IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 4 Issue 2 February, 2016

Estimation and modeling of bioaccumulation factor for a set of (PCBs): a QSPR study

Khadidja Amirat [1] ; Fatiha Mebarki[2] ;Nadia Ziani[3] ; Djelloul Messadi[4]

Environmental and food Security laboratory, Badji Mokhtar University, BP 12, 23000, Annaba, Algeria

1:khadidja_amirat@yahoo.fr

2:Fatiha_mebarki@yahoo.fr

3:Ziani_nadia84@yahoo.fr

4:d_messadi@yahoo.fr

Abstract

A structure/ bioaccumulation factor relationship was searched for 58 PCBs while promoting the hybrid genetic algorithm/simple linear regression approach, the structural parameters being calculated with the software Hyperchem and DRAGON. Among about a hundred of single regressor models gotten, we selected the one that present best values of the prediction parameter (Q^2) and of the determination coefficient (R^2).The reliability of the proposed model was further illustrated using various evaluation techniques: leave-many-out, cross-validation procedure, randomization test, and validation through the test set.

***Keywords*: PCBs, bioaccumulation factor, QSPR, molecular descriptors, software.**

I. INTRODUCTION

Polychlorinated biphenyls (PCBs), organic compounds with 1 to 10 chlorine atoms attached to biphenyl, have the general chemical formula $C_{12}H_{10-x}Cl_x$ (Figure 1). First manufactured by Monsanto in 1929, the PCBs production was banned in the 1970th due to the high toxicity of most PCBs (209) and mixtures [1]. PCBs were used as insulating fluids for industrial transformers and capacitors, and are known as persistent organic pollutants. Even if the production of the PCBs was stopped, they still have an influence on the human [2-4] and animal [5] health due to their accumulation in the environment. Moreover, the toxicity and carcinogenicity of PCBs could be related to mechanistic studies of their truncated analogue vinyl chloride [6]. Ecological and toxicological aspects of polychlorinated biphenyls (PCBs) in the environment are under investigation due to their worldwide distribution [7-10].Starting with the 20th century, several mathematical approaches, that link chemical structure and property/activity in a quantitative manner, have been introduced [11]. Nowadays, quantitative structure-property/activity relationships (QSPRs/QSARs) are currently used in pharmaceutical chemistry, toxicology and other related fields [12].

Bioaccumulation of chemicals is quantitatively expressed in terms of BCF, defined as the equilibrium of its concentration inside an organism (or in a certain tissue of the organism, usually in the fat) to that in the ambient environment [13]. The concentrations in tissues and in the environment are measured at steady-state after chronic exposure. However, the real test period may be too short to achieve steady-state. In addition, metabolism and chemical degradation may occur and large molecules may not permeate sufficiently through membranes into the organism, often lowering BCF values. Thus, experimental determination of BCF may underestimate the environmental risk [14] . In ideal case, the measured value of BCF should be strongly related to the high complexity of bioaccumulation process, taking into account such factors as metabolism, organ-specific bioconcentration, irreversible binding onto proteins, incomplete depuration, and kinetic effects [15] Fish with an average lipid content of 4.8% are preferred model animals for bioconcentration studies due to their relevance as food for many species, including humans [16] , and to the availability of standardized testing protocols. Bioaccumulation is a thermodynamically driven partitioning process between aquatic environment and the lipid tissues of fish, thus, n-octanol is often a satisfactory surrogate for biological lipids [17]. As demonstrated earlier, it is important to know the BCF of all PCBs congeners. The literature data on experimental BCF of PCBs are limited and their measurement is difficult and expensive. Thus, quantitative structure-property relationship (QSPR) methods based on the descriptors derived directly from the molecular structure are vital to supply the missing data independently of experimentation.

The BCF of a chemical is most commonly estimated from established correlations between logBCF and logKOW [17, 18] .However, multilinear QSAR/QSPR models including logKOW are valid only for compounds with logKOW values<6 [19, 20] For highly hydrophobic chemicals (logKOW>6) non-linear [18,21] bilinear [18] and polynomial [22] equations relate logBCF and logKOW. While logKOW models indicate priorities for assessing dangerous substances, they may not provide reliable predictions for unknown BCF.

