Research Article

Journal of Environmental And Sciences (ISSN 2836-2551)

Prediction of ecotoxicological classification risk index for soil of some organic compounds.

Mohammad Hosein Fatemi , Hamideh Ahangar Darabi

Chemometrics laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran

***Corresponding author**

Fatemi Mohammad Hossein ,
Chemometrics laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran.
Tel : 00981125242931,
Fax : 00981125342350,
Email : mhfatemi@umz.ac.ir

Received Date : June 11, 2024
Accepted Date : June 12, 2024
Published Date : July 12, 2024

ABSTRACT

Ecotoxicological classification risk index for soil (ECRIS), is a new classification system specific for soil risk assessment, which gives a comparative indication of the risk linked to environmental contamination by any chemical. In this work this parameter was estimated by quantitative structure–activity relationship approaches by using interpretable molecular descriptors. Linear and nonlinear models were developed using multiple linear regressions (MLR) and artificial neural network (ANN) methods. Robustness and reliability of the constructed MLR and ANN models were evaluated by using the leave-one-out cross-validation method, which produces the statistics of Q2 MLR = 0.84, Q2 ANN = 0.93. Furthermore, the chemical applicability domains of these models were determined via leverage approach. The results of this study indicated the ability of developed QSPR models in the prediction of ECRIS of various chemicals from their calculated molecular structural descriptors.

Keywords : Ecotoxicological classification risk index for soil; quantitative structure–activity relationship; artificial neural network; molecular descriptor; multiple linear regression.

1. INTRODUCTION

Today consideration of environmental risk assessment of chemical pollutants is very important. Several indicators for reporting environmental and human health conditions have been published and indicator frameworks have also been published for chemicals (Bunke and Oldenburg 2005), hazardous wastes (Peterson and Granados 2002) and hazardous material at landfill sites (Peterson and Williams 1999). Some scoring and ranking systems have been adopted by authorities and regulatory centres mainly as first screening tools to identify the chemicals with greatest potential for adverse effects (Huijbregts et al. 2000). For instance, the SCRAM scoring and ranking assessment model (Snyder et al. 2000) is one of these and one of the few systems that also takes the uncertainty into account when there is no data available. SCRAM is limited to chemicals found in the environment, because its aim is mainly to screen and order chemicals based on their profile of persistence, bioaccumulation and toxicity. Some indicators have been developed as decision support system tools, to assess the potential environmental or economic consequences of pesticide management systems (HAIR 2006; United Nation 2007). The indicators should track temporal risk trends in agricultural pesticide usage on different geographical scales (field scale, regional scale, national scale) and should follow up the progress in meeting pesticide reduction goals. HAPERITIF (Calliera et al. 2006) is one of these indicators for monitoring pesticide risk trends attributable to dietary pesticide exposure on various geographic and temporal scales, while ERIP (Finizio et al. 2001) is related to the ecotoxicological effects in soil. Soil contamination from point sources is a worldwide problem most often related to current activities, industrial plants no longer in operation, past industrial accidents and improper municipal and industrial waste disposals. One important criteria for assessment of chemicals, is ecotoxicological classification risk index for soil (ECRIS) (Senese et al. 2010). It is a semi-quantitative index for estimating risk based on several toxicity data and on various kinds of exposure information. Evaluating ecological risk is complex, since it requires detailed knowledge of the biotic and abiotic components of the considered ecosystem, in order to obtain a realistic estimate of all the exposure pathways of the contaminants. Such an approach is not only very expensive in terms of human and economic and time resources, but it also needs support by developments and integration of different scientific areas. Therefore developing of theoretical methods for prediction of environmental risk of pollutant is very important. One of these methods is quantitative structure – property relationship (QSPR) approaches. Quantitative structure–property relationship (QSPR) is one of the most promising methods, which explore a pattern in data by using descriptors derived from molecular structure to predict the activity/property of new and untested chemicals possessing similar molecular features.

A number of QSPR studies reported:

In a promising work, QSPR modeling of soil sorption coefficients (KOC) of Pesticides using SPA-ANN and SPA-MLR were reported by N. Goudarzi and co-workers (Goudarzi et al. 2009). In this study A quantitative structure−property relationship (QSPR) study was conducted to predict the adsorption coefficients of some pesticides. The successive projection algorithm feature selection (SPA) strategy was used as descriptor selection and model development method. Modeling of the relationship between selected molecular descriptors and adsorption coefficient data was achieved by linear (MLR) and nonlinear (ANN) methods. The QSPR models were validated by cross-validation as well as application of the models to predict the KOC of external set compounds, which did not contribute to model development steps. Both linear and nonlinear methods provided accurate predictions, although more accurate results were obtained by the ANN model. The root-mean-square errors of test set obtained by MLR and ANN models were 0.3705 and 0.2888, respectively. Another work is Development of QSAR's in soil ecotoxicology: Earthworm toxicity and soil sorption of chlorophenols, chlorobenzenes and chloroanilines were reported by A.M. Van Gestel and W.C. Ma (Van Gestel and Ma 1993). In this study Soil adsorption and the toxicity of four chloroanilines for earthworms were investigated in two soil types. The toxicity tests were carried out with two earthworm species, Eisenia andrei and Lumbricus rubellus. LC50 values in mg kg−1 dry soil were recalculated towards molar concentrations in pore water using data from soil adsorption experiments. An attempt has been made to develop Quantitative Structure Activity Relationships (QSAR's) using these results and data on five chlorophenols and dichloroaniline in four soils and five chlorobenzenes in two soils published previously. Significant QSAR relationships were obtained between 1) adsorption coefficients (log K om) and the octanol/water partition coefficient (log k ow), and 2) LC50 values (in itμmol L−1 soil pore water) and log K ow . It can be concluded that both earthworm species tested are equally sensitive to chlorobenzenes and chloroanilines, E. andrei is more sensitive than L. rubellus to chlorophenols. Moreover QSPR study on the soil-water partition coefficient of polychlorinated biphenyls by using artificial neural network were done by L. Jiao (Jiao 2012). They reported the practicable quantitative structure property relationship (QSPR) model for predicting the soil-water partition coefficient, Koc, of 16 polychlorinated biphenyls (PCBs). The structure of the investigated PCBs is encoded by five quantum structural descriptors and on topological index. The calibration model of Koc was developed by using artificial neural network (ANN). The input variables of ANN were generated from 6 structural descriptors by using principal component analysis (PCA). Leave one out cross validation was carried out to assess the predictive ability of the developed model. The prediction RMS%RE for the 16 PCBs is 6.35. The R2 between the predicted and experimental logKoc is 0.8522. It is demonstrated that ANN combined with PCA is a practicable method for developing QSPR model for Koc of these PCBs. Also Development of QSARs for the toxicity of chlorobenzenes to the soil dwelling springtail Folsomia candida were reported by D. Giesen and co-workers (Giesen et al. 2012). The purpose of their study was to developed quantitative structure-activity relationships (QSARs) for the toxicity of nine chlorinated benzenes to the soil-dwelling collembolan Folsomia candida in natural LUFA2.2 (Landwirtschaftliche Untersuchungs und Forschungsanstalt [LUFA]) standard soil and in Organisation for Economic Co-operation and Development artificial soil. Toxicity endpoints used were the effect concentrations causing 10% (EC10) and 50% (EC50) reduction in the reproduction of the test organism over 28 d, while lethal effects on survival (LC50) were used for comparisons with earlier studies. Chlorobenzene toxicity was based on concentrations in interstitial water as estimated using nominal concentrations in soil and literature soil–water partition coefficients. Additionally, for LUFA2.2 soil the estimated concentrations in interstitial water were experimentally determined by solid-phase microextraction measurements. Measured and estimated concentrations showed the same general trend, but significant differences were observed. With the exception of hexachlorobenzene, estimated EC10 and EC50 values were all negatively correlated with their logKow and QSARs were developed. However, no correlation for the LC50 could be derived and 1,2,4,5-tetrachlorobenzene and hexachlorobenzene had no effect on adult survival at all. The derived QSARs may contribute to the development of better ecotoxicity-based models serving the REACH program. In the present work we try to generate QSPR models based on MLR and ANN to predict the ECRIS of some organic compounds.

