Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 157 (2019) 551–560

Energy Procedia

www.elsevier.com/locate/procedia

Technologies and Materials for Renewable Energy, Environment and Sustainability, TMREES18, 19–21 September 2018, Athens, Greece

Soil contamination by pesticides: molecular modeling of octanol / organic carbone partition coefficient

Amel BOUAKKADIA[a,b*], Noureddine KERTIOU [a], Hayette BOUAKKADIA[c], Djelloul MESSADI [a]

[a]*Environmental and Food Safety Laboratory, Badji Mokhtar University, Faculty of sciences, Department of chemistry, Annaba, Algeria,*
.[b]Abbes Laghrour University, Faculty of sciences and technology, Khenchela, Algeria.
[c] Badji Mokhtar University, Faculty of sciences, Department of biology, Annaba, Algeria.

Abstract

QSPR methods are often used to estimate the physicochemical properties of organic compounds and to predict their behavior in the environment. QSPR models were developed for the prediction of octanol/organic carbone partition coefficient (Koc) of an heterogeneous set of pesticides. The approaches based on multilinear regression (MLR), artificial neural networks (ANN), every time associated with genetic algorithm (GA) selection of the most important variables, lead to models of very different qualities. The modeling of octanol/organic carbone partition coefficient of a heterogeneous mixture of pesticides show that the various statistics for the sets of training and validation (multiple coefficients of determination and prediction; roots of squared errors averages) attest to the superiority of non-linear models (ANN) and their relevance.

Keywords: Octanol/organic carbone partition coefficient; Molecular descriptors; QSPR methods; Pesticides; Multiple linear regression; artificial neural networks.

*Corresponding author. Tel: + 0-000-000-0000; Fax: +0-000-000-0000
E-mail address: amelbouakkadia@yahoo.fr

10.1016/j.egypro.2018.11.219

1. INTRODUCTION

Physicochemical properties of an organic chemical compound play an important role in determining its distribution and fate in the environment. Vapor pressures (Pv), aqueous solubility ($S_{w,L}$) and n-octanol/water partition coefficients (K_{ow}) are key physicochemical properties that can be used for assessing environmental partition and transport of organic substances [1]

In addition, knowledge of the octanol/organic carbone partition coefficient is necessary for predicting the evolution of pollutants in the environment. The retention of a substance in the soil depends on the properties of the substance and the composition of the soil (usually organic matter content), K is the ratio of the product concentrations in the soil, in g / g of soil and in water, in g / cm3 of solution.

QSPR methods are often used to estimate the physicochemical properties of organic compounds and to predict their behavior in the environment. Chemometric methods can be used to describe how the physicochemical properties vary according to the characteristics of the molecular structure expressed in terms of appropriate molecular descriptors. QSPR models can also provide a general overview of the molecular structure that influences these properties. This statistical technique is often used to replace costly biological tests or experiments of a given physico-chemical property with calculated descriptors, and can also be used to predict responses of interest for new compounds [2]. The results of these models reveal substantial differences in the fields of application and in the predictive capabilities [3-5]. The aim of this study is to found a statistical model for the prediction of the octanol/organic carbone partition coefficient (Koc) of 78 pesticides. The QSPR model was constructed using multiple linear regression (MLR), and artificial neural network (ANN) methods, and its performance validated. The model obtained shows which descriptors play a significant role in the Koc variation of these pesticides.

2. MATRIALS AND METHODS

The 78 pesticides are taken from [6] and are listed in Table I. The data were presented as the logarithm of Koc to reduce the range of variation. The data set is randomly separated into a training set of 60 compounds and a test set of 18 compounds.

All the two-dimensional structures of the molecules were drawn and the geometry optimizations of molecules were performed by HyperChem [7] using the MM+ molecular mechanics force field. Then a more precise optimization is done with Polack-Ribiere algorithm until the root mean square gradient 0.01, at the Austin model 1 (AM1), semi- empirical method level.. The HyperChem output files are transferred to Dragon [8] to calculate twenty classes of the descriptors. Descriptors having a constant value for all structures in the data set are also discarded, many of descriptors were removed because they were including zero values and they did not have important information of molecular structures [9].

Once the molecular descriptors are generated, multiple linear analysis regression and variable subset selection were performed by MOBYDIGS [10], using ordinary least squares regression (OLS) method and Genetic Algorithms for variable subset selection (GA-VSS) [11].

