JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY
ARTICLE

DOI:10.1021/jf9008839

QSPR Modeling of Soil Sorption Coefficients (K_{OC}) of Pesticides Using SPA-ANN and SPA-MLR

NASSER GOUDARZI,*[†] MOHAMMAD GOODARZI,[§] MARIO CESAR UGULINO ARAUJO,[#] AND ROBERTO KAWAKAMI HARROP GALVÃO[⊥]

[†]Faculty of Chemistry, Shahrood University of Technology, P.O. Box 316, Shahrood, Iran, [§]Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), UNLP, CCT La Plata-CONICET, Diag. 113 y 64, C.C. 16, Suc. 4, 1900, La Plata, Argentina, [#]Departamento de Química, Universidade Federal da Paraíba, CCEN, Caixa Postal 5093, CEP 58051-970 Joao Pessoa, PB, Brazil, and [⊥]Divisão de Engenharia Eletrônica, Instituto Tecnológico de Aeronáutica, CEP 12228-900 São José dos Campos, SP, Brazil

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 (multiple linear regression; MLR) and nonlinear (artificial neural network; ANN) methods. The QSPR models were validated by cross-validation as well as application of the models to predict the K_{OC} 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.

KEYWORDS: Quantitative structure–activity relationship; soil sorption coefficients; successive projection algorithm; artificial neural network

INTRODUCTION

The widespread use and distribution of pesticides, chemicals released deliberately and in large quantities into the environment, are of great relevance today. The potential for adverse consequences of pesticides, such as environmental impact on the quality of water and wildlife habitats, has led to the development of detailed analyses of potential environmental hazards associated with the use of these compounds (*1–3*). Knowledge of pesticide environmental behavior, such as adsorption to soil, leaching to groundwater, and volatility in the atmosphere, is of primary concern for an accurate assessment of the risk to the environment and humans. Pesticide distribution and fate in various environmental media and compartments are strongly influenced by the inherent properties of the compounds themselves, particularly by basic physicochemical properties such as solubility in water, vapor pressure (VP), and partitioning coefficients between organic matter (in soil or sediment) and water. In the assessment of pesticide environmental behavior it is important to understand the properties that control pesticide partitioning tendencies. Recently, some molecular modeling methods based on widespread quantitative structure–property/activity relationships (QSPR/QSAR) techniques have found their place as important tools for chemists (*4*, *5*).

The reliable estimation of soil sorption coefficients (K_{OC}) for organic pesticides plays a fundamental role in agriculture, especially for describing the pollution impact of the pesticides and their tendency for biodegradation. This partition coefficient represents a measure of the retaining of a chemical by the organic matter of soils and sediments through a wide variety of possible intermolecular interactions (*6*, *7*). Nowadays, fewer than 300 chemicals have measured K_{OC} values, and little information is available on the sorption behavior of their metabolites (*8*). Clearly, the prediction of the K_{OC} parameter for a wide number of chemical structures is very convenient for application in risk assessment.

A generally accepted remedy to surmount the lack of availability of experimental data in contemporary chemistry is the application of quantitative structure–property relationships (QSPR) analysis (*9*), in the present case to obtain adequate predictions for soil sorption coefficients. The ultimate role of the different formulations of the QSPR theory is to suggest mathematical models for estimating relevant end points of interest, especially when these cannot be experimentally determined for some reason. These studies simply rely on the assumption that the structure of a compound determines the physicochemical properties it manifests. The molecular structure is therefore translated into the so-called molecular descriptors through mathematical formulas obtained from several theories, such as chemical graph theory, information theory, and quantum mechanics (*10*, *11*). There exist more than a thousand theoretical descriptors available in the literature, and one usually faces the

*Author to whom correspondence should be addressed (telephone +982733335441; fax +982733335441; e-mail goudarzi10@yahoo.com or goudarzi@shahroodut.ac.ir).

Published on Web 07/13/2009

problem of selecting those that are the most representative for the property under consideration. In this investigation, we introduce the successive projections algorithm (SPA) (*12*) as a feature selection, due to its ability in solving the descriptor selection problems in QSPR model development.