2. MATERIALS AND METHODS

The main steps involved in developing a QSPR model are (a) selection of the data set, (b) calculation of molecular descriptors, (c) fitting the statistical model, (d) validation of the model and (e) Assessing the applicability domain.

2.1. Data set

Data set included 60 common molecules that were found in various landfills leachate of north Italy and are shown in **Table 1** (Senese et al. 2010). The ECRIS values of data set ranged from 1.32 to 58.44 for 2-Imidazolidinthyone and 4.4'-(Methylethylidene) bis-phenol, respectively. Data set was splitted to training, internal and external test sets by Y- ranking method, that each of them has 49, 6 and 5 members, respectively.

Table 1

Data set and corresponding observed MLR and ANN predicted values of ECRIS

Number	Chemical name	(ECRIS) EXP	(ECRIS) MLR	(ECRIS) ANN
1	4.4'-(Methylethylidene)bis-phenol	58.44	53.23	56.98
2	Dichloro-Benzophenone	57.00	48.47	57.46
3	N,N'-dicyclohexyilthiourea	44.51	40.86	44.15
4	4-Chloro-3-methyl-phenol	44.47	31.85	42.49
5^{i}	p-Terbuthyl-phenol	43.90	24.42	37.24
6	2,4-Bis-1-methylethylphenol	43.82	41.71	45.31
7	Isothyocyanate cyclohexane	40.38	43.78	41.65
8	4.4'-Methylenebis-phenol	39.02	43.46	38.80
9	2,6-Bis-(1,1-dimethylethyl)-phenol	37.52	38.41	36.23
10^{e}	Benzyl-butyl-phthalate	37.40	43.04	56.66
11	N,N'-Dicyclohexylurea	35.49	37.96	35.10
12	2-Methyl-thyobenzothiazole	28.53	21.97	28.99
13	Dimethylphenol	26.58	14.93	28.92
14	a,a,a,a-Tetramethylbenzen-dimethanol	23.58	19.87	24.76
15^{i}	a,a-Dimethylbenzen-methanole	23.39	21.85	35.27
16	4',2-Methylpropyl-acetophenone	22.88	25.00	20.13
17	(1-Methylethyl)-phenol	22.85	23.82	23.06
18	2(3H)-Benzothiazolone	20.64	12.41	19.95
19	2-Mercaptobenzothiazole	19.41	21.45	19.36
20^{e}	1,3-Bis(1-methylethenyl)-benzene	19.01	25.38	28.95
21	1-Ethyl-4-methoxy-benzene	17.33	9.50	17.35
22	Cumaranone	15.32	10.67	9.44
23	4-Methylphenol	15.14	14.97	14.93
24	Indole	14.25	14.58	14.38
25i	1-[4-(10-Hydroxy-1-methylethyl)phenyl]-ethanone	13.79	16.41	20.65
26	4-Ethyl-2-methoxy phenol	13.21	14.81	13.35
27	Phenol	11.63	8.36	10.91
28	1-Methyl-1-phenyl-idrazyne	10.76	6.10	7.39
29	1-Ethenyl-4-methoxy benzene	10.34	13.20	11.74
30^{e}	m-Xylene	9.33	10.85	11.44
31	Di-2-phenyl-1,2-propandiole	9.32	21.02	9.75
32	3,5,5-Trimethyl hexanoic acid	8.91	14.36	8.70
33	Benzothiazole	8.54	12.88	12.66
34	Hexanoic acid	8.10	-0.20	8.43
35i	1-Methoxyethylbenzene	8.07	7.61	8.47
36	Toluene	7.68	9.90	5.67

37	o-Xylene	7.67	13.85	10.67
38	p-Xylene	7.67	12.62	6.77
39	1,3-Dihydro-2H-indolone	6.38	9.72	7.62
40[e]	4-Piperidinole	6.23	0.79	8.78
41	Tetracloroethylene	5.92	4.09	5.25
42	Acetophenone	5.74	8.49	3.66
43	Benzene propanoic acid	4.40	10.16	6.98
44	2-Hexanole	3.57	3.91	3.53
45[i]	Trichloroethylene	3.55	0.94	-0.34
46	Benzene acetic acid	2.88	15.79	2.92
47	Carbon tetrachloride	2.50	4.13	3.62
48	1,2-Dichloropropane	2.50	3.40	1.16
49	Chloroform	2.45	0.65	1.18
50[e]	Trichlorofluoromethane	2.36	4.29	-6.95
51	Tetramethylthyourea	2.12	10.93	2.42
52	Freon 113	2.12	11.11	1.93
53	4-Methylbenzen-solfonamyde	2.04	9.78	6.35
54	2,2-Dimethyl-1,3-propandiole	2.00	3.29	1.35
55[i]	Caprolactame	1.74	-1.06	4.19
56	1,3-Propandiole-2-ethyl-2-hydroxymethyl	1.72	11.49	2.04
57	Tetrahydro-1,1-dioxydethiofene	1.72	-10.48	3.46
58	1,1,1-Trichloro ethane	1.64	1.03	2.13
59	1-(2-Methoxy propoxy)-propanole	1.64	3.39	1.29
60	2-Imidazolidinthyone	1.32	1.18	0.79

[i] Internal test set.