The outcome of the application of the genetic algorithms is a population of 100 regression models, ordered according to their decreasing internal predictive performance, verified by Q^2. The models with lower Q^2 are those with fewer descriptors. First of all, models with 1- 2 variables were developed by the all- subset- method procedure in order to explore all the low dimension combinations. The number of descriptors was subsequently increased one by one, and new models were formed. The best models are selected at each rank, and the final model must be chosen from among them. This has to be sufficiently correlated and, at the same time, protect against any over parameterization, which would lead to a loss of predictive power for molecules outside training set [9].

From a statistical view point the ratio of the number of samples (n) to the number of descriptors (m) should not be too low. Usually, it is recommended that $n/m \geq 5$ [12]. The GA was stopped when increasing the model size did not increase the Q^2 value to any significant degree. Particular attention was paid to the collinearity of the selected molecular descriptors: by applying the QUIK rule (Q Under Influence of K) [13] a necessary condition for the model validity. Acceptable models are only those with a global correlation of [X + y] block (K_{XY}) greater than the global correlation of the X block (K_{XX}) variable, X being the molecular descriptors and y the response variable. Therefore, when there were models of similar performance, those with higher ΔK ($K_{XY} - K_{XX}$) were selected and further verified.

Table I. Names, values of log Koc of pesticides examined in the study.

N°	Composé	log Koc_{Exp},	N°	Composé	log Koc_{Exp},
1	Ametryn	2,48	41	Prometryn	2,58
2	Bensulfuron methyl	2,57	42	Propazine	2,08
3	Bromacil	1,51	43	Prosulocarb	3,23
4	Butylate	2,6	44	Prosulfuron	1,48
5	Carbetamide	1,77	45	Quizalofop ethyl	2,71
6	Chorimuron ethyl	2,04	46	Rimsulfuron	1,78
7	Chloroxuron	3,26	47	Siduron	2,62
8	Clopyralide	0,66	48	Simazine	2,18
9	Clopyralid olamine	0,78	49	Terbacil	1,51
10	Cyanazine	2,26	50	Terbutylazine	2,55
11	Cycloate	2,34	51	Terbutryn	3,3
12	2,4-D	1,22	52	Thifensulfuron methyl	1,7
13	2,4-D dimethylammonium	2,02	53	Triasulfuron	2,15
14	2,4,5-T-trolamine	1,41	54	Tribenuron methyl	1,72
15	Desmedipham	3,47	55	Triclopyr	1,77
16	Desmetryn	2,18	56	Triclopyr butotyl	2,89
17	Dichlorprop	1,41	57	Triflusulfuron methyl	1,76
18	Dipropetryn	2,95	58	Vernolate	2,41
19	Diuron	2,6	59	Vinclozolin	2,39
20	EPTC	2,24	60	Propaquizafop	2,61
21	Ethametsulfuron methyl	2,03	61	Atrazine*	2,05
22	Fluazifop	1,31	62	Chlorpropham*	2,64
23	Fluazifop butyl	3,77	63	Chlorsulfuron*	1,6
24	Fluazifop P butyl	3,76	64	2,4-D-methyl*	2
25	fluoxypyr eptyl	4,02	65	Dichlorprop P*	1,76
26	Glyphosate	2,2	66	Difenxuron*	2,3
27	Haloxyfop ethoxyethyl	2,03	67	Fenoxaprop*	2,16
28	Hexazinone	1,57	68	Fenoxapropethyl*	4,05
29	Isoproturon	1,72	69	Fluometuron*	1,87
30	Lenacil	2,22	70	MCPA*	1,45
31	Linuron	2,54	71	Mecoprop*	1,3
32	MCPA isoctyl	3	72	MecopropP*	1,3
33	MCPB	1,3	73	Metamitron*	2,78
34	Methabenzthiazuron	2,15	74	Metoxuron*	2,05
35	Methazol	3,48	75	Pebulate*	2,63
36	Metribuzin	1,59	76	Tebuthiuron*	1,9
37	Metsulfuron methyl	1,65	77	Thiobencarb*	2,95
38	Napropemide	2,48	78	Triallate*	3,3
39	Phenmedipham	3,38			
40	Prometon	2,18			

*Test

The robustness of the models and their predictivity were evaluated by both Q^2_{LOO} and bootstrap, the bootstrapping was repeated 8000 times for each selected model. Moreover, to avoid overestimating, the predictive power of the models using the leave- more- out procedure was also performed.

