QSPR study of the water solubility of a diverse set of agrochemicals: hybrid (GA/ MLR) approach

Etude QSPR de la solubilité aqueuse d'un ensemble de divers produits agrochimiques: approche hybride (AG/RLM)

Amel Bouakkadia, Hamza Haddag, Nabil Bouarra & Djelloul Messadi*

Environmental and Food Safety Laboratory, Badji Mokhtar University, Annaba , PO Box 12, 23000, Algeria.

Soumis le : 05/03/2015 *Révisé le : 21/03/2016* *Accepté le : 12/04/2016*

ملخص

أجريت علاقة بين كمية الهيكل و الخاصية للتنبؤ بلنحلالية المبيدات منتمية إلى أربعة أقسام كيميائية: أحماض، اليوريا، تريازين، وكاربامات. المجموعة المكونة من 77 مبيد قسمت إلى مجموعة بناء من 58 مبيد ومجموعة اختبار من 19 مبيد بتقنية. النموذج بستة متغيرات بمعامل ارتباط (R^2) يساوي 0.8895 و خطا معيار التقدير (s) يساوي 0.52 وحدة تم تطويره بتطبيق التراجع المتعدد الخطي باستخدام المربعات الصغرى واختيار مجموعة المتغيرات تم باستعمال الخوارزمية الجينية. قوة النموذج المقترح تأكدت باستخدام عدة تقنيات للتقييم leave- one- out ، bootstrap ، الاختبارات العشوائية، والتحقق من خلال مجموعة الاختبار.

الكلمات الدالة: *المبيدات ، الانحلالية، QSPR ، الموصفات الجزيئية، التراجع المتعدد الخطي.*

Abstract

A quantitative structure- property relationship (QSPR) was performed for the prediction of the aqueous solubility of pesticides belonging to four chemical classes: acid, urea, triazine, and carbamate. The entire set of 77 pesticides was divided into a training set of 58 pesticides and a test set of 19 pesticides according to the Snee technique. A six descriptor model, with squared correlation coefficient (R^2) of 0.8895 and standard error of estimation (s) of 0.52 log unit, was developed by applying multiple linear regression analysis using the ordinary least square regression method and genetic algorithm- variable subset selection. The reliability of the proposed model was further illustrated using various evaluation techniques: leave- one- out cross- validation, bootstrap, randomization tests, and validation through the test set.

Key Words: *pesticides- aqueous solubility- QSPR- molecular descriptors- multiple linear regression.*

Résumé

Une relation quantitative structure-propriété (QSPR) a été réalisée pour la prédiction de la solubilité aqueuse des pesticides appartenant aux quatre classes chimiques: acide, urée, triazine, et carbamate. L'ensemble des 77 pesticides a été divisé en un ensemble de calibrage de 58 pesticides et un ensemble de test de 19 pesticides selon la technique de Snee. Un modèle de six descripteurs, avec un coefficient de corrélation (R^2) de 0,8895 et une erreur standard d'estimation (s) de 0,52, a été développé en appliquant une analyse de régression linéaire multiple en utilisant la méthode de régression des moindres carrés ordinaires et les algorithme-génétiques pour la sélection des sous-ensembles de variables. La fiabilité du modèle proposé a été en outre illustrée en utilisant diverses techniques d'évaluation: validation croisée par leave- one- out, bootstrap, tests de randomisation, et la validation par l'ensemble de test.

Mots clés: *pesticides- solubilité aqueuse- QSPR- descripteurs moléculaires- régression linéaire multiple.*

*Corresponding author : d_messadi@yahoo.fr

1. INTRODUCTION

The massive use of agrochemicals, known generically as pesticides [1], has allowed significant reduction in the agricultural plagues, and consequently, increased the productivity. On the other hand, the massive use of these agrochemicals has an environmental cost (due to their toxicity, their persistence, or their tendency to bioaccumulation), which is necessary to evaluate to conciliate productivity and environment protection [2].

Solubility in water is an important physicochemical property, having numerous applications to the modeling of the environmental effects of chemicals [3]. It is a direct measurement of hydrophobicity, that is, the tendency of water to exclude the substance from solution. Although the experimental determination of solubility is not difficult, there are some justifications to develop models that can predict it. This is especially important in environmental studies where the compounds are toxic, carcinogenic, or undesirable for some or other reason.