The objective of this work is to develop a robust QSPR model that could predict the bioaccumulation factor for a diverse set of PCBs, using the general molecular descriptors and to seek

the important features related to the bioaccumulation factor values.

II Methodology

II.1. Dataset:

Known experimental logarithmic BCF values of the 58 PCBs were taken from the literature [23]. The whole set of experimentally observed values lies in the range from 2.64 to 5.97 (<6) clearly indicating that the application of the linear QSAR approach to the studied property is relevant [19, 20].A complete list of the compounds name and their corresponding experimental values of PCBs is shown in Table 1. The data set was divided into two subsets: a training set of 30 compounds and a test set of 28 compounds according to Kennard and Stones algorithm. The training set was used to build the genetic algorithm /SLR and the test set was used to evaluate its prediction ability of the model.

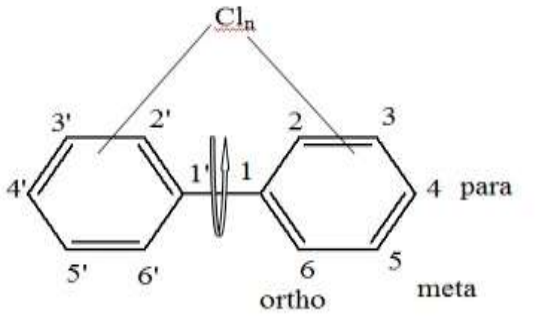

Figure 1. General structural formulae and substitution positions of the PCBs.

Table 1: Values of LogBCF(exp), HATS0v for a set of 58 PCBs. The last 28 chemicals are the test set.

Congener Number	LogBCFexp	HATS0v
PCB 0	2.64	0.063
PCB 3	2.77	0.085
PCB15	3.28	0.106
PCB 4	3.38	0.099
PCB 8	3.57	0.106
PCB 54	3.85	0.129
PCB 18	4.11	0.132
PCB28	4.2	0.126
PCB 40	4.23	0.156
PCB29	4.26	0.142
PCB 77	4.59	0.174
PCB 64	4.6	0.151
PCB 52	4.63	0.16
PCB 87	5.38	0.181
PCB 157	5.39	0.216
PCB 101	5.4	0.18
PCB 136	5.43	0.176
PCB 209	5.44	0.251
Congener Number	**LogBCFexp**	**HATS0v**
PCB 208	5.71	0.237
PCB 174	5.8	0.215
PCB 180	5.8	0.225
PCB 126	5.81	0.209
PCB 141	5.81	0.208
PCB 194	5.81	0.248
PCB 202	5.82	0.222
PCB 198	5.88	0.237
PCB 195	5.92	0.231
PCB 196	5.92	0.232
PCB 169	5.97	0.238
PCB14	3.78	0.143
PCB 7	3.55	0.106
PCB 6	3.8	0.12
PCB 9	3.89	0.121
PCB 5	4.11	0.117
PCB31	4.23	0.143
PCB 70	4.77	0.175
PCB 44	4.84	0.162
PCB 49	4.84	0.154
PCB 47	4.85	0.142
PCB 155	4.93	0.168
PCB 48	5	0.153
PCB 90	5	0.184
PCB 99	5	0.172
PCB 105	5	0.186
PCB 109	5	0.177
PCB 118	5	0.193
PCB 138	5.39	0.199
PCB 148	5.39	0.194
PCB 156	5.39	0.219
PCB 97	5.43	0.181
PCB 151	5.54	0.199
PCB 153	5.65	0.198
PCB 128	5.77	0.194
PCB 182	5.8	0.211
PCB 187	5.8	0.217

PCB 183	5.84	0.211
PCB 191	5.84	0.223
PCB 137	5.88	0.201

2.2. Descriptor Generation:

The structures of the molecules were drawn using Hyperchem 6.03 software [24]. The final geometries were obtained with the semi empirical method PM3.All calculations were carried out at the RHF (restricted Hartree–Fock) level with non configuration interaction. The molecular structures were optimized using the algorithm Polak-Ribiere and a gradient norm limit of 0.001 kcal.A°$^{-1}$.mol^{-1}The resulted geometry was transferred into the software Dragon version 5.3 [25] to calculate 1600 descriptors of the type Geometrical and GETAWAY (Geometry, Topology and Atoms Weighted AssemblY). Descriptors with constant ornear constant values inside each group were discarded. For each pair of correlated descriptors (with correlation coefficient r ≥ 0.95), the one showing the highest pair correlation with the other descriptors was excluded. The GA (Genetic Algorithm) [26] has been considered superior to other methods of variable selection techniques. So, variable selection was performed on the training set, using GA in the MobyDigs version of Todeschini [27] by maximizing the cross-validated explained variance Q^2_{LOO}.