SPA is a technique specifically designed to select subsets of variables with small collinearity and to improve the conditioning of multiple linear regression (MLR) models. This algorithm was originally proposed for wavelength selection in spectroscopic data sets, especially under conditions of strong spectral overlapping (*12*). MLR models obtained by using SPA have been shown to be superior, in terms of prediction ability, to full-spectrum partial least squares (PLS) models in a variety of applications, including UV−vis (*12−15*), ICP-OES (*16*), FT-IR (*17*), and NIR spectrometry (*17−19*). SPA has also been successfully employed in various classification studies (*20*, *21*).

SPA comprises three phases (*22*). Initially, the algorithm builds candidate subsets of variables on the basis of a collinearity minimization criterion. Such subsets are constructed according to a sequence of vector projection operations applied to the columns of the matrix of available predictor data. In the second phase, the best candidate subset is chosen according to a criterion that evaluates the prediction ability of the resulting MLR model, such as the root-mean-square error obtained in a validation set (*23*). In the third phase, the selected subset is subjected to an elimination procedure to determine whether any variables can be removed without significant loss of prediction ability. Each of these phases is explained in detail elsewhere (*24*).

Although SPA was initially designed for use with MLR models, it may be worth investigating whether it could be employed with different modeling techniques. In the present paper, the variables selected by SPA will be used to build MLR and artificial neural network (ANN) models.

MATERIALS AND METHODS

Data. Experimental soil sorption coefficient (K_{OC}) data of the 124 pesticides were used in this work (*25*). A QSPR model for the estimation of the soil sorption coefficient of pesticides is established in the following six steps: the molecular structure input and generation of the files containing the chemical structures is stored in a computer-readable format; quantum mechanics geometry is optimized with a semiempirical (AM1) method; structural descriptors are computed; structural descriptors are selected; and the structure− K_{OC} model is generated by MLR, ANN, and statistical analysis. The names of these compounds and their experimental and calculated values by ANN and MLR methods are shown in **Table 1**. As can be seen, this set contains 124 soil sorption coefficients (K_{OC}) data of pesticides. The data set was split into training, validation, and test sets. The training set of 62 compounds, with log K_{OC} values ranging from 0.301 to 6.000, was used to construct the model. The validation set of 31 compounds, with log K_{OC} values ranging from 0.531 to 5.00, was used to prevent overtraining/overfitting of the ANN model. The test set of 31 compounds, with K_{OC} values ranging from 0.301 to 4.204, was used as an external set to evaluate the predictive ability of the model.

Descriptor Generation and Screening. For calculating K_{OC}, we have used the well-known EPI K_{OC} estimation tool (*26*). The soil sorption coefficients (K_{OC}) of solutes are related to some of their structural, electronic, and geometric properties. The value of these properties can be encoded quantitatively by numerical values named molecular descriptors. These molecular parameters are to be used to search for the best QSPR model of the soil sorption coefficients. The 2D structures of the molecules were drawn using Hyperchem 7 software (*27*). The final geometries were obtained with the semiempirical AM1 method in the Hyperchem program. The molecular structures were optimized using the Polak−Ribiere algorithm until the root-mean-square gradient was 0.001 kcal mol^{-1}. The resulting geometry was transferred into the Dragon program package, which was developed by a Milano chemometrics and QSPR group (*28*), to calculate 1457 descriptors in constitutional, topological, geometrical, charge, GETAWAY (Geometry, Topology and Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), molecular walk count, BCUT, 2D autocorrelation, aromaticity index, randic molecular profile, radial distribution function, functional group, and atom-centered fragment classes.

It is worth mentioning that in the first preselected analysis we removed 647 descriptors because many of them included zero or other constant/near-constant values and did not have enough information of structure. On the other hand, to decrease the redundancy existing in the descriptor data matrix, the correlation coefficient *r* of the descriptors with each other was examined and the collinear descriptors (with $r > 0.9$) were removed.

Upon application of SPA, seven descriptors were selected for model building. These descriptors were as follows: H total index/weighted by atomic polarizabilities (HTp); Moran autocorrelation lag-6/weighted by atomic Sanderson electronegativities (MATS6e); third-component symmetry directional WHIM index/weighted by atomic van der Waals volumes (G3v); 3D-MoRSE-signal05/weighted by atomic masses (Mor05m); first-component symmetry directional WHIM index/weighted by atomic masses (G1m); Moran autocorrelation lag-4/weighted by atomic polarizabilities (MATS4p); and highest eigenvalue n.2 Burden matrix/weighted by atomic masses (BEHm2).

