Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. 80 (4) 499–508 (2015)
JSCS–4733
UDC 510.644:539.16:544.112:66–914.4+
536.421.4:536.423+532+532.11
Original scientific paper

QSPR study of supercooled liquid vapour pressures of polybrominated diphenyl ethers using the molecular distance–edge vector index

LONG JIAO[1,2]*, XIAOFEI WANG[1], SHAN BING[1], ZHIWEI XUE[3] and HUA LI[2]

[1]*College of Chemistry and Chemical Engineering, Xi'an Shiyou University, Xi'an 710065, P. R. China,* [2]*College of Chemistry and Materials Science, Northwest University, Xi'an 710069, P. R. China and* [3]*No. 203 Research Institute of the Nuclear Industry, Xianyang 712000, P. R. China*

(Received 17 July, revised 30 August, accepted 1 September 2014)

Abstract: The quantitative structure property relationship (QSPR) for supercooled liquid vapour pressures (p_L) of polybrominated diphenyl ethers (PBDEs) was investigated. The molecular distance–edge vector (MDEV) index was used as the structural descriptor. The quantitative relationship between the MDEV index and log p_L was modelled by multivariate linear regression (MLR) and an artificial neural network (ANN). The leave-one-out cross validation and *k*-fold cross validation were performed to assess the prediction ability of the developed models. For the MLR method, the prediction root mean square relative error (*RMSRE*) of the leave-one-out cross validation and the *k*-fold cross validation were 9.95 and 9.05, respectively. For the ANN method, the prediction *RMSRE* of the leave-one-out cross validation and the *k*-fold cross validation were 8.75 and 8.31, respectively. It was demonstrated that the established models were practicable for predicting the log p_L values of PBDEs. The MDEV index was quantitatively related to the log p_L of PBDEs. MLR and linear ANN were practicable for modelling this relationship. Compared with the MLR, the ANN method exhibited slightly higher prediction accuracy. Subsequently, an MLR model, the regression equation of which was log p_L = = $0.2868M_{11} - 0.8449M_{12} - 0.0605$, and an ANN model, which was a two--input linear network, were developed. The two models could be used to predict the log p_L value of each PBDE.

Keywords: QSPR; PBDEs; supercooled liquid vapour pressures; molecular distance–edge vector index; artificial neural network.

INTRODUCTION

Polybrominated diphenyl ethers (PBDEs) are a group of brominated organic compounds that have been widely used as flame-retardants in many products,

* Corresponding author. E-mail: mop@xsyu.edu.cn
doi: 10.2298/JSC140716087J

such as building materials, electronics, furnishings, coatings, plastics, textiles, *etc*.[1,2] There are 209 PBDE congeners, distinguished by the number and position of bromine atoms. Although the production of some PBDEs has been restricted under the Stockholm Convention since 2010, PBDEs already exist in many environmental compartments, such as water, air, soil, vegetation, animals and humans.[3–6] They have gained increasing attention due to their environmental persistence, tendency to bioaccumulate through the food chain and risk to ecosystems and human health.[1,7,8] Being volatile organic compounds, the volatility of PBDEs may affect their fate and distribution in the environment. The super-cooled liquid vapour pressures at 298.15 K (p_L) is usually used to describe the volatility of non-polar organic pollutants and the equilibrium partitioning of a compound between the vapour phase and the liquid state. Thus, a quantitative study on the p_L of PBDEs is of great importance to understand the distribution, fate and transportation of PBDEs in the environment. A study on the p_L is also necessary for the design of extraction and cleanup techniques and for the environmental modelling of PBDEs.[4,9,10] Many studies focused on the p_L of PBDEs have been reported.[1,4–14] However, the determination of p_L is still a hard difficult due to the complexity of the analytical methods, high cost of experiments, and lack of chemical standards of some PBDEs.[4,9,10] To overcome this problem, much attention has been given to the quantitative structure property relationship (QSPR) method for a preliminarily estimate the p_L of PBDEs. Several QSPR models for the p_L of PBDEs were proposed.[9–14] These studies demonstrated that it is practicable to predict the p_L of PBDEs by using QSPR methods. However, some of the proposed QSPR models were not global models for the P_L of PBDEs and the used structural descriptors were not satisfactory. There are many limitations in the use of such models.[4] In addition, the development of some QSPR models was not easy because quantum chemical descriptors were used.[9,10] Actually, the generation and selection of quantum chemical descriptors are always time-consuming and complicated in QSPR studies. The need to develop more easy-to-use QSPR models for the p_L of PBDEs still exists. Therefore, the QSPR model for the p_L of PBDEs was investigated in this work. The molecular distance–edge vector (MDEV) index[15–19] was used as the structural descriptor of PBDEs. Multivariate linear regression (MLR) and linear artificial neural network (L-ANN) were employed to model the quantitative relationship between the p_L and the MDEV index of PBDEs.