The proposed model was also checked for reliability and robustness by permutation testing: new models are recalculated for randomly reordered response (Y scrambling). Evidence that the proposed model is well founded, and not just the result of chance correlation, is provided by obtaining new models on the set with randomized responses that have significantly lower R^2 and Q^2 than the original model. When checked by the Y- scrambling procedure our suggested QSPR model verifies this condition [9].

The predictive power of the regression model developed on the selected training set is estimated on the predicted values of prediction set chemicals, by the external Q^2 that is defined [14] as:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} \left(\hat{y}_{i/i} - y_i \right)^2 / n_{ext}}{\sum_{i=1}^{n_{tr}} \left(y_i - \overline{y}_{tr} \right)^2 / n_{tr}} \quad (1)$$

Where y_i and $\hat{y}_{i/i}$ are, respectively, the measured and predicted (over the prediction set) values of the dependent variable, and $\overline{y}_{tr}$ the averaged value of the dependent variable for the training set.n_{tr} and n_{ext} are the number of training set objects and the number of objects in the external set, respectively.

The chemical domain of the studied compounds in the model was verified by the leverage approach to verify prediction reliability [15, 16]. The Williams plot shows standardized residuals in prediction plotted against leverage of each compound and makes it possible to verify the presence of outliers (compounds with jackknifed residual greater than 3 standard deviation units) and objects very influential in the determination of model parameters (compounds with leverage greater than 3(m+1)/n, where m is the number of the model parameters and n the number of the objects used to calculate the model). Conversely, when the leverage value of a compound is lower than the critical value.

Artificial neural networks (ANN) are computer models derived from a simplified concept of the brain [17- 19]. The introduction of ANN in quantitative structure- activity relationship (QSAR) studies and quantitative structure-property relationship (QSPR) studies can be traced back to the early 1990s [20].

The artificial neural networks (ANN) are a very powerful tool in QSAR and QSPR studies. Indeed, a ANN can find nonlinear relationships between the structure of the molecules and their activity or property [20].

In contrast to MLR, artificial neural networks (ANN) are capable of recognizing highly nonlinear relationships. The flexibility of ANN enables it to discover more complex relationships in experimental data, when it is compared with the traditional statistical models [21].

The basic entity of a ANN is the formal neuron. Its action consist in summing weighted inputs and producing an output signal through an activation function. Although the activation functions can have several forms, the most commonly used is the sigmoid function. The nonlinear nature of the sigmoid function plays an important role in the performances of the ANN.

Topologically a ANN presents three types of layers:

- One input layer (with number of neurons corresponding number of molecular descriptors);
- One (or more) hidden layer(s) with adjustable numbers of neurons;
- One output layer with a number of neurons depending on the modeled activity or property. This layer generates the calculated outputs [20].

3. RESULTS AND DISCUSSION

3.1. MLR analysis

A best six- parameters equation was obtained, which is as the following:

log Koc = - 0,065 (±0,4316) + 0,0420 (±0,007) **ALOGP2** - 6,31 (±1,500) **HATS5v** + 3,04 (±0,7237) **E2e** + 0,564 (±0,0861) **ATS6m**+ 0,181 (±0,0322) **SEigv** + 0,806 (±0,4807) **R5m** (2)

R^2= 75,84 Q^2_{LOO} = 68,02 Q^2_{LMO} (20%) = 65,10 Q^2_{BOOT} = 64,72

EQMP = 0,411 EQMC= 0,358 K_{xx}= 35,26 K_{xy} = 40,45

n= 60 S= 0,380 F= 27,72

n_{ext}= 18 Q^2_{ext} = 61,71 $EQMP_{ext}$ = 0,452

Correlation matrix as shown in Table II suggests that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation.