An extensive series of studies for the prediction of aqueous solubility has been reported in the literature [4- 10]. These methods can be categorized into three types:

1 - Correlation of solubility with experimentally data such as melting point (MP) and log P (logarithme of octanol/ water partition coefficient). However, this approach is of little use because it requires a knowledge of the compound's experimental melting point which is not available for virtual compounds. The melting point is a key index of the cohesive interactions in the solid and it is difficult to estimate.

2 - Estimation of solubility by group contribution methods. The group contribution method allows the approximate calculation of solubility by summing up fragmental values associated with substructural units of the compounds. The disadvantages of the group contribution method are that: 1/ the groups included must be defined in advance and therefore the solubility of a new compound containing new groups cannot be estimated; and 2/ the different effects of a group in different chemical environments are not considered.

3 - Correlation of solubility with descriptors derived from the molecular structure by computational methods. This third approach has been proven to be particularly successful for the prediction of solubility because it does not need experimental descriptors and can therefore also be applied to collections of virtual compounds.

The aim of the present work is to develop a robust QSPR model that could predict the aqueous solubility values for a diverse set of agrochemicals (which consists of 26 acids, 25 ureas, 13 triazines and 13 carbamates) using the general molecular descriptors computed with the help of DRAGON software [11].

2. METHODS

2.1 Experimental Data

The experimental S values (mg/l) of 77 selected, structurally heterogeneous, pesticides were taken from Hansen [12]. The water solubility values (log S) span between -1.05 and 5.90 (Table 1). The detailed structures of all studied compounds are available as Supporting Information.

2.2 Descriptor Generation

The chemical structure of each compound was sketched on a PC using the HYPERCHEM program [13] and preoptimized using MM+ molecular mechanics method (Polack- Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi- empirical PM3 method at a restricted Hartree- Fock level with no configuration interaction, applying a gradient norm limit of 0.01 kcal.Å^{-1}.mol^{-1} as a stopping criterion. Then the geometries were used as input for the generation of 1664 descriptors using the Dragon software (version 5.4) [11]. Quantum-chemical descriptors such as HOMO (highest occupied molecular orbital), LUMO (lowest unoccupied molecul arorbital), HOMO – LUMO gap (DHL), and ionization potential (P_{ion}), calculated by the semi empirical PM3 method using [13], were added and used for descriptor selection during model development. Constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (when there was more than 98% pairwise correlation, one variable was deleted), and the genetic algorithm was applied for variables selection to a final set of 1230 descriptors.

2.3 Selection of the training and test sets

It is important to rationally define a training set from which the model is built and external test set on which to evaluate its prediction power. The object of this selection should be to generate two sets with similar molecular diversity, in order to be reciprocally representative and to cover all the main structural and physicochemical characteristics of the global data set.

Several procedures can be adopted for the selection of the training and test sets, the later should contain between 15 and 40% of the compounds in the full data set.

DUPLEX algorithm adopted in this study proceeds as follows. In the first step, the two points which are furthest away from each other are selected for the training set. From the remaining points, the two- objects which are furthest away are included in the test set. In the third step, the remaining point which is furthest away from the two previously selected for the training set is included in that set. The procedure is repeated selecting a single point for the test set which is furthest from the existing points in that set. Following the procedure, points are added alternately to each set [14]. This algorithm was applied in the present study to separate data into two independent subsets: a training set of 58 compounds to build the model and a test set of the remained 19 compounds to evaluate its prediction ability.

2.4 Model Development and Validation

Multiple linear regression analysis (MLR) and variable selection were performed by the software MobyDigs [15] using the Ordinary Least Square regression (OLS) method and Genetic Algorithm-Variable Subset Selection (GA-VSS) [16].

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 overparameterization, which would lead to a loss of predictive power for molecules outside training set. 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 $n/\ m \geq 5$ [17]. 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) [18] a necessary condition for the model validity. Acceptable model is only that 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.

The collinearity in the original set of molecular descriptors results in many similar models that more or less yield the same predictive power (in MOBYDIGS software 100 models of different dimensionality). Therefore, when there were models of similar performance, those with higher ΔK (K_{xy}- K_{xx}) were selected and further verified.

The models were justified by the R^2, the adjusted R^2, the cross-validated values of Q^2 by leave-one-out (LOO), the F ratio values and the standard error s.

The robustness of the models and their predictivity were evaluated by both Q^2_{LOO} and bootstrap. In this last procedure K n-dimensional groups are generated by a randomly repeated selection of n- objects from the original data set.

The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 8000 times.