CALCULATION METHODS

Data set

The MDEV index was calculated according to the approach presented in the following section. The calculated MDEV indexes are listed in Table S-I of the Supplementary material to this paper. The experimental p_L values of the 22 PBDEs given in Table S-II of the

Supplementary material are the super-cooled liquid vapour pressures at 298 K. The values of the p_L were taken from the literature.[9]

Root mean square relative error (*RMSRE*) was used to indicate the prediction performance of the obtained QSPR models. The *RMSRE* is defined as:

$$RMSRE = \sqrt{\frac{\sum (RE_i)^2}{n}} \tag{1}$$

where RE_i is the relative error of the i^{th} sample, and n is the number of samples.

MDEV index

When calculating the MDEV index of a PBDE molecule, each non-hydrogen atom is regarded as a point and each chemical bond is considered as an edge. The whole molecule is regarded as a topological graph. The relative electronegativity of each bromine atom and benzene ring is defined as 1. Correspondingly, the MDEV index is defined as:

$$M_{kl} = \sum_{j \geq i} \frac{1}{d^2_{ik,\,jl}}, \; k, l = 1,2 \text{ and } l \geq k \tag{2}$$

In Eq. (2), k and l are the type of atoms (k = 1 or l = 1 denotes the bromine atom, and k = 2 or l = 2 denotes the benzene ring); items i and j are the coding number of a bromine atom or benzene ring. Additionally, i and j belongs to the k^{th} and l^{th} type, respectively. The $d_{ik,jl}$ value represents the nearest relative distance between the i^{th} and j^{th} atom. For example, $d_{i1,j1}$ means the shortest relative distance between the i^{th} and j^{th} bromine atom. The relative distance between the two adjacent non-hydrogen atoms is defined as d = 1. According to Eq. (2), there are three elements, M_{11}, M_{12} and M_{22}, in the MDEV index for a PBDE molecule. For instance, the MDEV index of 2,2′,4,4′-PBDE should be calculated as follows:

$$M_{11} = \left(\frac{1}{4}\right)^2 + \left(\frac{1}{6}\right)^2 + \left(\frac{1}{8}\right)^2 + \left(\frac{1}{8}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{4}\right)^2 = 0.1940$$

$$M_{12} = \left(\frac{1}{1}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{1}\right)^2 + \left(\frac{1}{6}\right)^2 + \left(\frac{1}{1}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{1}\right)^2 + \left(\frac{1}{6}\right)^2 = 4.1806$$

$$M_{22} = \left(\frac{1}{2}\right)^2 = 0.25 \tag{3}$$

Obviously, the M_{22} value of each PBDE is equal to 0.25. Thus, M_{11} and M_{12} were used to quantitatively describe the structure of the PBDEs.

Artificial neural network

ANN[20-30] is a multivariate calibration method capable of modelling complex functions. Its basic processing unit is the neuron (node). An ANN comprises a number of neurons organized in different layers. A linear artificial neural network,[26-30] is a kind of neural network having no hidden layers, but an output layer with fully linear neurons (*i.e.*, linear neurons with a linear activation function). It is the simplest ANN and is usually used to develop a linear model, In a L-ANN, the neurons between the input and output layers fully connect, while the neurons in the same layer do not. The basic architecture of a L-ANN is shown in Fig. 1.