2.3. Kennard and Stone algorithm

Kennard and Stones algorithm [28] has been widely used for splitting datasets into two subsets. This algorithm starts by finding two samples, based on the input variables that are the farthest apart from each other .these two samples are removed from the original dataset and put into the calibration set. This procedure is repeated until the desired number of samples has been selected in the calibration set. The advantages of this algorithm are that the calibration samples always map the measured region of the input variable space completely with respect to the induced metric and that the no validation samples fall outside the measured region. Kennard and Stones algorithm has been considered as one of the best ways to build training and test sets [29,30]. Using Kennard and Stones algorithm the entire set divided into two subsets: training set of 40 compounds, and a test set including the remaining 18 compounds.

2.4. Model Development and Validation:

Models with one variable were performed by the software MOBYDYGS [31] using the Ordinary Least Square regression (OLS) method.

The goodness of fit of the calculated models were assessed by means of the multiple determination coefficients, R^2, and the standard deviation error in calculation (SDEC).

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

Cross validation techniques allow the assessment of internal predictivity (Q^2_{LMO} cross validation; bootstrap) in addition to the robustness of model (Q^2_{LOO} cross validation).

Cross validation methods consist in leaving out a given number of compounds from the training set and rebuilding the model, which is then used to predict the compounds left out. This procedure is repeated for all compounds of the training set, obtaining a prediction for every one. If each compound is taken away one at a time the cross validation procedure is called leave-one-out technique (LOO technique), otherwise leave-more-out technique (LMO technique). An LOO or LMO correlation coefficient, generally indicated with Q^2, is computed by evaluating the accuracy of these "test" compounds prediction.

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{PRESS}{TSS} \qquad (2)$$

The "hat" of the variable y, as is the usual statistical notation, indicates that it is a predicted value of the studied property, and the sub index "i/i" indicates that the predicted values come from models built without the predicted compound.

TSS is the total sum of squares.

The predictive residual sum of squares (PRESS) measures the dispersion of the predicted values. It is used to define Q^2 and the standard deviation error in prediction (SDEP).

$$SDEP = \sqrt{PRESS/n} \qquad (3)$$

A value $Q^2 > 0.5$ is generally regarded as a good result and $Q^2 > 0.9$ as excellent [32, 33].

However, studies [34, 35] have indicated that while Q^2 is a necessary condition for high predictive power a model, is not sufficient.

To avoid overestimating the predictive power of the model LMO procedure (repeated 5000 times, with 5 objects left out at each step) was also performed ($Q^2_{L(5)O}$).

In bootstrap validation technique K n-dimensional groups are generated by a randomly repeated selection of n-objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 8000 times for each validated model.

By using the selected model the values of the response for the test objects are calculated and the quality of these predictions is defined in terms of Q^2_{ext}, which is defined as

$$Q^2ext = 1 - \frac{\sum_{i=1}^{next}(\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{ntr}(y_i - \bar{y}_{tr})^2 / n_{tr}} = 1 - \frac{PRESS / n_{ext}}{TSS / n_{tr}} \quad (4)$$

Here n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively.
The data set was divided according to kennard and stone algorithm into a training set (30 objects) used to develop the QSAR models and a validation set (28 objects), used only for statistical external validation.
Other useful parameters are R^2, calculated for the validation chemicals by applying the model developed on the training set, and external standard deviation error of prediction ($SDEP_{ext}$), defined as:

$$SDEP_{ext} = \sqrt{\frac{1}{n_{ext}}\sum_{i=1}^{n_{ext}}\left(yi - \bar{y}\right)^2} \quad (5)$$

where the sum runs over the test set objects (n_{ext}).