Table II. Correlation matrix

	log Koc	ALOGP2	HATS5v	E2e	ATS6m	SEigv
ALOGP2	0,618					
	0,000					
HATS5v	-0,468	-0,156				
	0,000	0,234				
E2e	0,245	0,190	-0,065			
	0,059	0,146	0,619			
ATS6m	0,399	0,266	-0,122	-0,062		
	0,002	0,040	0,354	0,640		
SEigv	0,024	-0,141	0,018	-0,195	-0,652	
	0,856	0,283	0,894	0,136	0,000	
R5m	-0,322	-0,091	0,653	-0,043	0,266	-0,524
	0,012	0,491	0,000	0,745	0,040	0,000

The high absolute t-values shown in Table III express that the regression coefficients of the descriptors involved in the MLR model are significantly larger than the standard deviation. The t- probability of a descriptor cans describle 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 percent confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance [22]. The smaller t- probability suggests the more significant descriptor. The t- probability values of five descriptors are very small, indicating that all of them are highly significant descriptors. Models would not be accepted if they contain descriptors with VIFs above a value of five [23].

Table III. Characteristics of the selected descriptors in the best MLR model

	X	Dx	t- value	t- probability	VIF
Constant	-0.0655	0.4316	-0.15	0.880	
ALOGP2	0.0419	0.0070	5.96	0.000	1.164
HATS5v	-6.3070	1.5000	-4.21	0.000	2.576
E2e	3.0402	0.7237	4.20	0.000	1.229
ATS6m	0.5635	0.0861	6.54	0.001	2.069
SEigv	0.1814	0.0322	5.63	0.019	3.224
R5m	0.8062	0.4807	1.68	0.099	3.672

Strong correlations appear between several descriptors confirmed by a small difference ΔK (Kxy - Kxx), lower than the limit value ΔK = 5.

The calculated and experimental log Koc values from Eq. (2) for the training and test set are showed in Figure 1.

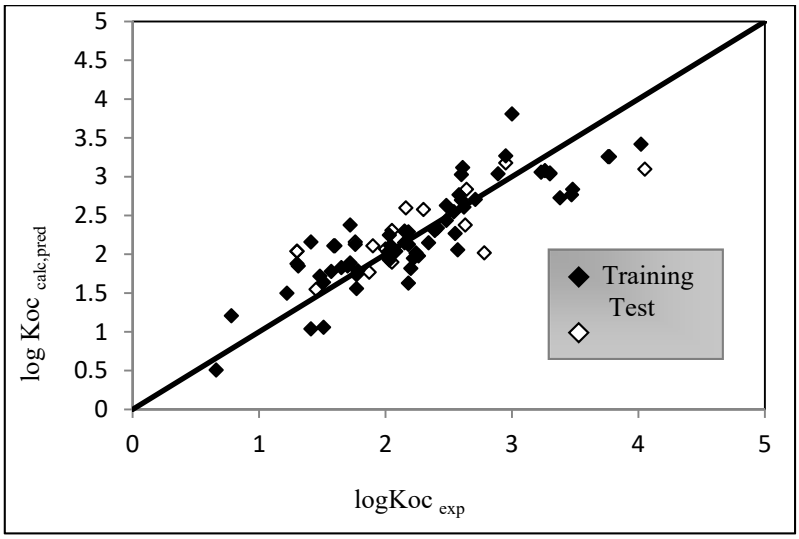

Fig1. Plot of predicted *vs*. experimental log Koc for the entire data set.

According to the figure 1 it was clear that the calculated log Koc values were very similar to the experimental values.

Before a QSPR model is put into use for screening compounds, its applicability domain must be defined and predictions for only those compounds that fall in this domain can be considered as reliable.

The AD of the MLR model model was analyzed in the Williams plot (shown in Figure 2).

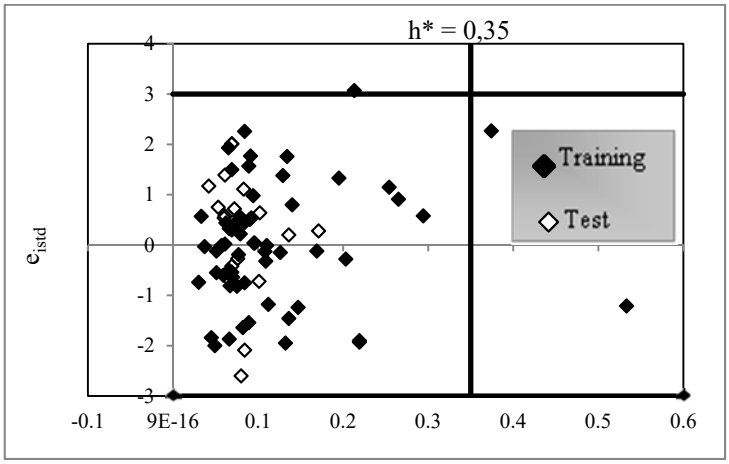

Fig 2. Williams plot for the entire data set.