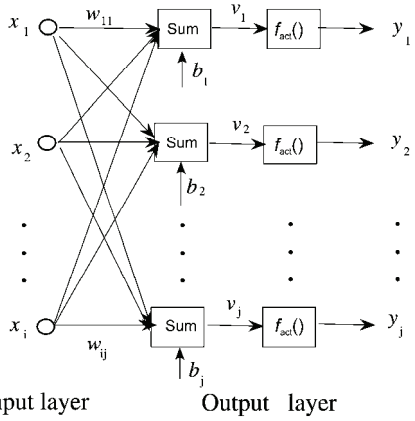

Fig. 1. The architecture of the employed linear artificial neural network.

In Fig. 1, x_i (i = 1, 2,…, n), y_j (j =1, 2,…, m) and w_{ij} are the input variables, output variables and the element of connection weight matrix **W**, respectively, and b_j is the bias vector, which corresponds to the thresholds. The symbol $f_{act}()$ denotes the activation function. Prior to a training procedure, the input and output variables require normalization. When the network is executed, it effectively multiplies the input variables by the weights matrix, and then adds the bias vector. Hence, the post synaptic potential (PSP) function of a neuron can be described as follows:

$$v_j = \sum_{i=1}^{n} x_i w_{ij} + b_j \tag{4}$$

Routinely, the activation function used in a L-ANN is a linear function described by:

$$y_j = v_j \tag{5}$$

As there are no non-linear functions or hidden neurons in the network, a L-ANN is ideal for dealing with linear problems. Actually, the training of a linear network means finding the optimal setting for the weight matrix **W** to minimize the root-mean-squared-error of the calibration set. In order to achieve this aim, known samples are generally divided into two sets: a training set and a verification set. The network is trained using the training set, but is also tested after each epoch using the verification set. The training should be stopped once a deterioration in the verification error is observed. Overfitting and overlearning can be effectively avoided in this way. Although the verification set is used to identify the best network, actually, training algorithms do not use the verification set to adjust the weights of the network. A standard pseudo-inverse linear optimization algorithm[26] is usually employed to train the network. This algorithm uses the singular value decomposition technique to calculate the pseudo-inverse of the matrix needed to set the weights in a linear output layer, so as to find the least mean squared solution. Essentially, it guarantees finding the optimal setting for the weights in a linear layer.

The main difference between a MLR and a L-ANN is the optimization algorithm. In a MLR, the aim of the least square algorithm is to minimize the sum of the squared residuals of the training set. As for L-ANN, the aim of the training algorithm is to minimize the root mean squared error of the verification set.[26] Thus, the prediction ability of a L-ANN is usually superior to that of a MLR.

Leave-one-out cross validation

The leave-one-out cross validation[18-20,27,31] is a conventional algorithm for estimating the predictive performance of a multivariable calibration model. Usually, practical calibration experiments have to be based on a limited set of available samples. The idea behind the leave--one-out cross validation algorithm is to predict the property value of each sample in turn with the calibration model developed with the remaining samples. When applying the algorithm to a dataset with N samples, the calibration modelling is performed N times, each time using N–1 samples for the modelling and one sample for testing. Thus, the procedure of leave-one-out cross validation can be divided into N segment. In each segment i (i = 1,…, N), there are three steps: 1) taking sample i out as temporary 'test set', which is not used to develop the calibration model, 2) developing a calibration model with the remaining N–1 samples, 3) testing the developed model with sample i, calculating and storing the prediction error of the sample. The advantage of leave-one-out cross validation over random sub-sampling is that each sample is used for both training and validation, and each sample is used for validation exactly once.

k-*Fold cross validation*

In k-fold cross validation,[32,33] the original N samples are randomly partitioned into k equal size subgroups, called folds (if k = N, this is equivalent to the leave-one-out cross validation). Of the k subgroups, a single subgroup is retained as the test set for testing the model, and the remaining k−1 subgroups are used as the training set for generating the model. The cross validation process is then repeated k times, with each of the k subgroups used exactly once as the test set. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation.