The proposed model was also checked for reliability and robustness by permutation testing: new models are recalculated for randomly recorded response (Y- scrambling) by using the same original independent variable matrix. After repeating this test several times (100 times in this work) it is expected to obtain new models that have significantly lower R^2 and Q^2 than the original model. If this condition is not verified the original model is not acceptable, as it was due to a chance correlation or a structural redundancy in the training set.

Obtaining a robust model does not give real information about its prediction power. This is evaluated by predicting the compounds included in the test set.The external Q^2_{ext} for the test set is determined with equation (1):

$$Q^2_{ext} = 1 - [(\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_{(i)})^2 / n_{ext}) / (\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2 / n_{tr})] \qquad (1)$$

Here n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively.

2.5 Applicability Domain Analysis

The applicability domain (AD) [19, 20] is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSPR should make reliable predictions. In this work, the structural AD was verified by the leverage (h_{ii}) approach [21].

The warning leverage h^* is, generally, fixed at $3(m + 1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation.

The presence of both the response outliers (Y outliers) and the structurally influential compounds (X outliers) was verified by the Williams plot [22], the plot of standardized residuals versus leverage values.

Table 1: Experimental and calculated logS for the studied pesticides.

No	Expt. logS	Calc. logS	Residual	No	Expt. logS	Calc. logS	Residual
1	3.89	3.86	0.03	40	1.88	2.47	-0.59
2	2.18	1.97	0.21	41	2.87	3.03	-0.16
3	2.95	3.44	-0.49	42 *	1.64	1.45	0.19
4	5.90	5.05	0.85	43	2.87	2.63	0.24
5	1.66	1.50	0.16	44	2.93	2.63	0.30
6	2.00	2.76	-0.76	45	3.23	2.72	0.51
7	2.27	1.85	0.42	46	1.77	2.56	-0.79
8	1.52	1.71	-0.19	47 *	2.83	2.10	0.73
9 *	2.08	2.86	-0.78	48 *	3.09	2.33	0.76
10 *	2.85	3.55	-0.70	49	3.98	3.53	0.45
11	1.64	1.75	-0.11	50 *	1.87	1.26	0.61
12	3.54	3.68	-0.14	51	2.00	1.78	0.22
13 *	0.40	0.85	-0.45	52	0.67	1.41	-0.74
14 *	1.95	1.88	0.07	53	2.63	3.27	-0.64
15 *	4.45	3.48	0.97	54 *	2.39	2.93	-0.54
16	3.08	2.79	0.29	55	2.86	1.94	0.92
17	5.75	5.71	0.04	56	1.52	1.49	0.03
18	5.16	4.95	0.21	57	-0.20	-0.11	-0.09
19	2.23	1.81	0.42	58	0.93	1.22	-0.29
20 *	1.98	1.80	0.18	59	3.60	2.91	0.69
21	0.95	0.27	0.68	60	1.12	0.90	0.22
22	2.76	2.10	0.66	61	-0.51	-0.14	-0.37
23	2.54	2.55	-0.01	62	3.86	3.32	0.54
24 *	2.77	2.55	0.22	63 *	0.79	1.94	-1.15
25 *	1.30	1.14	0.16	64	3.40	3.68	-0.28
26	1.20	0.93	0.27	65 *	2.85	3.48	-0.63
27	1.62	1.40	0.22	66	1.34	1.75	-0.41
28 *	2.54	2.35	0.19	67 *	0.93	0.81	0.12
29	1.70	2.81	-1.11	68	3.80	4.13	-0.33
30	-0.10	-0.66	0.56	69	1.45	1.26	0.19
31	0.00	-0.18	0.18	70	0.60	0.68	-0.08
32	0.30	-0.18	0.48	71	2.91	2.40	0.51

33	2.04	2.38	-0.34
34	-1.05	-0.59	-0.46
35	4.08	4.29	-0.21
36	1.64	1.81	-0.17
37	0.11	0.74	-0.63
38	1.81	2.38	-0.57
39	0.78	1.61	-0.83
72 *	3.18	3.44	-0.26
73	3.91	3.17	0.74
74	1.36	0.74	0.62
75	2.04	2.19	-0.15
76 *	2.03	2.33	-0.30
77	0.53	1.47	-0.94

* Members for the test set.

3. RESULTS AND DISCUSSION

3.1 Results of the MLR Model

The dissolving process is the establishment of equilibrium between the phase of solute and its saturated aqueous solution. Aqueous solubility is almost exclusively dependent on the intermolecular forces that exist between the solute molecules and the water molecules. The solute- solute, solute- water, and water- water adhesive interactions determine the amount of compound dissolving in water. Additional solute- solute interactions are associated with the lattice energy in the crystalline state.