The first selected descriptor is H total index/weighted by atomic polarizabilities (HTp), which is a GETAWAY descriptor. GETAWAY tries to match the 3D molecular geometry provided by the molecular influence matrix and atom relatedness by topology with chemical information by using different atomic weighting schemes (unit weights, mass, polarizability, electronegativity).

The molecular influence matrix **H** is defined by

$$\mathbf{H} = \mathbf{M}(\mathbf{M}^T\mathbf{M})\mathbf{M}^T \tag{1}$$

where **M** is the molecular matrix. The resultant $A \times A$ matrix is invariant to rotation of the molecular coordinates. The diagonal elements h_v are termed leverages and represent the influence of each atom in determining the shape of the molecule. Each off-diagonal element h_v represents the degree of accessibility of the j^{th} atom to interactions with the i^{th} atom (*29*).

The second and third descriptors that were selected for the QSPR model were Moran autocorrelation lag-6/weighted by atomic Sanderson electronegativities (MATS6e) and Moran autocorrelation lag-4/weighted by atomic polarizabilities (MATS4p). These descriptors are 2D autocorrelation descriptors. The structural variables introduced by Moran correspond to bidimensional autocorrelations between pairs of atoms in the molecule and are also defined to quantify the contribution of a considered atomic property to the analyzed property. These can be readily calculated, that is, by summing products of terms including the atomic weights for the terminal atoms in all of the paths of a prescribed length. For the case of MATS6e, the path connecting a pair of atoms has a length of 6 and involves the atomic Sanderson electronegativities as weighting scheme.

The next selected descriptors were third-component symmetry directional WHIM index/weighted by atomic van der Waals volumes (G3v) and first-component symmetry directional WHIM index/weighted by atomic masses (G1m), which are a kind of WHIM descriptor. WHIM descriptors are molecular descriptors based on statistical indices calculated on the projection of the atoms along principal axes. Also, WHIM descriptors are built in such a way as to capture relevant molecular 3D information regarding molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. The algorithm consists of performing a principal component analysis on the centered Cartesian coordinates of molecules by using a weighted covariance matrix obtained from different weighting schemes for atoms

$$S_{jk} = \frac{\sum_{i=1}^{A} w_i(q_{ij} - \overline{\bar{q_j}})(q_{jk} - \overline{\bar{q_k}})}{\sum_{i=1}^{A} w_i} \tag{2}$$

where S_{jk} is the weighted covariance between the j^{th} and k^{th} atomic coordinates, A is the number of atoms, w_i the weight of the i^{th} atom, q_{ij}

Table 1. Data Set with Experimental and Calculated Soil Sorption Coefficients (Log K_{OC})[a]