Software

All data processing was realized with subroutines developed under Matlab (ver. 7.0). The computation was performed on a personal computer equipped with Core2 T9400 processor.

RESULTS AND DISCUSSION

The MDEV index of the investigated 22 PBDEs was calculated. The obtained MDEV indexes are listed in Table S-I of the Supplementary material. Clearly, the MDEV indexes of different PBDE molecules are quite different. It was demonstrated that MDEV index can describe the structural differences among these PBDE molecules. Thus, it is reasonable to use the MDEV index as the structural descriptor for developing a QSPR model of PBDEs.

MLR model

Generally, a simple model should always be chosen in preference to a complex model unless the latter fits the data better. Thus, first, it was investigated whether MLR is practicable to model the quantitative relationship between the MDEV index and log p_L value of these PBDEs. The MDEV index was used as the independent variable and log p_L as the dependent variable to develop the regression model. In order to assess the predictive ability of the developed model, two validation methods, leave-one-out cross validation and k-fold cross validation, were performed. The results of the leave-one-out cross validation are presented in Table S-II of the Supplementary material, from which it could be seen

that the predicted log p_L values were in agreement with the experimental values of log p_L. For the 22 compounds, the *RMSRE* of prediction was 9.95. Moreover, the predicted log p_L values were plotted *versus* the experimental values (shown in Fig. 2a) and the plot showed a linear relationship ($y = 0.9470x - 0.1433$ with $R = 0.9853$) between the predicted and experimental log p_L values. Subsequently, *k*-fold cross validation was performed to assess further the predictive ability of the MLR model. In this procedure, the 22 samples were partitioned into 7 folds. There were 4 samples in each fold. The prediction results of the *k*-fold cross validation are given in Table S-III of the Supplementary material. As shown in this table, the predicted log p_L values were still in agreement with the experimental values. For the 7 folds, the total prediction *RMSRE* was 9.05. A plot of the predicted log p_L values *versus* the experimental values (presented in Fig. 2b) shows a linear relationship ($y = 0.9591x - 0.0722$ with $R = 0.9818$). The results of the leave-one-out and *k*-fold cross validation demonstrate that the MDEV index is quantitatively related to the log p_L of these PBDEs. MLR was shown to be practicable for modelling the quantitative relationship between the MDEV index and log p_L. Obviously, a linear QSPR model based on the MDEV index could be used to predict the log p_L values of PBDEs. Hence, an MLR model was developed using the 22 PBDEs listed in Table S-I of the Supplementary material. The obtained regression equation was $\log p_L = 0.2868M_{11} - 0.8449M_{12} - 0.0605$. The R^2, standard error of the estimate and the F value of the regression model were 0.9800, 0.2007 and 466.5, respectively. Thus, this model could be used to predict the log p_L value of all the PBDEs.

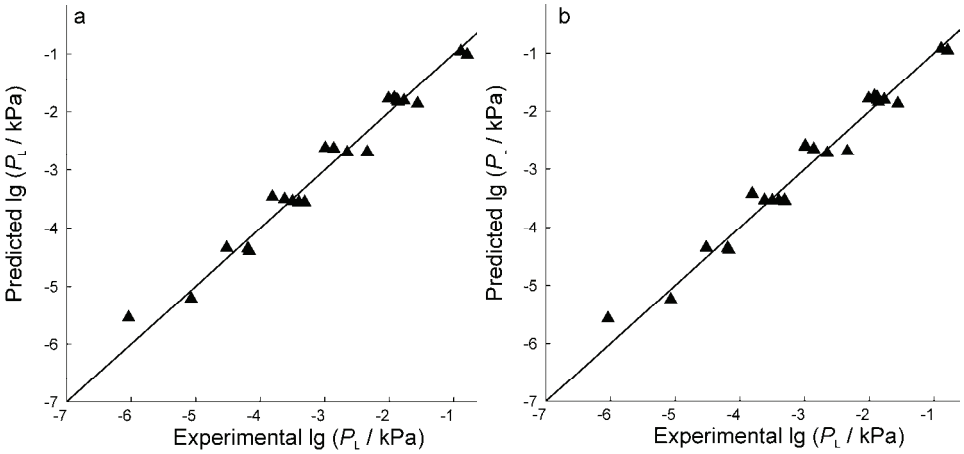

Fig. 2. Experimental log p_L values *vs*. the log p_L values predicted by the MLR model: a) results of the leave-one-out cross validation and b) results of the *k*-fold cross validation.