According to Golbraikh and Tropsha[35]. A QSPR model is successful if it satisfies several criteria as follows :

$R^2_{CVext} > 0.5$ (6)
$r^2 > 0.6$ (7)
$(r^2 - r^2_0)/r^2 <$ or $(r^2 - r^{2\prime}0)/r^2 < 0.1$ (8)
$0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ (9)

Here:

$$r = \frac{\sum(y_i - \tilde{y}_i)(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum(y_i - \tilde{y})^2 \sum(\tilde{y}_i - \bar{\tilde{y}}_i)^2}} \quad (10)$$

$$r_0^2 = 1 - \frac{\sum(\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum(\tilde{y}_i - \bar{\tilde{y}}_i)^2} \quad (11)$$

$$r_0'^2 = 1 - \frac{\sum(y_i - y_i^{r_0})^2}{\sum(y_i - \bar{y})^2} \quad (12)$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \quad (13)$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad (14)$$

$$T1 = \frac{\left(r^2 - r^2_0\right)}{r^2} \quad (15)$$

$$T2 = \frac{\left(r^2 - r'^2_0\right)}{r^2} \quad (16)$$

$$Ab = \left[r^2_0 - r'^2_0\right] \quad (17)$$

where r is the correlation coefficient between the calculated and experimental values in the test set; r^2_0 (calculated versus observed values) and r'^2_0 (observed versus calculated values) are the coefficients of determination; k and k' are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively

$y_i^{r_0}$, $\tilde{y}_i^{r_0}$; are defined as $y_i^{r_0} = k\tilde{y}_i$ and $\tilde{y}_i^{r_0} = k'y_i$ and the summations runs over the test set.

2.5. QSAR AD (Applicability Domain)

The AD was discussed by the Williams plot [31,32] of jacknified residuals versus leverages (hat diagonal values (h_i). The jacknifed residuals (or Studentized residuals) are the standardized cross-validated residuals. Each residuals is divided by its standard deviation, which is calculated without the i-th observation. The leverage(h_i). value of a chemical in the original variable space is defined as :

$$h_i = x_i (X^T X)^{-1} x_i^T \quad (i=1,\ldots,n) \quad (18)$$

Where x_i is the descriptor row-vector of the query compound, and X is the n (p+1) matrix of p model parameter values for n training set compounds. The superscript T refers to the transpose of the matrix/vector. The warning leverage value (h^*) is defined as 3(p +1)/ n. When h value of a compound is lower than h^*, the probability of accordance between predicted and actual values is as high as that for the compounds in the training set. A chemical with $h_i > h^*$ will reinforce the model if the chemical is in the training set. But such a chemical in the validation set and its predicted data may be unreliable. However, this chemical may not appear to be an outlier because its residual may be low. Thus the leverage and the jacknified residual should be combined for the characterization of the AD.

3. Results and Discussion

Application of the GA-VSS led to several good models for the prediction of based on different sets of molecular descriptors. The best single dimensional model was constructed using the descriptor HATS0e.All data concerning value of this descriptor and the dependent variable (LogBCF) are summarized in Table 1.
The equation of the optimal model can be written as:

LOG BCF = 1.5779 ± **(**0.1958**)** + 18.538± **(**1.065) HATS0v. (19)

HATS0v is calculated in Dragon software. More information about this descriptor can be found in Dragon software user's guide [25] and the references therein.

All relevant statistical parameters are reported in Table 2.

Values of $R^2$and R^2_{adj} attest the good fitting performances of the model which, moreover, is very highly significant (great value of the Fisher parameter F).

The model is robust, the difference between R^2 and Q^2 is small (1.25%). The model demonstrates a very good stability in internal validation (difference between Q^2_{LOO} and Q^2_{LMO} is about 0.76(%). While bootstrapping confirms the internal predictivity and stability of the model.$SDEP_{ext}$is a little bit different from SDEP. The model works slightly worse in external prediction than in internal prediction.