no.	compound	log K_{OC} (exptl)	ANN	MLR	EPI	no.	compound	log K_{OC} (exptl)	ANN	MLR	EPI
1	acephate	0.301	0.819	0.324	1.338	63	aldicarb	0.903	1.376	1.668	3.371
2	aldrin	3.699	3.397	3.784	3.494	64	anilazine	2.344	2.464	2.177	2.558
3	atrazine	2.093	1.827	2.018	2.267	65	benazolin	1.344	1.683	1.278	1.137
4	bendiocarb	2.756	2.610	2.104	1.512	66	bifenox	3.383	3.150	2.889	2.922
5	bromacil	1.505	1.485	1.349	5.025	67	butylate	3.602	3.116	3.415	1.493
6	captan	2.104	2.656	2.841	2.648	68	carbetamide	1.946	1.959	1.767	1.274
7	carbofuran	1.342	2.113	2.100	1.386	69	chloroneb	3.217	3.237	3.283	2.788
8	chlorothalonil	3.932	3.771	3.401	3.161	70	chlorpyrifos	3.697	3.112	3.09	2.647
9	chlorsulfuron	1.602	1.879	2.390	2.362	71	cypermethrin	5.000	4.549	4.612	2.638
10	diazinon	2.434	2.724	2.873	2.375	72	1,3-dichloropropene	1.505	1.861	2.106	2.149
11	dichlorprop	3.000	2.454	2.457	1.843	73	dinoseb	1.591	1.428	1.564	3.387
12	dimethoate	1.301	1.754	2.286	1.000	74	disulfoton	2.778	2.638	2.690	2.399
13	diuron	2.657	2.768	3.243	1.207	75	EPTC	2.082	2.233	2.370	3.528
14	esfenvalerate	3.724	3.830	4.078	2.716	76	ethofumesate	2.170	2.486	2.311	2.544
15	ethoprophos	2.004	1.715	2.276	1.574	77	fenoxaprop-P	3.652	3.716	2.813	1.484
16	fenoxycarb	3.000	3.042	2.792	3.578	78	fluometuron	2.000	2.148	2.081	1.686
17	fluoroxypyr	1.531	1.611	1.755	2.017	79	glyphosate	2.223	2.220	2.455	3.187
18	hexazinone	1.732	1.606	1.829	3.991	80	iprodione	2.679	3.034	2.73	1.000
19	isofenphos	2.778	3.038	3.165	2.638	81	linuron	2.597	2.323	2.129	2.260
20	malathion	3.255	3.059	3.493	2.789	82	methidathion	2.212	2.289	2.55	1.079
21	methiocarb	2.751	2.710	2.358	2.935	83	metolachlor	2.243	2.818	3.046	4.628
22	metsulfuron	1.544	1.493	1.896	2.384	84	parathion	3.698	3.423	3.31	2.465
23	parathion-methyl	2.373	2.333	2.211	2.061	85	phorate	3.000	2.905	2.365	1.110
24	phosalone	4.255	3.923	4.122	2.209	86	pirimiphos-methyl	3.000	3.032	2.923	2.456
25	prochloraz	3.864	3.645	4.047	1.850	87	propachlor	1.832	2.244	2.135	4.060
26	propanil	2.173	2.214	2.870	1.903	88	propiconazole	2.957	3.170	3.474	3.250
27	propoxur	1.477	1.931	2.057	2.739	89	sulfometuron	1.892	1.711	2.177	2.718
28	terbacil	1.74	1.995	1.660	2.359	90	terbutryn	2.817	2.836	2.957	2.676
29	thifensulfuron	1.653	1.636	2.093	3.378	91	thiram	0.531	0.846	0.984	5.250
30	tralomethrin	5.000	5.050	4.982	2.133	92	trichlorfon	1.462	2.010	1.453	2.647
31	tridemorph	3.308	2.951	3.230	2.318	93	vinclozolin	2.426	2.574	2.641	1.767
32	alachlor	2.079	2.581	2.646	2.707	94	acifluorfen	2.053	2.292	2.536	1.257
33	amitrole	2.105	2.116	1.965	2.378	95	ametryn	2.477	2.227	1.85	1.519
34	azinphos-methyl	3.166	2.735	3.069	3.124	96	azinphos-ethyl	3.166	2.979	3.389	2.143
35	bentazone	1.531	1.375	1.389	2.262	97	benomyl	3.278	3.240	2.740	3.428
36	bromoxynil	2.230	2.336	2.000	5.033	98	bromofenoxim	0.334	0.436	0.732	2.196
37	carbendazim	2.110	1.911	1.740	3.127	99	carbaryl	2.093	2.083	1.908	2.837
38	chloridazon	2.037	1.869	1.824	1.459	100	carboxin	2.415	2.255	2.253	2.455
39	chlorpropham	2.602	2.608	2.373	2.428	101	chlorotoluron	2.243	2.542	2.274	2.294
40	cycloate	2.633	2.232	2.116	1.907	102	cyanazine	2.029	2.177	2.041	2.551
41	dichlobenil	2.326	2.481	1.931	1.686	103	dicamba	0.301	0.416	0.321	2.107
42	diflubenzuron	4.000	3.719	3.593	2.878	104	diclofop	4.204	3.708	3.918	3.745
43	diquat	6.000	5.757	5.586	3.025	105	diphenamid	2.079	2.111	2.44	1.645
44	endosulfan	4.093	4.138	4.008	3.549	106	DNOC	1.806	1.706	1.421	3.201
45	ethion	4.000	4.018	3.626	1.389	107	ethalfluralin	3.602	3.179	3.275	2.173
46	fenamiphos	2.426	2.587	2.445	4.187	108	etridiazole	3.000	3.132	3.021	1.984
47	fenthion	3.176	3.079	2.346	2.918	109	fenpropimorph	3.403	3.477	3.477	1.891
48	glufosinate	2.000	2.049	1.899	2.913	110	fonofos	2.939	2.402	2.716	2.991
49	ioxynil	2.301	1.987	1.846	2.133	111	imazapyr	2.000	1.977	1.714	2.517
50	lindane	3.041	2.912	3.031	2.779	112	isoproturon	2.029	2.086	2.07	2.803
51	methabenzthiazuron	2.722	2.252	2.064	2.895	113	mecoprop	2.103	2.281	2.06	1.000
52	methoxychlor	4.903	4.766	4.26	2.412	114	methomyl	1.857	2.084	1.862	3.428
53	napropamide	2.845	3.023	2.866	5.644	115	molinate	2.278	1.993	1.764	3.672
54	permethrin	5.000	4.442	4.238	3.909	116	pebulate	3.155	2.342	2.824	1.000
55	pirimicarb	2.869	2.537	2.819	4.117	117	picloram	1.204	1.607	1.265	1.419
56	prometryn	2.411	2.199	2.266	2.462	118	prometon	3.176	2.791	3.012	3.447
57	propham	1.272	1.214	1.535	1.564	119	propazine	2.187	2.453	2.609	3.215
58	simazine	2.061	2.408	2.190	2.212	120	propyzamide	2.403	2.513	2.745	1.731
59	terbuthylazine	2.485	2.683	2.920	2.352	121	terbufos	2.698	2.666	3.117	4.155
60	thiodicarb	2.544	2.941	2.892	3.439	122	thiobencarb	2.954	2.504	2.464	3.986
61	triallate	3.268	3.293	3.933	3.917	123	triadimenol	2.357	2.527	2.969	2.677
62	vernolate	2.415	2.277	2.330	4.429	124	trifluralin	3.807	3.546	2.821	2.461