L-ANN model

Besides MLR, L-ANN, which is the simplest artificial neural network, is a commonly used linear calibration method in QSPR studies. Thus, it was investigated whether a better model could be established using a L-ANN. A 2–1 L-ANN (*i.e.*, 2 input variables and 1 output variable in the network) was used to establish the calibration model. The used activation function is presented as Eq. (5). The MDEV index and log p_L were used as input and output variables, respectively. Prior to the training procedure, the input and output variables were normalized. In each run of the ANN, the verification set comprised four randomly selected samples.

The leave-one-out cross validation and *k*-fold cross validation were still performed to assess the prediction performance of the developed model. The results of the leave-one-out cross validation are listed in Table S-II of the Supplementary material. As shown in Table S-II, the predicted log p_L values were in good agreement with the experimental log p_L values. For all 22 compounds, the *RMSRE* of the prediction was 8.75. Moreover, the predicted log p_L value were plotted *versus* the experimental ones (presented in Fig. 3a) and the plot showed a linear relationship ($y = 0.9515x - 0.1309$ with $R = 0.9873$) between the predicted and experimental log p_L values. The results of the *k*-fold cross validation are given in Table S-III. As shown in the table, the predicted log p_L values were also in good agreement with the experimental log p_L values. For the 7 folds, the total prediction *RMSRE* was 8.31. The plot of the predicted log p_L *versus* the experimental log p_L values (presented in Fig. 3b) showed a linear relationship ($y = 0.9643x - 0.0557$ with $R = 0.9826$). Obviously, the prediction accuracy of the L-ANN model was slightly better than that of the MLR model. Using the L-ANN was slightly better than using the MLR for the prediction of the log p_L values of

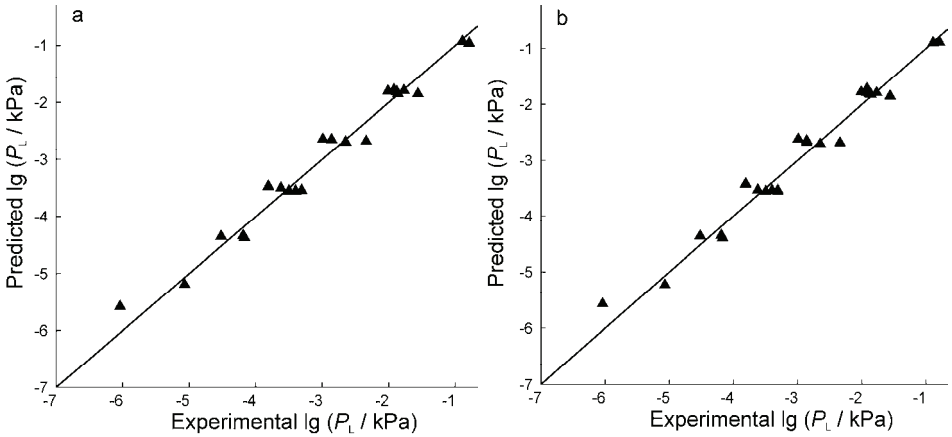

Fig. 3. Experimental log p_L values *vs.* the log p_L values predicted by the L-ANN model: a) results of the leave-one-out cross validation and b) the results of the *k*-fold cross validation.

the PBDEs. It was demonstrated that L-ANN is a practicable and promising method for modelling the quantitative relationship between the MDEV index and log p_L of PBDEs. Thus, a 2–1 L-ANN model was developed by using the 22 PBDEs listed in Table S-1 of the Supplementary material as the calibration set. The verification set for training the model comprises 4 randomly selected samples. The input and output variables were normalized in the training procedure. This model could be used to predict the log p_L values of all PBDEs.