Some important statistical parameters (as given in Table 3) were used to evaluate the involved descriptor. The t-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute t-values shown in Table 3express that the regression coefficients of the descriptors involved in the GA/SLR model are significantly larger than the standard deviation. The t-probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (i.e., descriptors' interactions). Descriptors with t-probability values below 0.05 (95% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance [36] .The smaller t-probability suggests the more significant descriptor.The t-probability values of the two descriptors are equal to zero, indicating that all of them are highly significant descriptors.The VIF values and the correlation matrix as shown in Table 4 suggest that these descriptors are weakly correlated with each other. The distributions of errors for the entire dataset are given in **Figure2**.As the errors are distributed on both sides of the zero line,one may conclude that there is no systematic error in the model development.The model was also verified by Y-scrambling. **Figure 3** clearly ensures the existence of a linear relationship between LogBCF and the descriptor HATS0v ;As can be observed the permuted responses yield poor predictive models, all having Q^2< 0.2 On the other hand, the correctly ordered LogBCF yield good statistical parameters, and therefore it is located isolated in the plot.

The statistical parameters of Tropsha et all reported in Table 5 were obtained for the test set, which obviously satisfy the generally accepted condition and thus demonstrate the predictive power of the present model:

R^2cv_{ext}= 0.7793>0.5

r^2= 0.8707>0.6

$$T1 = \frac{\left(r^2 - r^2_0\right)}{r^2} = 0.0014<1$$

Or $$T2 = \frac{\left(r^2 - r'^2_0\right)}{r^2} = -0.0046<0.1$$

0.85≤ k= 1.0447<1.15 or 0.85≤ k'= 0.9552≤1.15

Table 2 statistical parameters of a developed model.

n_{tr}	n_{ext}	Q^2_{LOO} (%)	R^2(%)	$Q^2_{LMO/50}$ (%)	Q^2_{BOOT} (%)	R^2_{adj} (%)	Q^2_{ext} (%)
30	*28*	90.28	91.53	*89.52*	89.23	91.23	90.44
SDEC	*SDEP*	*SDEP ext*	*S*	*F*			
0.3	0.321	0.319	0.3106	302.7423			

Table 3. Characteristics of the selected descriptor in the best GA/ SLR model.

Predictor	Coef	SE Coef	T	P
Constant	1.5779	0.1958	8.06	0.000
HATS0v	18.538	1.065	17.40	0.000

Table 4: The statistical parameters of Tropsha *et al.*

R^2cv_{ext}	r^2	k	k'	r^2o
0.7793	0.8707	1.0447	0.9552	0.8695
$r^2'o$	Ab	T1	T2	
0.8747	0.0948	0.0014	-0.0046	

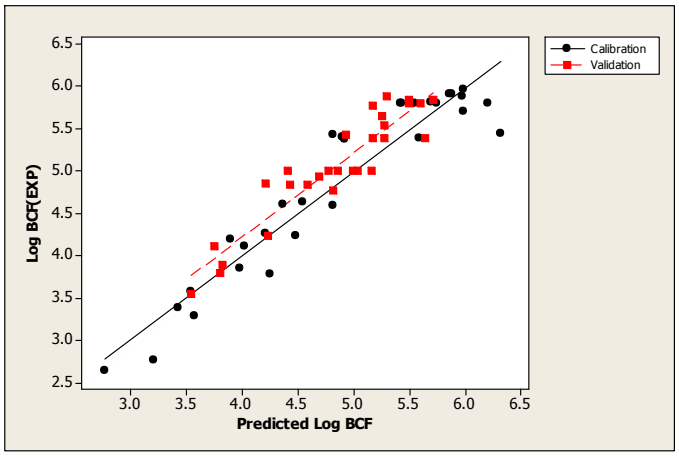

Fig:2 Predicted Log BCF versus experimental LogBCF for the entire dataset.