[a] Training set, molecules 1−62; validation set, molecules 63−93; test set, molecules 94−124.

and q^{jk} represent the j^{th} and k^{th} coordinates (j, $k = x, y, z$) of the ith atom, respectively, and $\overline{q}$ is the corresponding average value. The last descriptor that appeared in the model is highest eigenvalue n.2 Burden matrix/weighted by atomic masses (BEHm2). BCUT descriptors are defined as eigenvalues of a modified connectivity matrix, which could be called Burden matrix B. The B matrix is defined as follows: The diagonal elements B_{ii} are the atomic

Table 2. Descriptors Used in Model Construction

no.	symbol	class	meaning
1	HTp	GETAWAY	H total index/weigthed by atomic polarizabilities
2	MATS6e	2D autocorrelation	Moran autocorrelation lag-6/ weighted by atomic Sanderson electronegativities
3	G3v	WHIM	third-component symmetry directional WHIM index/weighted by atomic van der Waals volumes
4	Mor05m	3D-MoRSE	3D-MoRSE-signal05/weighted by atomic masses
5	G1m	WHIM	first-component symmetry directional WHIM index/weighted by atomic masses
6	MATS4p	2D autocorrelation	Moran autocorrelation lag-4/weighted by atomic polarizabilities
7	BEHm2	BCUT	highest eigenvalue n.2 Burden matrix/weighted by atomic masses

Table 3. Correlation Matrix for the Seven Selected Descriptors

	HTp	MATS6e	G3v	Mor05m	G1m	MATS4p	BEHm2
HTp	1						
MATS6e	0.0337	1					
G3v	0.209	0.001	1				
Mor05m	0.0056	0.0134	0.0072	1			
G1m	0.0575	0.0021	0.1073	0.0005	1		
MATS4p	0.0202	0.0001	0.0001	0.0331	0.0386	1	
BEHm2	0.1867	0.3702	0.0035	0.02	0.0004	0.0297	1

Table 4. Architecture and Training Settings for the ANN Models

no. of nodes in the input layer	7 + 1[a]
no. of nodes in the hidden layer	6
no. of nodes in the output layer	1
learning rate	0.376
momentum	0.628
no. of iterations	10
transfer function	sigmoid

[a] Bias.

number Z_i of the atoms; the off-diagonal elements B_{ij} representing two bonded atoms i and j are equal to $\pi^* \times 10^{-1}$, where π^* is conventional bond order, that is, 0.1, 0.2, 0.3, and 0.15 for single, double, triple, and aromatic bonds, respectively; off-diagonal elements B_{ij} corresponding to terminal bonds are augmented by 0.01; all other matrix elements are set to 0.001 (*29*).