[e] External test set.

2.2. Descriptor calculation and screening

Molecular descriptors are used to encode molecular structural features with QSPR aims. In order to calculate descriptors, the chemical structures of molecules were drawn by Hyperchem package (Version 7) and optimized by the AM1 semiempirical method (Hyperchem 2002). After geometry optimization, Hyperchem output files were used by Dragon program as input to calculate molecular descriptors (Todeschini et al. 2003). Then descriptors that have high correlation with each other (R>0.9), and descriptors with same or near the same values were eliminated from the pool of descriptors. Variable selection is one of the most important steps in QSPR model development, which is especially important when one is required to deal with a large or even over whelming variable set. In order to determine the optimum number of descriptors from the remaining 429 descriptors the stepwise multilinear regression was used.

In order to determine the optimum number of descriptors in the model the value of R2 was calculated and plotted versus the number of descriptors in the model (Figure 1, break- point procedure). As can be seen in this figure there is not any significant improvement in R2 by adding more than six descriptors to the model. Therefore these descriptors were selected to developing MLR and ANN models. The selected descriptors are; Radial Distribution Function - 045 / weighted (RDF045v), Moriguchioctanol-water partition coefficient (logP) (MLOGP), hydrophilic factor (Hy), 3D-MoRSE - signal 13 / unweighted (Mor13u), leverage-weighted autocorrelation of lag 4 / unweighted (HATS4u) and leverage-weighted autocorrelation of lag 5 / weighted by mass (HATS5m). Detailed description of these descriptors can be found in the hand book of molecular descriptors by todeschini (Todeschini and Consonni 2000). Table 2 indicate the correlations matrix between these descriptors. As can be seen in this table there is not any high correlations between selected molecular descriptors.

Figure 1

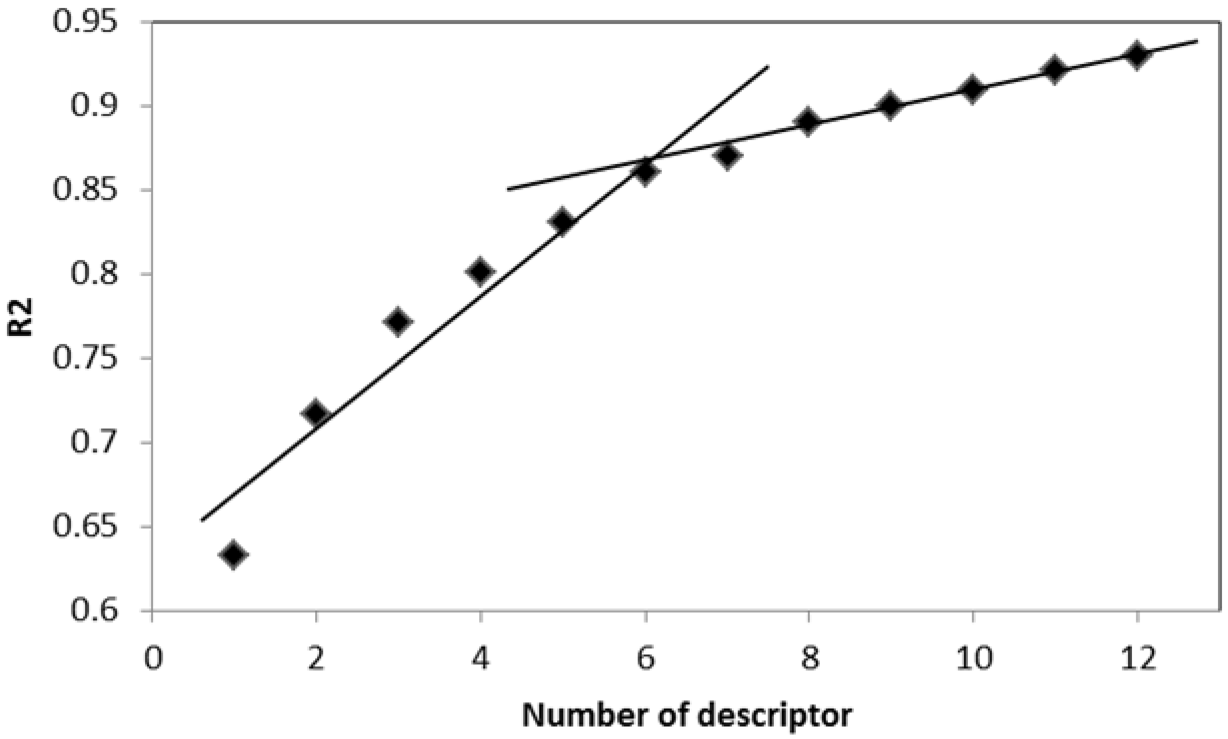

Figure 1. The plot of R2 against number of descriptor.

Table 2

The correlations matrix among selected descriptors

Descriptors	Mor13u	Hy	HATS4u	HATS5m	MLOGP	RDF045v
Mor13u	1	-0.197	0.311	0.071	0.183	0.230
Hy		1	0.230	-0.077	-0.451	-0.053
HATS4u			1	0.138	-0.367	0.046
HATS5m				1	0.272	0.509
MLOGP					1	0.492
RDF045v						1

2.3. Diversity analysis

In order to evaluate the prediction power of developed QSPR models (external validation test), data set must be divided to training and test sets. The common selection procedure, which is used for data set splitting is random selection. In this method the available data will be splitted without any bias for structure and there is a great probability of selecting chemicals outside the model structural application domain (AD) in the prediction set. Thus, the predictions for these chemicals could be unreliable, simply as they are extrapolated by the model. The other method is y- ranking procedure. In this method the data set is sorted in an ascending or descending manner according to their ECRIS value. Then test sets compounds were selected from this list by desirable distances from each other and remaining was considered as training set. This method was used to splitting of data set in the present work.