Only one compound shows a high residual (about 3 standard deviation units) as is also well evident in the Williams plot (Figure 2).The latter also shows that is an aberrant object in the determination of MLR model parameters. All the objects present a leverage smaller than the control value (h*) represented by the vertical straight line in the plot, excepted two compounds of the training set, are a structurally influential compounds.

The results for the randomized models can be compared with the real starting one only by representing in a plot the statistical coefficients R^2 and Q^2. This is depicted in figure 3.Tha figure shows weak statistics (Q^2 <0.2, R^2 <0.3) for the modified vectors, while the representative point of the real model, which is isolated in the graph, presents good statistical parameters, which guarantees the existence of a (multi) linear relationship between log Koc and the selected descriptors.

The modified vector statistics of the log octanol / organic carbon partition coefficient are smaller than those of the real model.

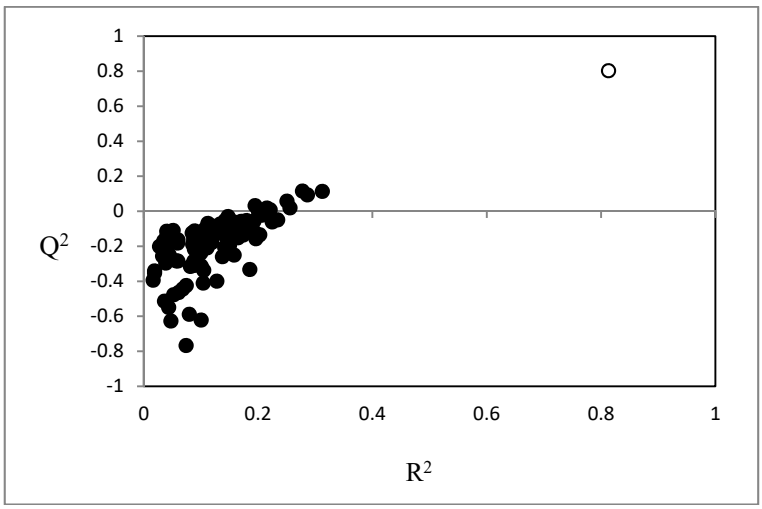

Fig 3. Randomization test associated to previous QSPR model. Circles represent the randomly ordered , and star corresponds to the real .

3.2. ANN analysis

Neural network with three layers and complete connections between neurons was trained. The input layer comprises the subset of descriptors used in MLR model (Eq 2), and the calculated octanol /organic carbon partition coefficient (log Koc) constitute the output layer. After several trials, we selected a hidden layer with five neurons and 600 cycles, the figure 4 explains this choice.

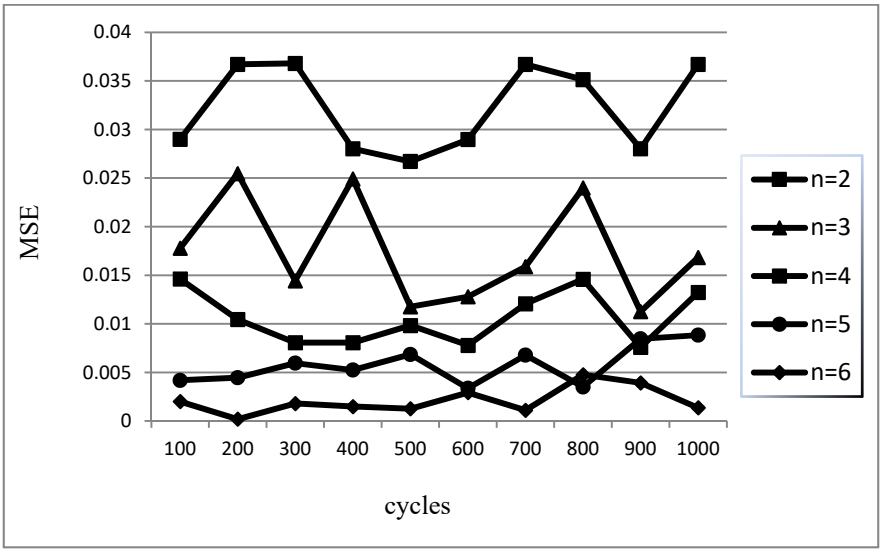

Fig 4. Choice of the number of neurons of the hidden layer and the number of cycles