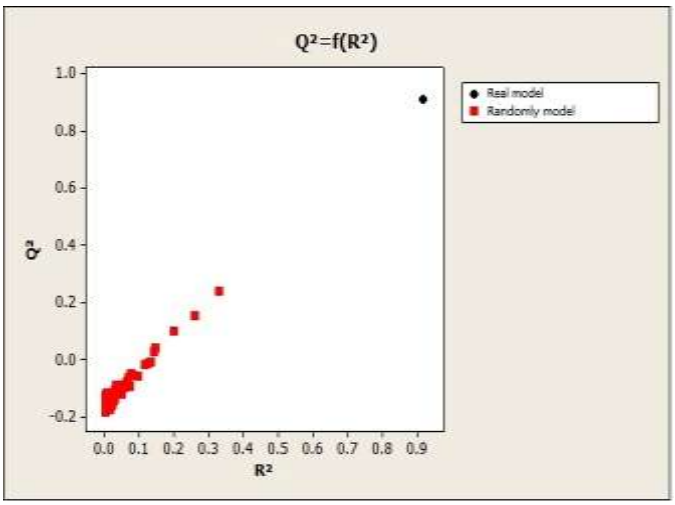

Fig:2 Randomization test associated to the previous QSPR model. Square represent the randomly ordered properties and the circle corresponds to the real properties.

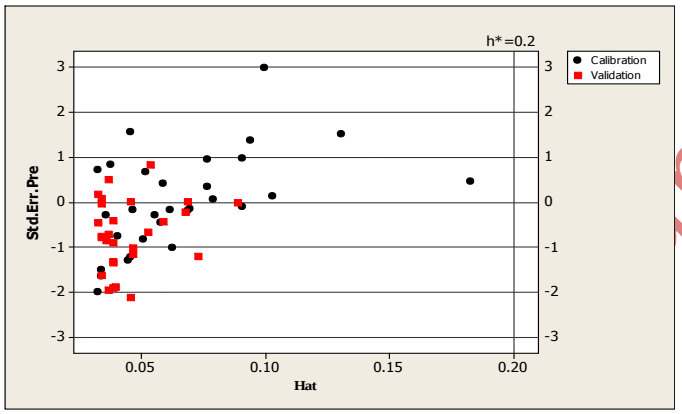

Fig: 3 Williams plot of the current QSRR model

3.2 Mechanistic Interpretation

The selected descriptor and its class and meaning are gathered in the Table 5.

Table 5: selected descriptor and its meaning and class for the best GA/ SLR model.

Descriptor	meaning	class
HATS0v	leverage-weighted autocorrelation of lag 0 / weighted by atomic van der Waals volumes	GETAWAY

HATS0v is a GETAWAY descriptor and correlates with the experimental Log BCF values of 0.957.The GETAWAY descriptors [37,38] have been proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular influence matrix. which is based on the spatial autocorrelation formulas, weighting the molecule atoms by the physic-chemical properties w together with 3D information encoded by the elements of the molecular influence matrix H and influence/distance matrix R.These descriptors, as based on spatial autocorrelation, encode information on the effective position of substituents and fragments in the molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties. The positive sign of HATS0v (equation 20) means that the increase in this descriptor increases the Log BCF.

$$HATS0V = \sum_{i=1}^{nat-1} \sum_{j>1} (V_i . h_{ij}).(V_j . h_j).\delta(0, d_j) \quad (20)$$

Where; nAT is the number of molecule atoms; d_{ij} is the topological distance between atoms i and j; V is the atomic van der Waals volumes ; d is the topological diameter; δ (k; d_{ij}) is a Dirac-delta function (δ =1 if d_{ij} = k, zero otherwise); δ (k; d_{ij}; h_{ij}) is another Dirac-delta function (δ = 1 if d_{ij} = k and h_{ij}>0, zero otherwise).

3.3 Applicability Domain

On analyzing the model applicability domain from Williams plot, all residuals were located within the range of three Standard deviations, and there is no influential compound both for training or prediction set (Figure.3), which means that the model has a good external predictivity.

4. Conclusion

A QSPR model for the estimation of the logarithm of bioaccumulation factor for 58 PCBs was established

According to obtained results it is concluded that the HATS0v can be used succesfully for modeling bioaccumulation factor (Log BCF) of the under study compounds. High correlation coefficient0.9153) and low prediction error (SDEP=0.321;SDEPext=0.319in Log unit)obtained confirm good predictive ability of the model.The QSPR model proposed with the simply calculated molecular descriptor can be used to estimate bioaccumulation factor for new compounds

References

[1] National Research Council (1979), Committee on the Assessment of Polychlorinated Biphenyls in the Environment. Polychlorinated biphenyls: a report; National Academy of Sciences: Washington, U.S.A.