The solubility of a compound is thus affected by many factors: the state of solute, the relative aromatic and aliphatic degree of the molecules, the size and shape of the molecules, the polarity of the molecule, steric effects, and the ability of some groups to participate in hydrogen bonding.

In order to predict solubility accurately, all these factors correlated with solubility should be represented numerically by descriptors derived from the structure of the molecule.

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

log S = - 2.80 - 1.27 **E_{HOMO}** - 0.182 **Mor02v** - 17.2 **G2e** - 9.56 **HATS7v** + 4.76 **RTu+** - 0.0821 **AlogP2** (2)

R^2= 0.8895 R^2_{adj}= 0.8765 Q^2_{LOO} = 0.8547 Q^2_{EXT} =0.8511 Q^2_{BOOT} = 0.8323 s = 0.52 log unit F = 68.42

K_{xx}= 37.68 K_{xy} = 45.67

Here, E_{HOMO} is the Highest Occupied Molecular Orbital energy [23, 24] ; Mor 02 v is the 3D-MoRSE- signal 02/ weighted by atomic van der Waals volume [25, 26] ; G2e is the second component symmetry directional WHIM index/ weighted by atomic Sanderson electronegativities [27, 28] ; HATS7v is the leverage weighted autocorrelation of lag 7/ weighted by atomic van der Waals volumes [29, 30] ; RTu+ is the R maximal index/ unweighted [29, 30] ; AlogP2 is the squared Ghose-Crippen-Viswanadhan octanol-water partition coefficient [31, 32].

More information about these descriptors can be found in [33] and the references therein.

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 1. The statistics for the modified logS vectors are clearly lower than the real QSPR model. This ensures that a real structure-property relationship has been found out.

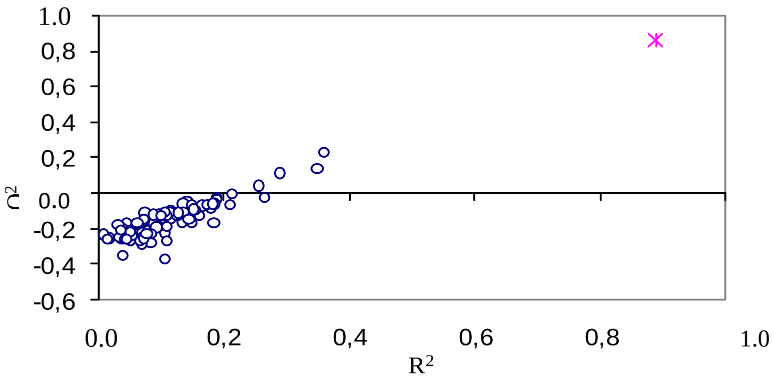

Figure1. Randomization test associated to previous QSPR model. Circles represent the randomly ordered solubilities, and star corresponds to the real solubilities.

Some important statistical parameters (as given in table 2) were used to evaluate the involved descriptors. The t-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute t-values shown in table 2 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 can describe the statistical significance when combined together within an overall collective QSPR model (i. e., descriptor's interactions). Descriptors with t-probability values below 0.05 (95% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance [34]. The smaller t-probability suggests the more significant descriptor. The t-probability values of the six descriptors are very small, indicating that all of them are highly significant descriptors. The VIF values suggest that these descriptors are weakly correlated with each others. Thus, the model can be regarded as an optimal regression equation.

The calculated log S values from equation (2) for the training and test set are showed in table 1 and figure 2. The distribution of errors for the entire data set is given in figure 3. As the errors are distributed on both sides of the zero line, one may conclude that there is no systematic error in the developed model.

Table 2. Characteristics of the selected descriptors in the best MLR model

Descriptor	Descriptor type	X	Dx	t- value	t-probability	VIF
Constant		-2.801	2.545	-1.1	0.276	
E_{HOMO}	Quantum-chemical descriptors	-1.267	0.245	-5.18	0.000	1.1
Mor02v	3D- MoRSE descriptors	-0.182	0.031	-5.78	0.000	4
G2e	WHIM Index	-17.202	4.131	-4.16	0.000	2.1
HATS7v	GETAWAY descriptors	-9.561	1.492	-6.41	0.000	1.2
RTu+	GETAWAY descriptors	4.762	1.909	2.49	0.000	3.2
AlogP2	Molecular properties	-0.082	0.014	-5.96	0.000	2.1