CONCLUSIONS

A method for developing a QSPR model of the supercooled liquid vapour pressures of PBDEs was investigated. The MDEV index, which can be generated easier than quantum chemical descriptors, was used as the structural descriptor of the PBDEs. Accordingly, the use of the MDEV index as structural descriptor was more convenient than using quantum chemical descriptor when developing a QSPR model for the p_L of PBDEs. Calibration models between the MDEV index and log p_L were established using MLR and ANN methods. The predictive ability of the developed models was assessed using leave-one-out cross validation and *k*-fold cross validation. The validation result demonstrated that the MDEV indexes of the PBDEs were quantitatively related to the log p_L values of the PBDEs. Thus, it was reasonable to establish the QSPR model for the log p_L values of the PBDEs based on the MDEV index. In addition, the result indicated that both MLR and L-ANN are practicable for modelling the quantitative relationship between the MDEV index and log p_L value of PBDEs. In summary, the proposed method should be easy-to-use and practicable for predicting the log p_L values of PBDEs. Therefore, an MLR model and an ANN model, which could be used to predict the log p_L values of all PBDEs, were developed. The regression equation of the MLR model was $\log p_L = 0.2868M_{11} - 0.8449M_{12} - 0.0605$. The ANN model was a 2–1 L-ANN that was trained using the 22 known PBDEs as a calibration set.

SUPPLEMENTARY MATERIAL

MDEV indexes of the investigated PBDEs and results of the leave-one-out cross validation and *k*-fold cross validation, Tables S-I–S-III. respectively, are available electronically from http://www.shd.org.rs/JSCS/, or from the corresponding author on request.

Acknowledgments. The work was supported by the National Natural Science Foundation of China through projects No. 21305108 and 20975081, the Project Supported by Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2014JM2039) and the Innovative Research Team of the Xi'an Shiyou University (Program No. 2013QNKYCXTD01).

И З В О Д

QSPR СТУДИЈА НАПОНА ПАРЕ СУПЕРОХЛАЂЕНИХ ПОЛИБРОМОВАНИХ ДИФЕНИЛ ЕТАРА ПРИМЕНОМ МОЛЕКУЛСКОГ ВЕКТОРСКОГ ИНДЕКСА

LING JIAO[1,2], XIAOFEI WANG[2], SHAN BING[1], ZHIWEI XUE[3] и HUA LI[2]

[1]Xi'an Shiyou University, Xi'an, P. R. China, [2]Northwest University, Xi'an, P. R. China и [3]Research lnstitute of Nuclear Industry, Xianyang, P. R. China

Испитивана је квантитативна релација између структуре и својстава (QSPR) за напон паре суперохлађених полибромованих дифенил етара (PBDE). Као структурни дескриптор, примењен је молекулски векторски индекс MDEV. Корелација између напона паре и MDEV анализирана је мултиваријантном линеарном регресијом (MLR) помоћу вештачких неуронских мрежа (ANN). Показано је да добијени модели омогућују квантитативно предвиђање напона паре за PBDE. Методом ANN се постижу нешто тачнији резултати него са MLR.

(Примљено 17. јула, ревидирано 30. августа, прихваћено 1. септембра 2014)