[2] Angulo Lucena, R., Farouk Allam, M., Serrano Jiménez, S. and Luisa Jodral Villarejo, M. A.(2007), "review of environmental exposure to persistent organochlorine residuals during the last fifty years". Current Drug Safety,Vol. 2 No.2, pp. 163-172.

[3] Roveda, A. M., Veronesi, L., Zoni, R., Colucci, M. E. and Sansebastiano, G. (2006), **"**Exposure to polychlorinated biphenyls (PCBs) in food and cancer risk: recent advances",Igiene e Sanita Pubblica,Vol. 62 No.6, pp. 677-696.

[4] Lundqvist, C.,Zuurbier, M., Leijs, M., Johansson, C.,Ceccatelli, S.,Saunders, M.,Schoeters, G., Ten Tusscher, G. and Koppe, J. G. (2006)," The effects of PCBs and dioxins on child health", Acta Paediatrica,Vol.95 No.453,pp.55-64.

[5] Poppenga, R. H. (2000), "Current environmental threats to animal health and productivity", The Veterinary Clinics of North America. Food Animal Practice ,Vol. 16 No.3, pp.545-558.

[6] Bren, U.,Zupan, M., Guengerich, F. P. and Mavri, J.(2006)" Chemical Reactivity as a Tool to Study Carcinogenicity: Reaction between Chloroethylene Oxide and Guanine", The Journal of Organic Chemistry ,Vol. 71 No.11,pp. 4078-4084.

[7] Lebeuf, M., Noël, M.,Trottier, S. and Measures, L. (2007), "Temporal trends (1987-2002) of persistent,bioaccumulative and toxic (PBT) chemicals in beluga whales (Delphinapterus leucas) from the St.Lawrence Estuary, Canada", Sciences of the Total
Environment, Vol.383 No. (1-3), pp.216-231.

[8] Tan, J.,Cheng, S. M., Loganath, A.,Chong, Y. S.and Obbard, J. P. (2007), "Selected organochlorine pesticide and polychlorinated biphenyl residues in house dust in Singapore",Chemosphere ,Vol. 68 No.9, pp.1675-1682.

[9] Borrell, A., Cantos, G., Aguilar, A., Androukaki, E. and Dendrinos, P. (2007) "Concentrations and patterns of organochlorine pesticides and PCBs in Mediterranean monk seals
(Monachus monachus) from Western Sahara and Greece", Science of the Total Environment , Vol.381 No. (1-3), pp.316-325.

[10] Klánová, J.,Kohoutek, J., Kostrhounová, R. and Holoubek, I. (2007),"Are the residents of former Yugoslavia still exposed to elevated PCB levels due to the Balkan wars?. Part 1: air sampling in Croatia, Serbia, Bosnia and Herzegovina", Environment International, Vol. 33 No. 6, pp.719- 726.

[11] Hansch, C. (1969),"Quantitative approach to biochemical structure-activity relationships", Accounts of Chemical Research, Vol.2 No.8, pp.232-239.

[12] Hansch, C.and Leo, A. (1979),"Substituent Constants for Correlation Analysis in Chemistry and Biology", John Wiley & Sons,New York.

[13] Voutsas, E., Magoulas, K. and Tassios, D.(2002), "Prediction of the bioaccumulation of persistent organic pollutants in aquatic food webs", Chemosphere ,Vol.48,pp. 645– 651.

[14] Franke, C. (1996), "How meaningful is the bioconcentration factor for risk assessment?" Chemosphere,Vol. 32 No.10,pp. 1897–1905.

[15] Franke, C., Studinger, G., Berger, G., Bohling, S., Bruckmann, U., Cohors-Fresenborg, D.and Jtihnck, U. (1994), "The assessment of bioaccumulation",Chemosphere ,Vol.29 No.7,pp.1501–1514.

[16] Schecter, A., Cramer, P., Boggess, K., Stanley, J., Päpke, O., Olson, J., Silver, A. and Schmitz, M. (2001), "Intake of dioxins and related compounds from food in the U.S. population". Journal of Toxicology and Environmental Health, Part A,Vol. 63 No.1.pp. 1–18.