The obtained training set consist of 49 molecules and was used for model generation, while the internal test set had 6 compounds and was used for preventing over training of ANN model and the external test set had 5 members and was used to evaluate the predictability of the ANN model. In the case of MLR model internal and external test sets were considered as test set. Even by this algorithm there is no guarantee that the training and test sets be scattered over the whole area occupied by representative points in thedescriptor space (representativity), and that the training set be distributed over an area occupied by representative points for the whole dataset. To examine the diversity of data set, the mean distances of one sample to the remaining ones (i) were computed from descriptor space matrix as follows:

$$\bar{d}_i = \frac{\sum_{j=1}^{n} d_{ij}}{n-1} \quad i = 1,2,\ldots,n \qquad (1)$$

Where is a distance score for two different compounds,which can be measured by the Euclideandistance norm based on thecompound'sdescriptors (and):

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \qquad (2)$$

Then the mean distances were normalized within the interval of zero to one and the resulting values were plotted against ECRIS values. Figure 2 indicates the results of diversity analysis on the data set. As can be seen from this figure, the structures of the compounds are diverse in all sets and the training set with a broad representation of the chemistry space was adequate to ensure the model's stability and the diversity of test sets can prove the predictive capability of the model.

Figure 2

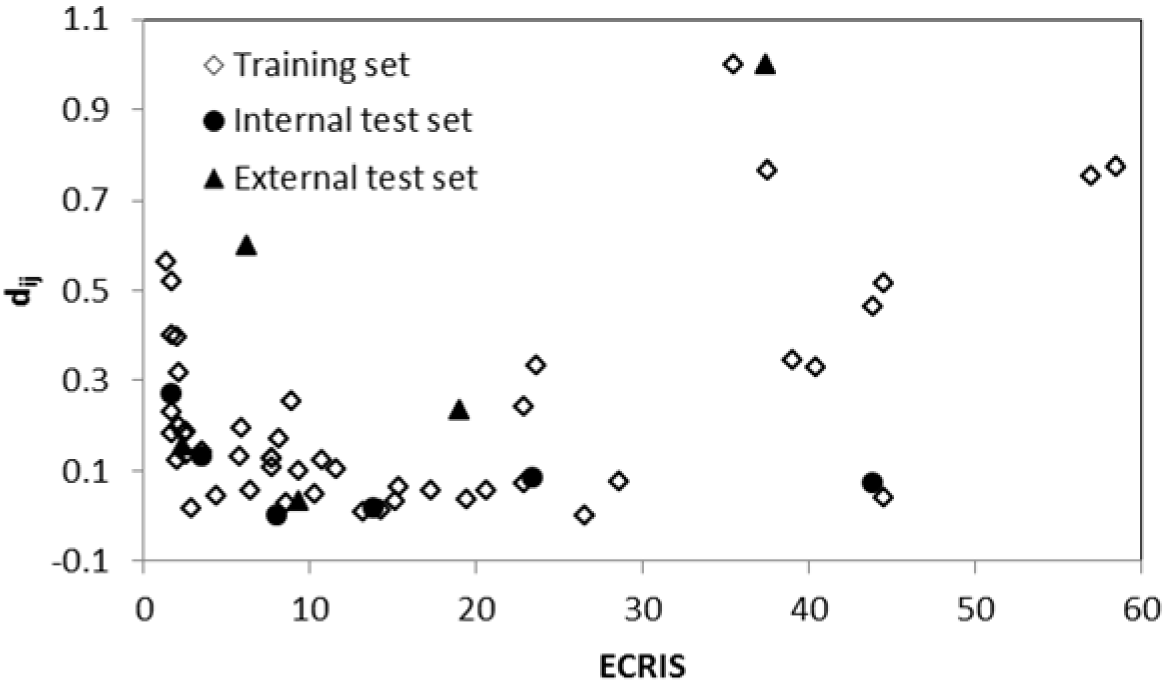

Figure 2. The results of diversity test.

3. RESULTS AND DISCUSSION

3.1. Linear modeling

Six selected descriptors were considered as independent variables and ECRIS value was considered as dependent variable for developing linear model. The specification of obtained MLR model is shown in eq. (3):

PECRIS = -9.797 (± 2.948) + 4.619 * RDF045v (± 0.676) + 7.351 * MLOGP (± 1.117) - 10.066 * Mor13u (± 2.455) + 6.047 * HATS4u (±2.645) + 56.753 * HATS5m (± 20.968) + 4.299 * Hy (± 1.670)   (3)

The calculated ECRIS values of molecules in data set by this model are shown in Table 1. The statistical parameters of this model are indicated in Table 3.

3.2. Nonlinear modeling

In order to check any nonlinear relationships between selected molecular structural descriptors and ECRIS values, artificial neural network (Hagan et al. 1996) was applied by using STATISTICA (ver.7) software (STATISTICA 2004). Generally, each network is built from several layers: one input layer, one or more hidden layers, and one output layer. The node in each layer is connected to the nodes of the next layer by weights. The number of neurons in input and output layers is equal to the number of independent variables and dependent variables, respectively. The number of neurons in hidden layer would should be optimized. A three-layer network with a sigmoid transfer function was designed, for which selected 6 selected descriptors were used as its inputs and ECRIS values as outputs. After optimization of topology and training of network, it was used for prediction of ECRIS values of data set. The predicted values of ECRIS for training, internal and external test sets were shown in Table 1. The statistical parameters of this model are shown in table 3.

Comparison between these values and those obtained by MLR model, indicates the superiority of ANN model over MLR ones. Figure 3 indicates the plot of ANN calculated versus experimental values of ECRIS. The correlation coefficient between calculated and experimented values of ECRIS is 0.99, 0.90 and 0.98 for training, internal and external test sets, respectively. Also the residuals of the ANN calculated ECRIS versus their experimental values are shown in Figure 4. Random propagation of residuals over zero line indicates that there is not any systematic error in developed ANN model.

