Chemosphere, Vol.22, Nos.9-10, pp 953-957, 1991
Printed in Great Britain

0045-6535/91 $3.00 + 0.00
Pergamon Press plc

ESTIMATIONS OF AQUEOUS SOLUBILITY FROM N-OCTANOL/WATER PARTITION COEFFICIENTS ANALYZED BY THE BOOTSTRAP METHOD

Efraim Halfon[1], Joachim Altschuh[2], Rainer Brüggemann[2,*] and Walter Karcher[3]

[1] Lakes Research Branch, National Water Research Institute
Canada Centre for Inland Waters, Burlington, Ontario, Canada L7R 4A6

[2] GSF - Forschungszentrum für Umwelt und Gesundheit
Projektgruppe Umweltgefährdungspotentiale von Chemikalien
Ingolstädter Landstr. 1, D-8042 Neuherberg, Federal Republic of Germany

[3] Joint Research Centre - Ispra, Varese, Italy

ABSTRACT

A statistical method, the bootstrap is used to test the reliability of QSAR models. Seven property-property relationships (PPRs) for the estimation of the aqueous solubility from the n-octanol/water partition coefficient (and the melting point) were used as an example. The testing data set includes over 350 substances. The result of the bootstrap analysis is straightforward: One PPR turned out to be more reliable than all the others.

1. INTRODUCTION

Property-property relationships (PPRs), as an essential part of QSARs, allow an estimation of substance properties without analyzing molecular structures. In this paper we combine two statistical techniques, the bootstrap and functional linear regressions, to predict the reliability of PPRs. First we describe the bootstrap method [1, 2]. Then we introduce the functional (geometric) regression method [3]. Third we assess which of seven linear regression models relating aqueous solubility and the n-octanol/water partition coefficient is the best.

* To whom reprint requests should be sent.

2. STATISTICAL METHODS

2.1 General Remarks on the Bootstrap Method

The bootstrap method, invented by Efron [1], generalizes the jackknife method and is used to estimate the uncertainties associated with a statistics of interest, for example with the slope and the intercept of a linear regression model. In ecotoxicology, for example, linear regression models are developed using data from a limited number of chemicals; but there are always chemicals not included in the model development. The bootstrap method infers the conclusions we could expect if all hundreds of chemicals of the same class would have been included in the statistical analysis. Recently the bootstrapping was applied to reexamine QSARs and to deduce confidence intervals for the coefficients of the predicting equations [4].

2.2 The Technique of the Bootstrap Method#

Although the bootstrap method is well explained in [4], we will briefly summarize some main steps. Assume that we have N observations of a substance property Y_{obs} and assume that a QSAR method has been used to predict these variables. For example,

$$Y_{pred} = f(\text{QSAR variables}) \quad (1)$$

where Eq. 1 can be any QSAR model. We want to know the reliability and applicability of these predictions. To assess the reliability of a QSAR we test the quality of the functional relation between Y_{pred} and Y_{obs}. For each QSAR we compute a functional regression (using the technique presented below) between predicted and observed data. The corresponding equation has the following general form.

$$Y_{pred} = a + b\,Y_{obs} \quad (2)$$

The coefficients of Eq. 2, a and b, are the intercept and slope, respectively. If a QSAR is appropriate we expect a to be zero and b to be one with a certain (95 %) degree of confidence.

To implement the bootstrap we store all the N data pairs into two arrays Y_{obs} and Y_{pred}. In the bootstrap procedure these N pairs are randomly sampled with replacement. A random number generator chooses the pairs. Since the choice is random some pairs might be chosen more than once and some not at all. Since we have N pairs in our analysis, the first time we run the program we have a set of N pairs chosen randomly from the original set. This procedure is repeated about M times (we recommend: $M \geq 10 \cdot N$). We now have a set containing $M \times N$ data pairs. For each of the M sets we compute a functional regression (using the technique presented in sect. 2.4) between predicted and observed data. Thus, in memory we keep M slopes and M intercepts. These slopes and intercepts represent the frequency distribution function F' that infers the real one F if we had in our hand all the possible data required for their analysis (this assumption is at the basis of the bootstrap, see references). From this frequency distribution we can compute the median, the average, the standard deviations (-1 SD, +1 SD), and the confidence limits (-95 %, +95 %). The confidence limits are usually not symmetrical since the probability distribution of the data is not normal.

