扑参数的稳健性低于极性校正因子。其中$^6\chi_{ch}$是模型中最不稳健的参数。稳健性较差的极性校正因子包括 SO_2、O、F、NH_2和 SO 等。

参　考　文　献

1. Gerstl Z. Estimation of organic chemical sorption by soils. J. Contam. Hydrol., 1990, 6:357～375
2. Muller M, W Kordel. Comparison of screening methods for the estimation of adsorption coefficients on soil. Chemosphere, 1996, 32:2493～2504
3. Meylan W, Howard. P H. Molecular topology / fragment contribution method for predicting soil sorption coefficients. Environ. Sci. & Technol., 1992, 26:1560～1567
4. Nirmalakhandan N N, Speece R E. Prediction of aqueous solubility of organic chemicals based on molecular structure. Environ. Sci. Technol., 1988, 22:328～338
5. Dietrich S W, Nicholas D D, Dreyer D, Hansch C. Confidence interval estimators for parameters associated with quantitative structure–activity relationship. J. Med. Chem., 1980, 23:1201～1205
6. Kenaga E E, Goring C A. Relationship between water solubility, soil sorption, octanol–water partitioning, and conc of chem in biota. In: Eaton J G, Parish P P and Hendricks A C. Eds. Aquatic Toxicol, Proc of the 3rd Annual Symp on Aqua Toxicol. ASTM, STP 707. Philadelphia. 1980, 78～115
7. Briggs G G. A simple relationship between soil sorption of organic chemicals and their Octanolwater partitioning behavior. Proc 7th British Insecticide and Fungicide Conf., 1973. 11:475～478
8. 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
9. Briggs G G, Theoretical and experimental relationships between soil adsorption, octanol–water partition coefficients, water solubilities, bioconcentration factors, and the parachor. J. Agric. Food Chem., 1981, 29: 1050～1059
10. Karickhoff S W, Brown D S, Scott T A. Sorption of hydrophobic pollutants on natural sediments. Water Res., 1979, 13:241～248
11. Karickhoff S W. Semiempirical estimation of sorption of hydrophobic pollutants on natural sediments and soils. Chemosphere, 1981, 10(8):833～846
12. Karickfoff S W, Morris K R. Sorption dynamics of hydrophobic pollutants in sediment suspensions. Env. Toxicol. Chem., 1985, 4(4):469～479
13. Vowles P D, Mantoura R F C. Sediment–water partition coefficients and HPLC–retention factors of aromatic hydrocarbons. Chemosphere, 1987, 16:109～116
14. Hassett J J. Sorption properties of sediments and energy–related pollutants. EPA–600/3–80–041, 1980
15. Hassett J J. Sorption of Alpha–naphthol:Implications concerning the limits of hydrophobic Sorption. Soil Sci. Soc. Am. J., 1981, 45(1):38～42
16. Chiou C T, Peters L J, Freed V H. A physical concept of soil–water equilibria for non–ionic organic compounds. Science, 1979, 206:831～832
17. Chiou C T, Porter P E, Schmeddi D W. Partition equilibria of non–ionic organic compounds between soil

organic matter and water. Env. Sci. & Tech., 1983, 17(4):227~231
18. Gustafson David L. Pesticides in Drinking Water. Van Nostrand Reinhold 1993
19. Montgomery J. Agrochemicals Desk Reference-environmental Data. Lewis Publishers 1994
20. Karel V. Handbook of Environmental Data on Organic Chemicals. Van Nostrand Reinhold 1996

ROBUSTNESS TEST OF A TOPOLOGICAL INDICES AND POLARITY FACTORS MODEL FOR ESTIMATING K_{OC} OF ORGANIC COMPOUNDS

Tao Shu, Lu Xiao-xia, Cao Jun, Hu Hai-ying

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

Summary

Modified jackknife test was applied to examine the robustness of a topology-polarity correction model for estimating the sorption coefficient (K_{oc}) of organic compounds. The test was performed by four methods: 1) random deletion of a single chemical from the data set (100 times); 2) deletion of a chemical with abnormal value (difference between predicted and measured values is larger than 0.9 log-units, 27 outliers); 3) deletion of 50 chemicals randomly selected from the data set (30 times); and 4) deletion of a set of chemicals belonging to the same class (17 classes). The results indicated that similar trends were observed with various jackknife tests. The coefficient of determination (R^2) of the multivariate regression model showed relatively high robustness compared to other regression coefficients. The sixth-order chain molecular connectivity index ($^6\chi_{ch}$) was the least robust parameter in the model. Among the 17 polarity factors, the robustness of SO_2、O、F、NH_2 and SO were lower than the others.

Key words Sorption coefficient, K_{oc}, Molecular connectivity indices, Polarity correction factor, Jackknife test, Robustness