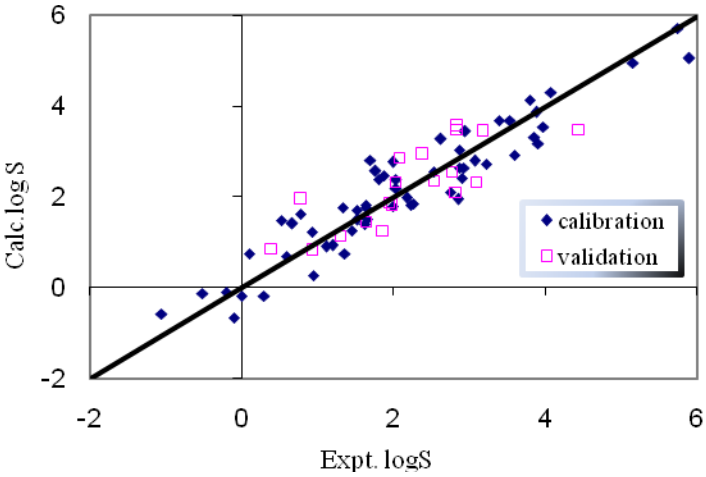

Figure 2. Plot of predicted *vs*. experimental logS for the entire data set.

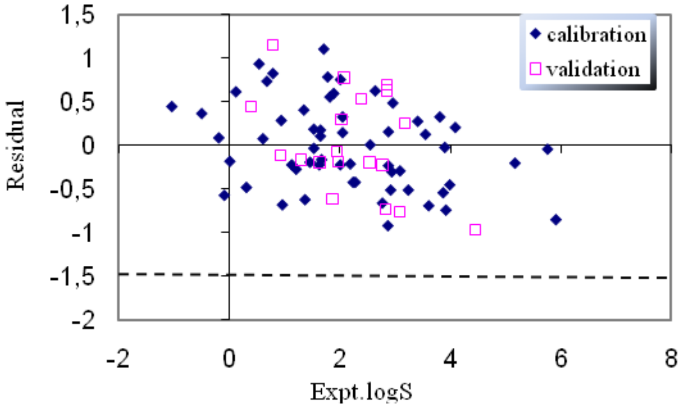

Figure 3.Plot of residual *vs*. experimental logS for the entire data set.

3.2 Descriptor Contribution Analysis and Interpretation

Based on a previously described procedure [35, 36], the relative contribution of the six descriptors to the model were determined and they decrease in the following order: HATS7v (17.91%) > Mor2v (17.67%) > HOMO (16.94%) > AlogP2 (16.80%) > G2e (15.76%) >RTu+ (14.89%). It should be noted that the difference in the descriptor contribution between any two descriptors used in the model is not significant, indicating that all of the descriptors are indispensable in generating the predictive model (Fig.4).

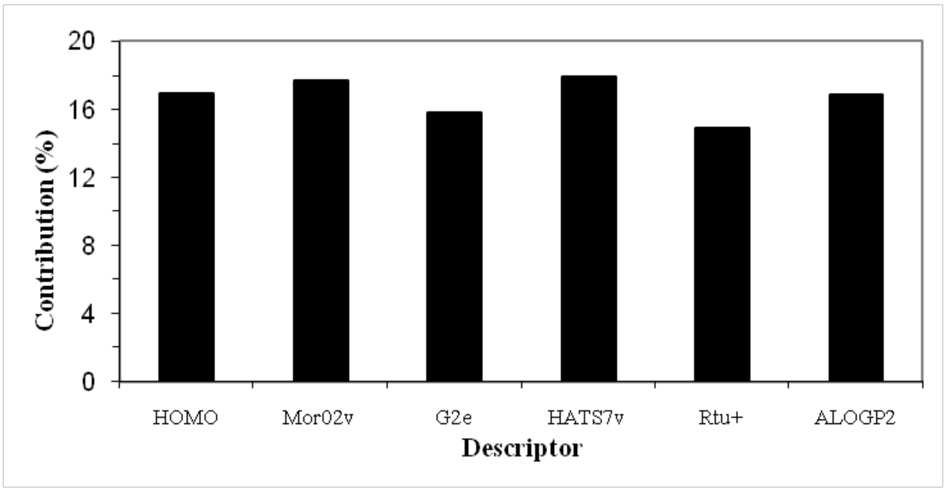

Figure 4.Relative contributions of the selected descriptors to the MLR model.