REFERENCES

1. H. A. Anderson, P. Imma, L. Knobeloch, M. Turyk, J. Mathew, C. Buelow, V. Persky, *Chemosphere* **73** (2008) 187
2. EPA, http://www.epa.gov/ace/pdfs/Biomonitoring-PBDEs.pdf (accessed in Jan, 2014)
3. Stockholm Convention, http://chm.pops.int/TheConvention/ThePOPs/ListingofPOPs/ /tabid/2509/Default.aspx (accessed on 10th July, 2014)
4. C. Y. Yue, L. Y. Li, *Environ. Pollut*. **180** (2013) 312
5. M. Alaee, R. J. Wenning. *Chemosphere* **46** (2002) 579
6. F. M. Jaward, N. J. Farrar, T. Harner, A. J. Sweetman, K. C. Jones, *Environ. Sci. Technol.* **38** (2004) 34
7. Y. W. Wang, C. Y. Zhao, W. P. Ma, H. X. Liu, T. Wang, G. B. Jiang, *Chemosphere* **64** (2006) 515
8. D. J. Watkins, M. D. McClean, A. J. Fraser, J. Weinberg, H. M. Stapleton, T. F. Webster, *Environ. Int.* **59** (2013) 124
9. H. Y. Xu, J. W. Zou, Q. S. Yu, Y. H. Wang, J. Y. Zhang, H. X. Jin, *Chemosphere* **66** (2007) 1998
10. Z. Y. Wang, X. L. Zeng, Z. C. Zhai, *Sci. Total Environ*. **389** (2008) 296
11. E. Papa, S. Kovarich, P. Gramatica, *QSAR Comb. Sci.* **28** (2009) 790
12. A. Wong, Y. D. Lei, M. Alaee, F. Wania, *J. Chem. Eng. Data* **46** (2001) 239
13. S. A. Tittlemier, G. T. Tomy, *Environ. Toxicol. Chem.* **20**(2001) 146
14. S. A. Tittlemier, T. Halldorson, G. A. Stern, G. T. Tomy, *Environ. Toxicol. Chem.* **21** (2002) 1804
15. S. S. Liu, H. L. Liu, Z. N. Xia, C. Z. Cao, Z. L. Li, *J. Chem. Inf. Comput. Sci*. **39** (1999) 951
16. C. S. Yin, W. M. Guo, T. Lin, S. S. Liu, R. Q. Fu, Z. X. Pan, L. S. Wang, *J. Chin. Chem. Soc*. **48** (2001) 739
17. H. H. Liu, X. Xiao, J. Qin, Y. M. Liu, *J. Chongqing Inst. Tech.* **19** (2005) 67 (In Chinese)
18. L. Jiao, X. Wang, H. Li, Y. Wang, *J. Serb. Chem. Soc*. **79** (2014) 965
19. L. Jiao, Z. W. Xue, G. F. Wang, X. F. Wang, H. Li, *Chemometr. Intell. Lab. Syst.* **137** (2014) 91
20. L. Jiao, *Chemosphere* **80** (2010) 671

21. M. Jalali-Heravi, M. Asadollahi-Baboli, P. Shahbazikhah, *Euro. J. Med. Chem.* **43** (2008) 548
22. M. H. Fatemi, Z. G. Chahi, *SAR QSAR Environ. Res.* **23** (2012) 155
23. R. Smiljanic, D. Lazic, M. Gligoric, M. Jotanovic, Z. Zivkovic, I. Mihajlovic, *J. Serb. Chem. Soc.* **76** (2011) 1163
24. X. J. Yao, A. Panaye, J. P. Doucet, H. F. Chen, R. S. Zhang, B. T. Fan, M. C. Liu, Z. D. Hu, *Anal. Chim. Acta* **535** (2005) 259
25. M. Fatemi, Z. Ghorbannezhad, *J. Serb. Chem. Soc.* **76** (2011) 1003
26. Statsoft, http://www.statsoft.com/textbook/neural-networks (accessed on 10th July, 2014)
27. L. Jiao, H. Li, Chemometr. *Intell. Lab. Syst.* **103** (2010) 90
28. Y. X. Zhang, H. Li, A. X. Hou, J. Havel, *Chemometr. Intell. Lab. Syst.* **82** (2006) 165
29. C. S. Yin, Y. Shen, S. S. Liu, Q. S. Yin, W. M. Guo, Z. X. Pan, *Comput. Chem.* **25** (2001) 239
30. W. J. Zhang, X. Q. Zhong, G. H. Liu, *Stoch. Environ. Res. Risk, A* **22** (2008) 207
31. H. A. Martens, P. Dardenne, *Chemometr. Intell. Lab. Syst.* **44** (1998) 99
32. M. Zounemat-Kermani, O. Kisi, T. Rajaee, *Appl. Soft Comput.* **13** (2013) 4633
33. Wikipedia, http://en.wikipedia.org/wiki/Cross-validation(statistics) (accessed on 10th July 2014).