Figure 3

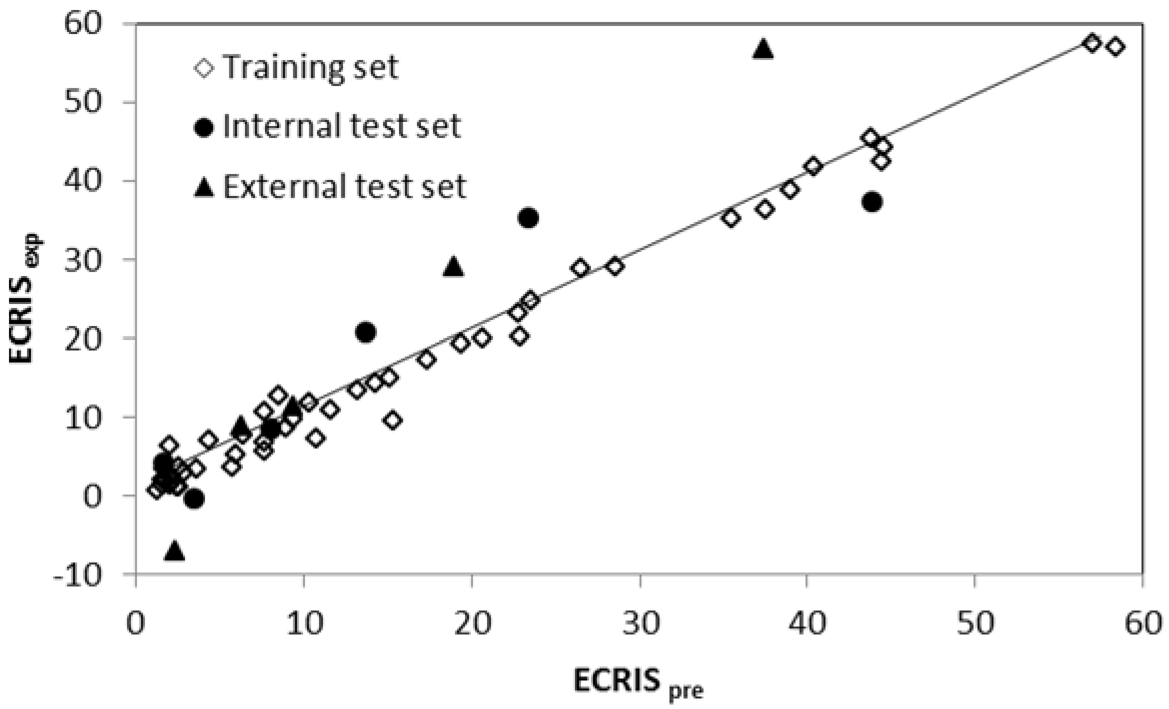

Figure 3. The plot of the ANN calculated ECRIS against the experimental values.

Figure 4

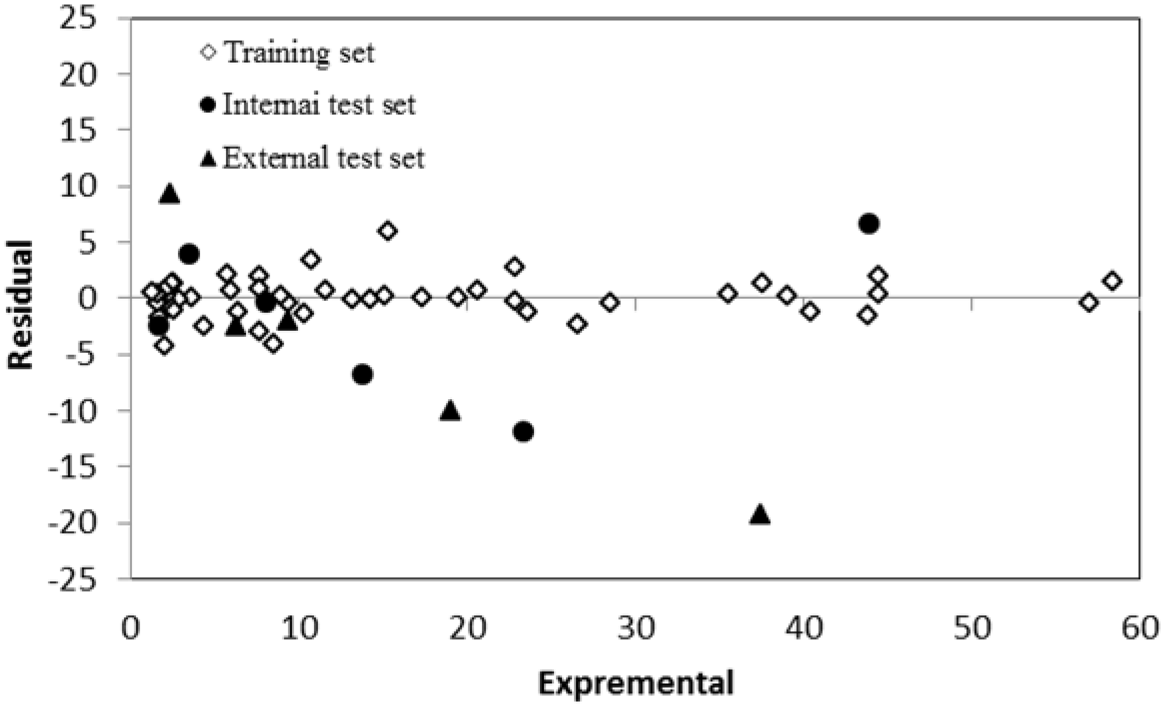

Figure 4. Plot of the ANN residuals against experimental values of ECRIS.