A sigmoid transfert function was used. The optimal structure adopted is reproduced in the table IV:

Table IV: Optimal structure adopted for the neural network

Inputs	06 (descriptors)
Output	01 (log Koc)
Hidden layer	One hidden layer
Number of neurons	05
Learning Algorithm	Retropropagation of the error gradient
Learning function	Hyperbolic tangent (hidden layer) Linear (output layer)

R^2= 81,30 Q^2_{LOO} = 80,22 Q^2_{LMO} (20%) = 79,32 Q^2_{BOOT} = 80,72

SDEP = 0,211 SDEC= 0,249 K_{xx}= 35,26 K_{xy} = 56,15

n= 60 S= 0,298 F= 47,72

n_{ext}= 18 Q^2_{ext} = 81,46 $SDEP_{ext}$ = 0,220

The values of R^2 show the quality of the fit, while the small difference between R^2 and Q^2_{LOO} gives information on the robustness of the model, which is also highly significant (high value of F). The small difference between Q^2_{LOO} and $Q^2_{LMO\ /\ 20}$ demonstrates good stability in internal validation, and bootstrap validation (Q^2_{boot}) confirms both the good internal prediction capability and the stability of the model. External statistical validation (Q^2_{ext}) attests to the good predictive ability of compounds that did not participate in the calculation of the model.

The plot in figure 5 indicates that there was a significative correlation between calculated and observed log Koc

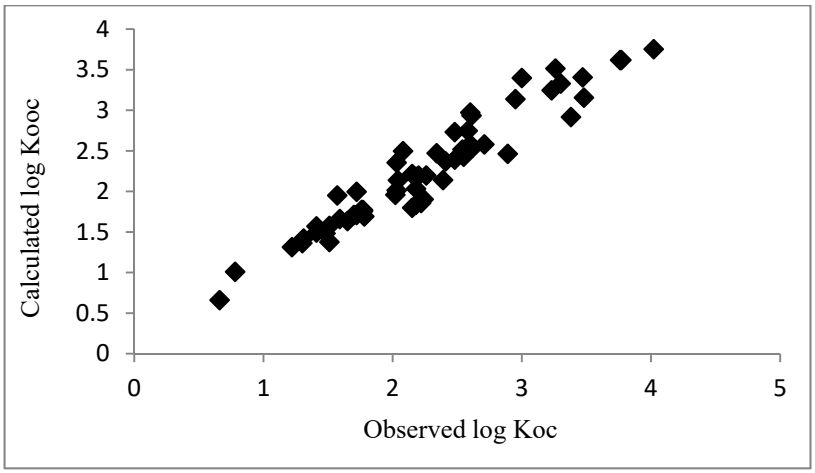

Fig 5. Calculated vs. observed log Koc values

4. Conclusion:

A quantitative- structure property relationship analysis has been performed on the logarithm of octanol/carbone partition corfficient for 78 different pesticide compounds by using MLR and ANN methods. The performance of the NN was superior to that of MLR and this may indicate the presence of non- linearity in the data since the efficiency of descriptors was increased.

REFERENCES

[1] H. Y. Xu, J. Y. Zhang, J. W. Zou, X. S. Chen,. "QSPR Models for The Physicochemical Properties of Halogenated Methyl-Phenyl Ethers". *Journal of Molecular Graphics and Modelling*, 26 (2008): 1076–1081.

[2] H. Kubinyi. "Variable Selection in QSAR Studies: I, An Evolutionary Algorithm". *Quant, Struct,- Act, Relat*, 13(1994): 285- 294.

[3] D. Mackay, W. S. Shiu, K. C. Ma. Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences. R. S. Boethling, D. Mackay, eds, Lewis, Boca Raton, FL, USA (2000).