RESULTS AND DISCUSSION

The prediction ability of QSAR/QSPR models is affected by two factors. One is the descriptors, which should carry enough information of molecular structure for the interpretation of the activity/property. The other is the modeling method employed (*30*). The number of descriptors available for QSAR/QSPR studies is often so large that it is difficult to obtain a model including all of them. Therefore, identifying important descriptors certainly plays an important role in QSAR/QSPR. Descriptors should represent the maximum information in activity variations, and collinearity among them must be kept to a minimum.

The seven descriptors that have been selected using SPA are shown in **Table 2**. As can be seen from the correlation matrix (**Table 3**), there is no significant correlation between the selected descriptors. These descriptors were used for both linear and nonlinear models.

To build and test the models, the data set comprising 124 compounds was separated into a training set of 62 compounds, a validation set of 31 compounds, and a test set of 31 compounds. By using the training set, with the seven selected descriptors, the following linear model was obtained:

$$\text{Log } K_{OC} = 1.72965 + 0.333407\text{HTp} + 0.936504\text{MATS6e} + 0.78916\text{G3v} + 0.122569\text{Mor05m} + 4.201647\text{G1m} + 1.210337\text{MATS4p} - 0.711055\text{BEHm2}$$

This model was then used to predict the validation and test data. The prediction results are given in **Table 1**.

To construct an ANN model, a three-layer network with a sigmoid transfer function was employed. Before the networks were trained, the input values were normalized between −1 and 1. The initial weights were selected randomly between −0.3 and 0.3. The number of nodes in the hidden layer, as well as the learning rates and momentum values, were optimized. For this purpose, the network was trained with different numbers of nodes in the hidden layer, and the root-mean-square error (RMSE) between network outputs and target values was employed as performance metric. **Table 4** shows the architecture and training settings of the optimized network.

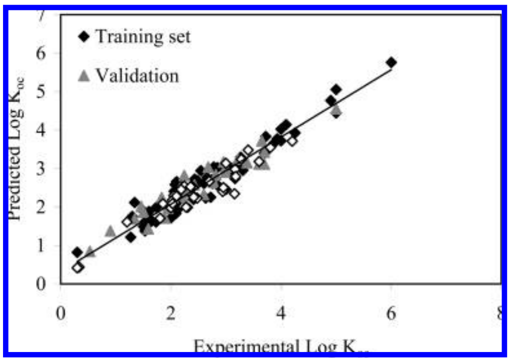

Figure 1. Plot of calculated soil sorption coefficients (log K_{OC}) against experimental values.

It should be noted that the training of the network for the prediction of log K_{OC} was interrupted when the RMSE of the validation set started to increase, to avoid overfitting. To select the best weight update function, two statistical methods were considered for evaluating the resulting models, namely, leave-one-out cross-validation (Q^2 LOO) and prediction standard error of estimation (SEP).

The compounds in the test set were not used during the training process and were reserved to evaluate the predictive power of the generated ANN. The ANN-calculated values of log K_{OC} for training, validation, and test sets are shown in **Table 1**. **Figure 1** shows the experimental values versus the predicted values by SPA-ANN. The residuals of the ANN calculated values of log K_{OC} are plotted against the experimental values in **Figure 2**. The presence of residuals at both sides of the zero line indicates that no systematic error exists in the development of the ANN model. For comparison, we also calculated K_{OC} by using the EPI package as shown in **Table 1**.

For evaluation of the predictive ability of the models, we employed the determination coefficient (R^2), root-mean-square error of prediction (RMSEP), relative standard error of prediction

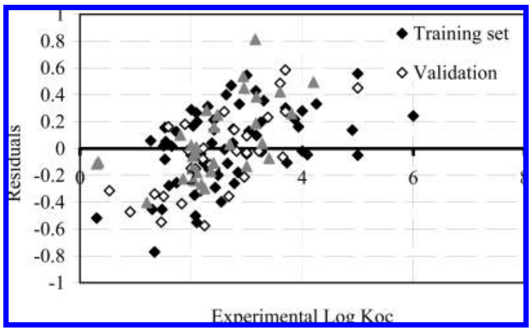

Figure 2. Plot of residuals versus experimental values of soil sorption coefficients (log K_{OC}).