3.3. Model Validation

Validation is a crucial aspect of quantitative structure–activity relationship modeling. Cross validation provides a reasonable approximation of ability with which the QSPR predicts the activity values of new compounds. Leave one out cross validation (LOO) and leave many out cross validation (LMO) tests are two methods, which frequently used to validate QSPR models (Roy 2007). In the case of leave-one-out cross-validation, each member of the sample in turn is removed, the full modeling method is applied to the remaining n-1 members, and the fitted model is applied to the holdback member. Cross-validated squared correlation coefficient Q2 is calculated according to the following formula:

$$Q^2 = 1 - \frac{\sum (Y_i - Y_i)^2}{\sum (Y_i - Y)^2} \qquad (4)$$

where and   indicate predicted and observed activity values, respectively and   indicate mean activity value. A model is considered acceptable when the value of Q2 exceeds 0.5. Also standardized predicted error sum of squares (SPRESS), are calculated according to the following equation:

$$SPRESS = \sqrt{\frac{(Y_i - Y_i)^2}{n - k - 1}} \qquad (5)$$

In the above expression, n is the number of observations, and k is the number of descriptors in the model. The calculated values of Q2CV and SPRESS for LOO test on the ANN model are; 0.93 and 4.27, while these values are 0.84 and 6.5, respectively for the MLR model. Comparison between these values and also statistics in Table 3, indicates the superiority of ANN over MLR model. Also the Y- scrambling procedure was performed to ensure that there is not any chance correlation within the data matrix (Rücker et al. 2007). The mean value of R2 after 30 Y-scrambling runs was 0.286, which does not indicate the probability of a chance correlation.

Table 3

Statistical results of MLR and ANN models

Model	Training set				Internal test set				External test set			
	R	SE	F	RMSE	R	SE	F	RMSE	R	SE	F	RMSE
MLR	0.92	6.36	50.31	5.94	-	-	-	-	0.82	8.63	18.49	6.82
ANN	0.99	1.77	3755.17	1.74	0.90	7.41	18.91	6.5	0.98	2.39	135.14	10.65

3.4. Applicability Domain

Before a QSPR model is put in to use for screening chemicals, its domain of application must be defined (Xia et al. 2009). A simple measure of a chemical being too far from the applicability domain of the model is its leverage (Gramatica 2007), which is defined as (Netzeva et al. 2005):

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (i = 1, ..., n) \qquad (6)$$

Where is the descriptor row-vector of the query compound and is the matrix of model descriptor values for training set compounds. The super script refers to the transpose of the matrix/vector. The warning leverage is, generally, fixed at . To visualize the applicability domain of non linear model, the standardized residuals versus leverage (Hat diagonal) values were plotted (William plot) for an immediate and simple graphical detection of both the response outliers (i.e., compounds with standardized residuals greater than three standard deviation units>) and structurally influential chemicals in the model (>). Figure 5 shows the results for this analysis of the nonlinear QSPR model. As can be seen from this figure, there is no response outlier compound both for training and test sets, which indicated further the reliability of the predictions from another aspect.

Figure 5

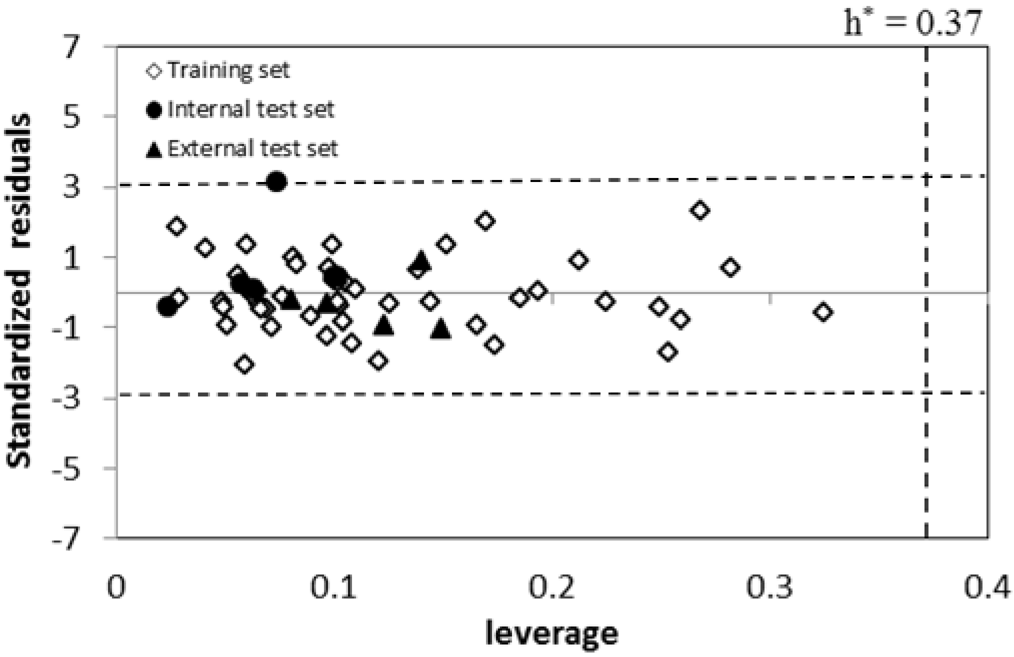

Figure 5. Applicability domain of non linear model; ( ).

3.5. Descriptors interpretation

In order to determine the relative importance of each variable in the ANN model, the sensitivity analysis was applied. This method is performed based on the sequential removal of variables by zeroing the specific connections weight for that specific input variable in the first layer of the ANN. For each sequentially zeroed input variable, root-mean-square error of prediction (RMSEP) as the prediction error of network was calculated. Generally RMSEP value increases in this way. Then, differences between RMSEP and root-mean-square error of established ANN was calculated and shown as DRMSE. Each variable, which causes greater value of DRMSE, is more important. This procedure was applied on the developed ANN model. The calculated values of DRMSE are plotted in Figure6. As can be seen in this figure the most important descriptor was RDF045v.This descriptor is the radial distribution function - 045 / weighted by van der Waals volume and is the topology type descriptors.

The RDF descriptors are based on the distance distribution in a three-dimensional representation of the molecule. Besides information about inter-atomic distances they also give information about ring types, planar and non-planar systems and atom-types (Hemmer et al. 1999). The second descriptor was MLOGP. This descriptor is the Moriguchi octanol-water partition coefficient which indicates the liphophilicity of molecule (Moriguchi et al. 1992). The 3D-molecular representation of structure based on electron diffraction (3D-MORSE)-type descriptors that represent the 3D structure of a molecule is another descriptor in the model (Mor13u) (Soltzberg and Wilkins 1997). These types of descriptors are based on the idea of obtaining information from the 3D atomic coordinates by transforming that used in electron diffraction studies for preparing theoretical scattering curves (Schuur et al. 1996; Soltzberg and Wilkins 1997). The others descriptors are; HATS4u which is leverage-weighted autocorrelation of lag 4 / unweighted and HATS5m which is leverage-weighted autocorrelation of lag 5 / weighted by mass. These type of descriptors are computed on the basis of Hydrogen-filled molecule.