The bootstrap computer program is obtainable from the first author. The program runs on any mainframe or MS-DOS microcomputer. As an example each case presented in this paper with 4000 Monte Carlo simulations and 355 pairs of data took about 20 minutes on an 286 machine.

2.3 Analysis of the Bootstrap Results

The most important features of the frequency distribution F' are the median and the following intervals I_i

$$I_1 = (\text{min, max}) \tag{3}$$

$$I_2 = (\text{confidence limit (- 95\%), confidence limit (+ 95\%)}) \tag{4}$$

$$I_3 = (\text{standard deviation (-1 SD), standard deviation (+1 SD)}) \tag{5}$$

Clearly

$$I_1 \supset I_2 \supset I_3. \tag{6}$$

These intervals and the median of the frequency distributions of the different PPRs to be reexamined for the slope and intercept have to be compared with the ideal values of a (=1) and b (=0). A QSAR, Eq. 1, is as more reliable, as the smaller interval includes the ideal values.

Conversely, if the ideal value is not enclosed by at least I_1, the corresponding QSAR cannot be recommended. Thus, we state that the original QSAR model has a bias.

2.4 The Functional Regression

The functional regression model has been described in the literature since the late 1940s. It has been applied to QSAR data by Halfon [3]. The method is based on the assumption that the coefficients of Eq. 2 must be computed taking into account that both the predicted values Y_{pred} and the observations Y_{obs} are uncertain. The Y_{pred}-values are uncertain because they are obtained by statistical estimations. The data, Y_{obs}, are also uncertain because of experimental problems.

3. THE PPRS AND THE DATA SET

The relation between aqueous solubility (WS) and n-octanol/water partition coefficient (K_{ow}) is given as

$$\log(WS) = c \cdot \log(K_{ow}) + d + e \cdot f(T_m) \tag{7}$$

$$f(T_m) = \begin{cases} T_m \text{ for } T_m > 25\ ^\circ C \\ 25\ ^\circ C \text{ elsewhere} \end{cases} \tag{8a}$$

$$f(T_m) = \begin{cases} T_m \text{ for } T_m > 25\ ^\circ C \\ 0 \text{ elsewhere} \end{cases} \tag{8b}$$

(WS is given in mol/L and the melting point T_m in °C.)

A number of PPRs for these two properties can be found in the literature. In a previous paper [5] we used 26 of them in a validation study based on an analysis of Eq. 2. The mean square error, the deviation of the slope/intercept from their ideal values, and the number of outliers were used as criteria in a ranking procedure to evaluate the "best" PPR. Based on the results of this study we selected seven PPRs (see Table 1) for the applica-

tion of the bootstrap method: PPRs No 3§) and 15, which are developed for mixed chemical classes but turned out to be not recommendable, but maintained in this study to verify the response of the bootstrap method with respect to such PPRs; PPR No 24, derived from theoretical (thermodynamical) considerations; finally, PPRs No 2, 17, 21, and 25, which turned out to be recommendable relations.

From a survey of the literature we compiled data for WS, K_{ow}, and T_m as testing sets for the PPRs [5]. The resulting number of compounds used in the bootstrap validation study is 355 and 374, respectively, depending on whether T_m is required or not.

Table 1: Characterization of some PPRs between WS and K_{ow} (c, d, and e refer to Eq. 7; WS in mol/L).

PPR	c	d	e	N	Range of applicability, reference
2	-1.12	1.3	-0.015[a)]	27	Mixed classes[b)], [6]
3[c)]	-0.922	1.184	0	90	Mixed classes, [7]
15	-1.339	0.978	-0.0095[d)]	156	Mixed classes, [8]
17	-0.9874	0.7178	-0.0095[a)]	35	Halobenzenes, [9]
21	-1.26	1.0	-0.0054[a)]	300	Mixed classes, [10]
24	-1.0	1.05	-0.01[a)]	-	Theoretical relation, [11]
25	-1.05	0.87	-0.012[a)]	155	Mixed classes, [11]

[a] Eq. 8a. [b] In the original paper calculated as $\log K_{ow} = f(WS)$. [c] WS in g/L. [d] Eq. 8b.