[17] Barron, M. G. (1990), "Bioconcentration. Will water-borne organic chemicals accumulate in aquatic animals? ", Environmental science & technology. Vol.24, pp. 1612–1618.

[18] Bintein, S., Devillers, J. and Karcher, W. (1993), "Nonlinear dependence of fish bioconcentration on n-octanol/water partition coefficient", SAR and QSAR in

Environmental Research,Vol. 1, pp.29–39.

[19] Isnard, P. and Lambert, S. (1988)," Estimating bioconcentration factors from octanol–water partition coefficient and aqueous solubility", Chemosphere,Vol. 17,pp. 21–34.

[20] Mackay, D. (1982), "Correlation of bioconcentration factors", Environmental Science &Technology,Vol. 16,pp.274–278.

[21] Dimitrov, S. D., Mekenyan, O. G. and Walker, J. D. (2002), "Non-linear modeling of bioconcentration using partition coefficients for narcotic chemicals",SAR and QSAR in Environmental Research,Vol. 13 No.1, pp.177–184.

[22] Connell, D. W. and Hawker, D. W. (1988)," Use of polynomial expressions to describe the bioconcentration of hydrophobic chemicals by fish",Ecotoxicology & Environmental
Safety, Vol.16 No.3, pp.242-257.

[23] Ivanciuc, T., Ivanciuc, O. and Klein, D. J. (2006), "Modeling the bioconcentration factors and bioaccumulation factors of polychlorinated biphenyls with posetic quantitative super-structure/activity relationships (QSSAR)", Molecular Diversity,Vol.
10,pp.133–145.

[24] Hyperchem™ (2000), Release 7, Hypercube for Windows, Molecular Modeling System.

[25] Todeschini, R., Consonni,V.and Pavan, M. (2006),DRAGON Software for the Calculation of Molecular Descriptors. Release 5.4 for Windows,Talete s.r.l.,Milano, Italy.

[26] Leardi, R., Boggia, R. and Terrile, M. (1992) "Genetic algorithms as a strategy for feature selection", Journal of Chemometrics, Vol.6, pp. 267-281.

[27] Todeschini, R., Ballabio, D., Consonni, V., Mauri, A. and Pavan, M. (2009), "MOBYDIGS, Software for Multilinear Regression Analysis and Variable Subset

Selection by Genetic Algorithm",Release 1.1 for windows,Milano,Italy.
[28] Kennard, R. and Stone, L.A (1969), Technometrics, Vol.11, p.137.
[29] Tropsha,A. ,Gramatica,P . and Grombar, V.K(2003),"The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models",QSAR and Combinatorial Science ,Vol.22,pp.69.
[30] Wu,W.,Walczac,B.,Massart,D.L;Heuerding,S.;Erni,F.,Last,I.R. and Prebble,K.A. (1996),"Artificial neural networks in classification of NIR spectral data: design of the training set",Chemometrics and Intelligent Laboratory Systems,Vol.33,pp. 35-46.

[31] MOBYDIGS – Models BY Descriptors In Genitic Selection – ver. 1.1 for Windows, Talete S.r.l., Milano, Italy.

[32] 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, Environmental Health Perspectives, Vol. 111(10),1361-1375.

[33]Tropscha A., Gramatica P. & Grombar V.K., 2003.The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR & Combinatorial Science, Vol. 22(1), 69-77
[34]Kubinyi H., Hamprecht F.A. & Mietzner T., 1998. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices, Journal of Medicinal Chemistry, Vol.41(14), 2553-2564.

[35] Golbraikh, A. and Tropsha, A. (2002), "Beware of q^2!",Journal of Molecular Graphics and Modelling, Vol.20 No.4, pp.269-276.

[36] Ramsey, L. F.and Schafer, W. D. (1997), "The Statistical Sleuth", Wadsworth Publishing Company,U.S.A.
[37] Consonni, V.; Todeschini, R.; Pavan, M. J. Chem. Inf. Comput. Sci. 2002, 42, 682
[38] Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. J. Chem.Inf. Comput. Sci. 2002, 42, 693.