They are belonged to geometry, topology, and atom-weighted assembly (GETAWAY) descriptors (Consonni et al. 2002). These types of descriptors encode geometrical information given from influence matrix, topological information given by molecular graph, and chemical information from selected atomic properties. Another descriptor is hydrophilic factor, Hy. This descriptor is an empirical descriptor that related to hydrophilicity of compounds (Todeschini et al. 1997), and defined as follows:

$$H_y = (1+N_{Hy})\log_2(1+N_{Hy}) + N_c\left(1/A\log_2 1/A + \sqrt{N_{Hy}/A^2}\right)/\log_2(1+A) \qquad (7)$$

Where, is the number of hydrophilic groups (-OH, -SH, -NH),   is the number of carbon atoms, and   the number of atoms (hydrogen excluded). The appearances of topological and electronic type descriptor in developed QSPR model indicates the role of steric and electronic interactions in ECRIS values of chemicals.

Figure 6

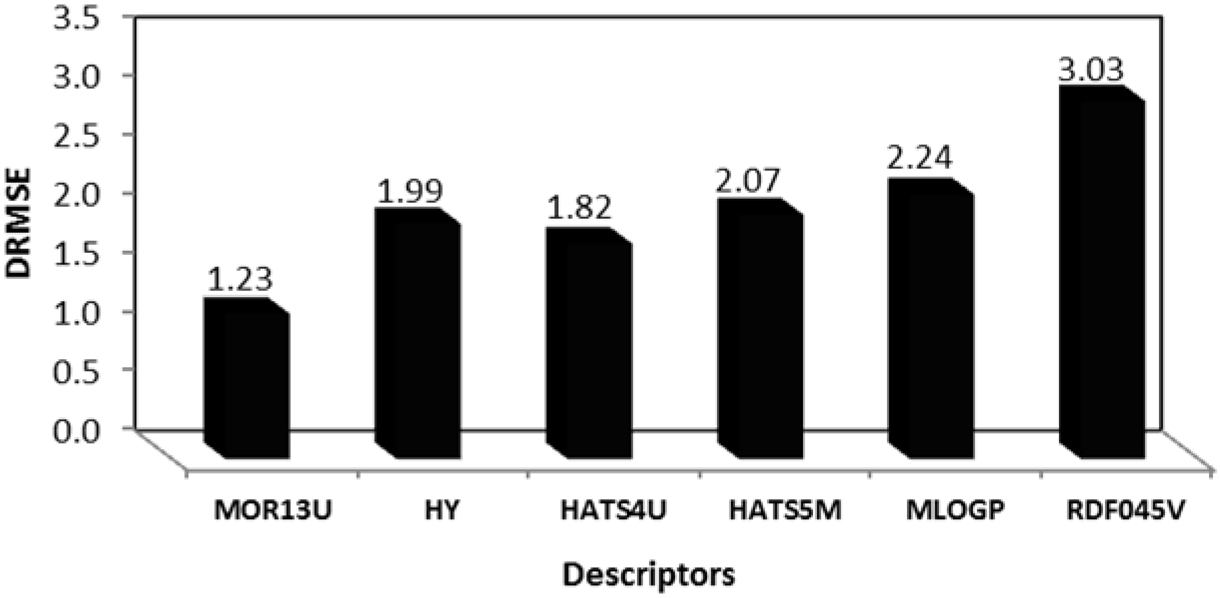

Figure 6. The results of sensitivity analysis on the ANN model.

4. CONCLUSION

In this study, MLR and ANN were used to build linear and nonlinear QSPR models to predict the soil contaminant index of some organic compounds. The statistical results of the developed models indicated the superiority of the nonlinear model over linear ones. These results revealed that there are some nonlinear relations between the soil contaminant index of some organic compounds and their structural molecular descriptors. Moreover, it was concluded that it was possible to predict the soil contaminant index of some organic compounds from their theoretical calculated molecular descriptors.

REFERENCES

1. Arnot JA, Meylan W, Tunkel J, Howard PH, Mackay D, Bonnell M, Boethling RS (2009) A quantitative structure–activity relationship for predicting metabolic biotransformation rates for organic chemicals in fish. Environ Toxicol Chem 28(6):1168–1177

2. Bunke D, Oldenburg C (2005) Indicators for chemicals: sources, impacts and policy performance. Environ Sci Pollut Res 12(5):310–314

3. Calliera M, Finizio A, Azimonti G, Benfenati E, Trevisan M (2006) Harmonized pesticide risk trend indicator for food (HAPERITIF): the methodological approach. Pest Manag Sci 62(12):1168–1176

4. Consonni V, Todeschini R, Pavan M (2002) Structure/responsecorrelation and similarity/diversity analysis by GETAWAYdescriptors. 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci 42(3):682–692

5. Finizio A, Calliera M, Vighi M (2001) Rating systems for pesticide risk classification on different ecosystems. Ecotoxicol Environ Safe 49(3):262–274

6. García-Domenech R, Alarcón-Elbal P, Bolas G, Bueno-Marí R, Chordá-Olmos FA, Delacour SA, Mouriño MC, Vidal A, Gálvez J (2007) Prediction of acute toxicity of organophosphorus pesticides using topological indices. SAR QSAR Environ Res 18(7-8):745-755

7. Giesen D, Jonker MTO, van Gestel CAM (2012) Development of QSARs for the toxicity of chlorobenzenes to the soil dwelling springtail Folsomia candida. Environ Toxicol Chem 31(5):1136–1142

8. Goudarzi N, Goudarzi M, Araujo MCU, Galvão RKH (2009) QSPR Modeling of Soil Sorption Coefficients (KOC) of Pesticides Using SPA-ANN and SPA-MLR. J Agric Food Chem 57(15):7153–7158

9. Gramatica P (2007) Principles of QSAR models validation:internal and external. QSAR Comb Sci 26(5):694–701

10. Hagan MT, Demuth HB, Beale M (1996) Neural Network Design, PWS, Boston

11. HAIR (2006) Harmonised Environmental Indicators for Pesticide Risks. Contract FP6-2002-501997-SSP-1.<http://www.rivm.nl/rvs/overige/risbeoor/Modellen/HAIR.jsp>

12. Hemmer MC, Steinhauer V, Gasteiger J (1999) Deriving the 3D structure of organicmolecules from their infrared spectra. Vib Spectrosc 19(1):151-164