4. RESULTS AND DISCUSSION

As Table 2 shows, the correlation coefficient is not very helpful to decide which PPR should be recommended. Therefore the main interest is directed toward the position of the intervals I_i (Eqs. 3-5) relative to the ideal values of the slope and intercept of Eq. 2 (cf. Table 2).

Only PPR No 25 has both characteristics within the confidence limits. This is the main result of the bootstrap analysis. Some further results for the PPRs No 3, 15, 17, and 24 may be discussed as examples. The worst PPRs are No 3 and 15, which agrees with results given in [5]. For No 3 the intercept deviates strikingly from its ideal value of zero. With respect to the intercept the PPRs No 15 and 17 are quite well. However the slopes are outside of the confidence limits. The theoretical relation (PPR No 24) is not the best one but ranks together with No 17 as one of the two second best relations. Whereas the intercept of PPR No 17 is located between 1 SD and +95 % and the slope is outside of the confidence limits, the PPR No 24 has a bias with respect to both. However, the slope of PPR No 24 is closer to the confidence limits than that of No 17.

In conclusion, the bootstrap analysis leads us to recommend PPR No 25 as the best one.

5. ACKNOWLEDGEMENT

We thank Prof. Lasser, GSF-Medis-Institut, for helpful discussions and the German Ministry for Research and Technology for supporting this work in the frame of Science & Technology Cooperation Germany/Canada.

§ We keep the numbers for the PPRs from the previous paper [5] to allow a better comparison of the results.

Table 2: Results of bootstrap. Median (Med) and confidence limits for the slope and the intercept, and correlation coefficient r of Eq. 2.

PPR	Intercept					Slope					r	
	-95 %	-1SD	Med	+1SD	+95 %	-95 %	-1SD	Med	+1SD	+95 %	Min	Max
2	0.06	0.15	0.21	0.30	0.39	1.04	1.06	1.10	1.13	1.18	0.88	0.96
3	-1.29	-1.22	-1.16	-1.10	-1.04	0.80	0.82	0.85	0.87	0.90	0.86	0.94
15	-0.02	0.03	0.12	0.18	0.26	1.16	1.19	1.21	1.24	1.29	0.89	0.96
17	-0.25	-0.18	-0.13	-0.07	0.00	0.87	0.89	0.91	0.94	0.97	0.89	0.96
21	0.01	0.08	0.17	0.22	0.30	1.03	1.06	1.08	1.11	1.14	0.88	0.96
24	0.06	0.13	0.18	0.24	0.32	0.88	0.90	0.93	0.96	0.99	0.89	0.96
25	-0.21	-0.16	-0.08	-0.03	0.08	0.94	0.97	1.00	1.03	1.07	0.89	0.96

6. REFERENCES

[1] (a) B. Efron, *Can. J. Statistics* **9** (1981) 139-172. (b) B. Efron, *Biometrika* **68** (1981) 589-599. (c) B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Methods*, SIAM Monograph No. 38, Society for Industrial and Applied Mathematics, 1982

[2] E. Halfon, *Chemosphere* **14** (1985) 1433-1440

[3] E. Halfon, *Environ. Sci. Technol.* **19** (1985) 747-749

[4] R. D. Crammer III, J. D. Bunce, D. E. Patterson, *Quant. Struct.-Act. Relat.* **7** (1988) 18-25

[5] R. Brüggemann and J. Altschuh, *Sci. Total Environ* accepted for publication

[6] S. Banerjee, S. H. Yalkowsky, and S. C. Valvani, *Environ. Sci. Technol.* **14** (1980) 1227-1229

[7] E. E. Kenaga and C. A. I. Goring, in J. G. Eaton, P. R. Parrish, and A. C. Hendricks, (Ed.), *Aquatic Toxicology*, ASTM STP 707, American Society for Testing and Materials, 1980, pp. 78

[8] C. Hansch, J. E. Quinlan, and G. L. Lawrence, *J. Org. Chem.* **33** (1968) 347-350

[9] S. H. Yalkowsky, R. J. Orr, and S. C. Valvani, *Ind. Eng. Chem. Fundam.* **18** (1979) 351-353

[10] P. Isnard and S. Lambert, *Chemosphere* **18** (1989) 1837-1853

[11] S. H. Yalkowsky and S. C. Valvani, *J. Pharm. Sci.* **69** (1980) 912-922

(Received in Germany 20 April 1991; accepted 6 May 1991)