[4] J. C. Dearden, G. Schüürmann. "Quantitative Structure- Property Relationships for Predicting Henry's law Constant From Molecular Structure". *Environ, Toxicol, Chem*, 22 (8)(2003): 1755- 1770.

[5] E. Estrada, E. J. Delgado, J. B. Alderate, G. A. Jana. "Quantum- Connectivity Descriptors in Modeling Solubility of Environmentally Important Organic Compounds". *J, Comput, Chem*, 25(14)(2004): 1787- 1796.

[6] O. C. Hansen. "Quantitative Structure-Activity Relationships (QSAR) and Pesticides". Teknologisk Institute. Pesticides Research No. 94(2004).

[7] Hyperchem TM Release 6,03 for windows, Molecular Modeling System (2000).

[8]] R. Todeschini, V. Consonni, A. Mauri, M. Pavan. DRAGON Software – version 5.4-TALETEsrl (2005)

[9] A. Bouakkadia, L. Lourici, D Messadi. "Modeling and Prediction of Octanol/Water Partition Coefficient of Pesticides Using QSPR Methods". *Management of Environmental Quality*, 28 (4)(2017): 579-592.

[10] R. Todeshini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan. MOBY DIGS Software for Multilinear

Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 1.1 for windows, Milano (2009).
[11] D. Leahy, J. J Morris, P. J Taylor. "Model Solvent Systems for QSAR. Part 3". An LSER Analysis of the "Critical Quartet:" New Light on Hydrogen Bond Strength and Directionality. *Journal of the Chemical Society, Perkin Transactions 1*, 2(1992): 705–731.
[12] J. Xu, H. Zhang, L. Wang, G. Liang, L. Wang, X. Shen, W. Xu. "QSPR Study of Absorption Maxima of Organic Dye- Sensitized Solar Cells Based On 3D Descriptors". *Spectrochimica Acta Part A*, 76(2010): 239-247.
[13] R. Todeschini, A. Maiocchi, V. Consonni. "The K Correlation Index: Theory Development and Its Application in Chemometrics". *Chemom, Int. Lab. Syst,* 46 (1999): 13 – 29
[14] L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. M. Blair, W. S. Branham, S. L. Dial, C. L. Moland, D. M. Sheehan. "QSAR Models Using a Large Diverse Set of Estrogens". *Journal of Chemical Information and Computer Science*, 41(2001): 186- 195.
[15] A. Tropsha, P. Gramatica, V. K. Gombar. "The Importance of Being Earnest: Validation is The Absolute Essential for Successful Application and Interpretation of QSPR Models". *QSAR & Combinatorial Science*, 22 (2003): 69- 76.
[16] L. Eriksson, J. ,Jaworska A. Worth, M. Cronin, R. M. McDowell, P. Gramatica. "Methods for Reliability, Uncertainty Assessment, and Applicability Evaluations of Regression Based and Classification QSARs". *Environmental Health Perspectives,* 111 (10)(2003): 1361-1375.
[17] T. Kohonen, G. Barna, R. Chrisley. Statistical Pattern Recognition with Neural Networks: Benchmarking Studies. Laboratory of Computer and Information Sciences. Helsinki University of Technology Rakentajanaukio 2C, SF- 02150 Espoo, Finland, (1988).
[18] J. A. Anderson. "Data Representation in neural networks". Al Expert June, (1990): 30- 37.
[19] L. Boddy C. W. Morris, J. W. T. Wimpenny. "Introduction to Neural Networks". Bnary 2(1990): 179- 185.
[20] J. Devillers. 3Neural Networks in QSAR and drug design3. Academic Press. Harcrout Brace & Company Publishers. London, San Diego, New York, boston, Sydney, Tokyo, Toronto (1996).
[21] O. Deeb, M. Goodarzi. "Predicting the Solubility of Pesticide Compounds in Water Using QSPR Methods". *Molecular Physics.* 108 (2)(2010): 181- 192.
[22] L. F. Ramsay, W. D. Schafer. The Statistical Sleuth, Wadsworth Publishing Company, Belmont. (1997).
[23] A. J Holder, D. M Yourtee, D. A White, A. G Galaros. R. J Smith. "Chain Melting Temperature Estimation for Phosphatidyl Cholines by Quantum Mechanically Derived Quantitative Structure Property Relationships". *Computer- Aided Molecular Design*, 17(2)(2003): 223- 230.