13. Huijbregts MAJ, Thissen U, Guine JB, Jager T, Kalfe D, van de Meent D, Ragas AMJ, Wegener Sleeswijk A, Reijnders L (2000) Priority assessment of toxic substances in life cycle assessment. Part I: calculation of toxicity potentials for 181 substances with the nested multi-media fate, exposure and effects model USES-LCA. Chemosphere 41(4):541–573

14. Hyperchem (2002), Release 7.0 for Windowse, Hypercube Inc

15. Jiao L (2012) QSPR Study on the Soil-Water Partition Coefficient of Polychlorinated Biphenyls by Using Artificial Neural Network. Adv Mater Res 455– 456:930-934

16. Kim JH, Gramatica P, Kim MG, Kim D, Tratnyek PG (2007) QSAR modeling of water quality indices of alkylphenol pollutants. SAR QSAR Environ Res 18(7-8):729-743

17. Moriguchi I, Hirono S, Liu Q, Nakagome I, Matsushita Y (1992) Simple method of calculating Octanol/Water Partition Coefficient. Chem Pharm Bull 40(1):127-130

18. Moudgal CJ, Young D, Nichols T, Martin T, Harten P, Venkatapathy R, Stelma G, Siddhanti S, Baier-Anderson C, Wolfe M (2008) Application of QSARs and VFARs to the rapid risk assessment process at US EPA. SAR QSAR Environ Res 19(5-6):579-587.

19. Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V (2010) Advances in computational methods to predict the biological activity of compounds. Expert Opin Drug Discov 5(7):633–654

20. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V (2009) A practical overview of quantitativestructure-activityrelationship.EXCLIJ8:74–88

21. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJM, Tong W, Veith G, Yang C (2005) Current status ofmethods for defining the applicability domain of (Quantitative) Structure–Activity Relationships: The report and recommendations of ECVAM Workshop 52. Altern Lab Anim 33(2):155–173

22. Peterson PJ, Granados A (2002) Towards sets of hazardous waste indicators: essential tools for modern industrial management. Environ Sci Pollut Res 9(3):204–214

23. Peterson PJ, Williams WP (1999) New indicator approaches for effective urban air quality management. Environ Sci Pollut Res 6(4):225–232

24. Peyret T, Krishnan K (2011) QSARs for PBPK modelling of environmental contaminants. SAR QSAR Environ Res 22(1-2):129-169

25. Qin LT, Liu SS, Liu HL, Ge HL (2008) A new predictive model for the bioconcentration factors of polychlorinated biphenyls (PCBs) based on the molecular electronegativity distance vector (MEDV). Chemosphere. 70(9):1577–1587

26. Roy K (2007) On some aspects of validation of predictive quantitative structure–activity relationship models. Expert Opin Drug Discovery 2(12):1567–1577

27. Rücker C, Rücker G, Meringer M (2007) Y–randomization and its variants in QSPR/QSAR. J Chem Inf Model 47(6):2345–2357

28. Ruiz P, Faroon O, Moudgal CJ, Hansen H, De Rosa CT, Mumtaz M (2008) Prediction of the health effects of polychlorinated biphenyls (PCBs) and their metabolites using quantitative structure–activity relationship (QSAR). Toxicol Lett 181(1):53–65

29. Sabljic A, Guesten H, Hermens J, Opperhuizen A (1993) Modeling octanol/water partition coefficients by molecular topology: chlorinated benzenes and biphenyls. Environ Sci Technol 27(7):1394–1402

30. Schuur JH, Selzer P, Gasteiger J (1996) The Coding of the Three-dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure - Spectra Correlations and Studies of Biological Activity. J Chem Inf Comput Sci 36(2):334-344

31. Senese V, Boriani E, Baderna D, Mariani A, Lodi M, Finizio A, Testa S, Benfenati E (2010) Assessing the environmental risks associated with contaminated sites: definition of an ecotoxicological classification index for landfill areas (ECRIS). Chemosphere 80(1):60–66

32. Snyder EM, Snyder SA, Giesy JP, Blonde SA, Hurlburt GK, Summer CL, Mitchell RR, Bush DM (2000) SCRAM: a scoring and ranking system for persistent, bioaccumulative, and toxic substances for the North American great lakes (Part I, II, III, IV). Environ. Sci Pollut Res 7(3):176-184

33. Soltzberg LJ, Wilkins CL (1997) Molecular Transforms: A Potential Tool for Structure-Activity Studies. J Am Chem Soc 99(2):439-443

34. STATISTICA (data analysis software system), StatSoft. Inc., Tulsa (2004) software available at http://www.statsoft.com

35. Todeschini R, Consonni V (2000) Handbook of Molecular Descriptors, Wiley–VCH, Weinheim

36. Todeschini R, Vighi M, Finizio A, Gramatica P (1997) 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico- chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. SAR QSAR Environ Res 7(1-4):173-193

37. Todeschini R, Consonni V, Mauri A, Pavan M (2003) DRAGON software version 3.0, Milano Chemometrics and QSAR Research Group, TALETE srl, Milan, Italy, software available at http://www.disat.unimib.it/chm

38. Tugcu G, Saçan MT, Vracko M, Novic M, Minovski N (2012) QSTR modelling of the acute toxicity of pharmaceuticals to fish. SAR QSAR Environ Res 23(3-4):297-310

39. United Nations (2007) Indicators of Sustainable Development: Guidelines and Methodologies, third ed. Available at: <http://www.un.org/esa/sustdev/ natlinfo/indicators/guidelines.pdf>

40. Van Gestel AM, Ma WC (1993) Development of QSAR's in soil ecotoxicology: Earthworm toxicity and soil sorption of chlorophenols, chlorobenzenes and chloroanilines. Water Air Soil Pollut 69(3-4):265-276

41. Xia B, Liu K, Gong Z, Zheng B, Zhang X, Fan B (2009) Rapid toxicity prediction of organic chemicals to Chlorella vulgaris using quantitative structure–activity relationships methods. Ecotox Environ Safe 72(3):787–794

42. Zhou W, Zhai Z, Wang Z, Wang L (2005) Estimation of n–octanol/water partition coefficients (Kow) of all PCB congeners by density functional theory. J Mol Struct (Theochem) 755(1-3):137–145