Table 5. Comparison of the Statistical Parameters Obtained Using the SPA-MLR and SPA-ANN Models

parameters		SPA-ANN	SPA-MLR	EPI
	training set	0.2824	0.4202	
RMSEP	validation set	0.3007	0.3992	1.3200
	test set	0.2888	0.3705	
	training set	9.8423	14.646	
RSEP (%)	validation set	11.377	15.102	48.085
	test set	11.164	14.324	
	training set	6.0120	7.4011	
MAE (%)	validation set	8.8812	10.149	9.0304
	test set	8.5345	9.7164	
	training set	0.9314	0.8402	
R^2	validation set	0.9196	0.8262	0.0052
	test set	0.9029	0.8199	
	training set	814.57	315.50	
F statistical	validation set	331.49	137.85	0.6333
	test set	269.77	132.05	
	training set	28.541	17.762	
t test	validation set	18.207	11741	0.7958
	test set	16.425	11.491	

(RSEP), and mean absolute error (MAE) values (*31*), which are defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{\text{pred}} - y_{\text{obs}})^2}{\sum_{i=1}^{n} (y_{\text{obs}} - y_{\text{meas}})^2}$$

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} (y_{\text{pred}} - y_{\text{obs}})^2}{n}}$$

$$\text{RSEP } (\%) = 100 \times \sqrt{\frac{\sum_{i=1}^{n} (y_{\text{pred}} - y_{\text{obs}})^2}{\sum_{i=1}^{n} (y_{\text{obs}})^2}}$$

$$\text{MAE } (\%) = \frac{100}{n} \times \sqrt{\sum_{i=1}^{n} |(y_{\text{pred}} - y_{\text{obs}})|}$$

where y_{pred} is the predicted value of the property under consideration, y_{obs} is the experimental value, and n is the number of samples in the set. These statistical parameters for SPA-MLR, SPA-ANN, and EPI are shown in **Table 5**. As can be seen, SPA-ANN provided the most accurate predictions of log K_{OC} for the pesticides employed in the study.

LITERATURE CITED

(1) Sabljic, A.; Piver, W. T. Quantitative modeling of environmental fate and impact of commercial chemicals. *Environ. Toxicol. Chem.* **1992**,*11*, 961−972.

(2) Altenburger, R.; Boedeker, W.; Faust, M.; Grimme, L. H. Comparative hazard identification for pesticides: interrelations between physico-chemical properties, tonnages, and occurrence in surface waters. *Sci. Total Environ. Suppl.* **1993**, 1633−1654.

(3) Tarazona, J. V.; Fresno, A.; Aycard, S.; Ramos, C.; Vega, M. M.; Carbonell, G. Assessing the potential hazard of chemical substances for the terrestrial environment. Development of hazard classification criteria and quantitative environmental indicators. *Sci. Total Environ.* **2000**, *247*, 151−164.

(4) Goodarzi, M.; Freitas, M. P. *QSAR Comb. Sci.* **2008**, *27*, 1092−1098.

(5) Fatemi, M. H.; Goudarzi, N. *Electrophoresis* **2005**, *26*, 2968−2973.

(6) Jury, W. A. Adsorption of organic chemicals onto soil. In *Vadose Zone Modeling of Organic Pollutants*; Henn, S. C., Melancon, S. M., Eds.; Lewis Publisher: Boca Raton, FL, 1986; pp 177−189.

(7) Hasset, J. J.; Banwart, W. L. *Reactions and Movement of Organic Chemicals in Soils*; Special Publication 22; Soil Science Society of America and American Society of Agronomy (SSSA): Madison, WI, 1989; pp 31−44.

(8) Augustijn-Beckers, P. W. N.; Hornsby, A. G.; Wauchope, R. D. *Rev. Environ. Contam. Toxicol.* **1994**, *137* 1−82.

(9) Hansch, C.; Leo, A. Exploring QSAR. *Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(10) Trinajstic, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992.

(11) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Rev. Soc.* **1995**, *24*, 279−287.

(12) Araujo,M. C. U.; Saldanha, T. C. B.; Galvão, R. K. H.; Yoneyama, T.; Chame, H. C.; Visani, V. *Chemom. Intell. Lab. Syst., Lab. Inf. Manag.* **2001**, *57*, 65−73.

(13) Dantas Filho, H. A.; Souza, E. S. O. N.; Visani, V.; Barros, S. R. R. C.; Saldanha, T. C. B.; Araújo, M. C. U.; Galvão, R. K. H. *J. Braz. Chem. Soc.* **2005**, *16*, 58−61.

(14) Di Nezio, M. S.; Pistonesi, M. F.; Fragoso, W. D.; Pontes, M. J. C.; Goicoechea, H.C.; Araújo, M. C. U.; Band, B. S. F. *Microchem. J.* **2007**, *85*, 194−200.

(15) Grunhut, M.; Centurión, M. E.; Fragoso, W. D.; Almeida, L. F.; Araujo, M. C. U.; Band, B. S. F. *Talanta* **2008**, *75*, 950−958.

(16) Galvão,R. K. H.; Pimentel, M. F.; Araujo, M. C. U.; Yoneyama, T.; Visani, V. *Anal. Chim. Acta* **2001**, *443*, 107−115.

(17) Honorato, F. A.; Galvão, R. K. H.; Pimentel, M. F.; Neto, B. B.; Araújo, M. C. U.; Carvalho, F. R. *Chemom. Intell. Lab. Syst., Lab. Inf. Manag.* **2005**, *76*, 65−72.

(18) Breitkreitz, M. C.; Raimundo, I. M., Jr.; Rohwedder, J. J. R.; Pasquini, C.; Dantas Filho, H. A.; José, G. E.; Araújo, M. C. U. *Analyst* **2003**, *128*, 1204−1208.

(19) Dantas Filho, H. A. D.; Galvão, R. K. H.; Araújo, M. C. U.; Silva, E. C.; Saldanha, T. C. B.; José, G. E.; Pasquini, C.; Raimundo, I. M., Jr.; Rohwedder, J. J. R. *Chemom. Intell. Lab. Syst., Lab. Inf. Manag.* **2004**, *72*, 83−91.

(20) Pontes, M. J. C.; Galvão, R. K. H.; Araújo, M. C. U.; Moreira, P. N. T.; Pessoa Neto, O. D.; José, G. E.; Saldanha, T. C. B. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 11−18.

(21) Gambarra-Neto, F. F.; Marino, G.; Araújo, M. C. U.; Galvão, R. K. H.; Pontes, M. J. C.; de Medeiros, E. P.; Lima, R. S. *Talanta* **2009**, *77*, 1660−1666.

(22) Galvão, R. K. H.; Araújo, M. C. U.; Fragoso, W. D.; Silva, E. C.; José, G. E.; Soares, S. F. C.; Paiva, H. M. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 83−91.

(23) Galvão, R. K. H.; Araújo, M. C. U.; Silva, E. C.; José, G. E.; Soares, S. F. C.; Paiva, H. M. *J. Braz. Chem. Soc.* **2007**, *18*, 1580−1584.

(24) Galvão, R. K. H.; Araújo, M. C. U. Linear regression modeling: variable selection. In *Comprehensive Chemometrics*; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2009.

(25) Barcelo, D.; Hennion, M. C. *Trace Determination of Pesticides and Their Degradation Products in Water*; Elsevier Sciences: Amsterdam, The Netherlands, 1997.

(26) http://epa.gov/opptintr/exposure/pubs/episuite.htm.

(27) HyperChem Release 7, HyperCube, Inc., http://www.hyper.com.

(28) Todeschini, R. Milano Chemometrics and QSPR Group, http://www.disat.unimib.it/vhml.

(29) Consonni, V.; Todeschini, R. *Handbook of MolecularDescriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(30) Goodarzi, M.; Freitas, M. P. *J. Phys. Chem. A* **2008**, *112*, 11263−11265.

(31) Goodarzi, M.; Goodarzi, T.; Ghasemi, N. *Ann. Chim* **2007**, *97*, 303−312.

Received December 2, 2008. Revised manuscript received May 26, 2009. Accepted May 29, 2009. M.G. thanks the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) of Argentina for the financial support.