The importance of atomic van der Waals volumes on the log S values is apparent, since the descriptors weighted by atomic van der Waals explain 35.58% of the contributions (17.91% of HATS7v, and 17.67% of Mor2v). The first important descriptor is HATS7v, which has a relatively high negative correlation with the experimental log S values (R= -0.328). The negative coefficient of HATS7v indicates that the agrochemicals with larger values for this descriptor would have lower log S values.

The second important descriptor is Mor02v, a 3D- MoRSE descriptor, which has a smaller negative correlation coefficient with the experimental log S values (R= -0.787). 3D-MoRSE descriptors are the 3D molecular representations of structure based on electron diffraction descriptor [25, 26], which are calculated by summing atomic weights viewed by a different angular scattering function. The values of these descriptor functions are calculated at 32 evenly distributed values of scattering angle (s) in the range of 0- 31A° from the three dimensional atomic coordinates of a molecule. The 3D- MoRSE descriptor is calculated using following expression:

$$\text{Morsw} = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i w_j (\sin(s.r_{ij}) / s.r_{ij}) \quad (3)$$

where s is the scattering angle, nAT is the number of atoms, r_{ij} is the interatomic distance between the i^{th} and the j^{th} atoms, w is an atomic property, including atomic number, masses, van der Waals volumes, Sanderson electronegativities, and polarizabilities. The coefficient of Mor02v is negative, indicating that an increase in Mor02v would result in a decrease in log S values.

Hence, as expected, atomic volumes have a specific effect on the log S values: an increase in Mor02v (or in HATS7v) would result in a decrease in log S values.

The Squared-Ghose- Crippen-Viswanadhan octanol-water partition coefficient (AlogP2) [31, 32] is calculated from a regression equation based on the hydrophobic character of the molecule. It reflects both the interactions of the solute with the bulk of the surrounding solvent (macroscopic or non specific solvent effects) and the specific bonding between the solute and individual solvent molecules (microscopic or specific solvent effects). When this descriptor increases, the log S decreases.

Highest occupied molecular orbital energy (E_{HOMO}) is a measure of the nucleophicity of a molecule. It should explain the differences in the tendency of solutes to take part in the charge transfer interactions, i. e. the ability of electron- donating to water molecules of solute molecules. According to the Koopmans theorem [37], the energy of the HOMO is directly related to the ionization potential IP (-E_{HOMO} = IP), provided that the ionization process is adequately represented by the removal of an electron from an orbital without change in the wave functions of the other electrons. The descriptor and its coefficient in the model are negative, so the contribution of E_{HOMO} is positive.

The importance of the axial shape and symmetry of the molecule on the log S values is apparent due to the presence of G2e. In the calculations Sanderson atomic electronegativity was used for each atom because it may determine, with other atomic properties, the macroscopic properties of a compound. The positive sign of G2e means that the increase in this descriptor decreases the log S.

RTu+, as HATS7v, is a GETAWAY descriptor and correlates with the experimental log S values of 0.490. The GETAWAY descriptors [29, 30] have been proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular influence matrix. These descriptors, as based on spatial autocorrelation, encode information on molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties.

HATS7v and RTu+ are calculated by Equations. (4) and (5) respectively.

$$\text{HATSk(w)} = \sum_{i=1}^{A} \sum_{j\neq i}^{A} (w_i.h_i)(w_j.h_j)\delta(d_{ij};k)$$

$$\text{for } k=0,1,2,3,\ldots D \qquad (4)$$

$$\text{RTu+} = \text{maxij}(\frac{\sqrt{h_i h_j}}{r_{ij}}.w_i.w_j.\delta(d_{ij};k)$$

$$i \neq j \qquad k= 0, 1, 2,3,\ldots D \qquad (5)$$

where A is the number of atoms, w is an atomic weighting scheme, d_{ij} is the topological distance, δ (k, d_{ij}) is a Dirac- delta function (δ=1 if d_{ij}=k, zero otherwise), r_{ij} is the interatomic distance. D is the molecule topological diameter that is the maximum topological distance in the molecule.The coefficient of RTu+ is positive, meaning that the pesticides with larger values for this descriptor have larger log S values.

3.3 Applicability Domain of the MLR Model

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 was analyzed in the Williams plot (shown in Fig.5). Clearly observation 35 of the training set with leverage higher than the warning limit of 0.36 is a structurally influential compound. Deleting observation 35 could alter slightly R^2 between the experimental logS values and the selected descriptors to 0.8866 (Q^2= 0.8485) and increase the standard error to 0.524, while utilization of a higer energy conformation geometry for this observation alter negatively the calculated model.

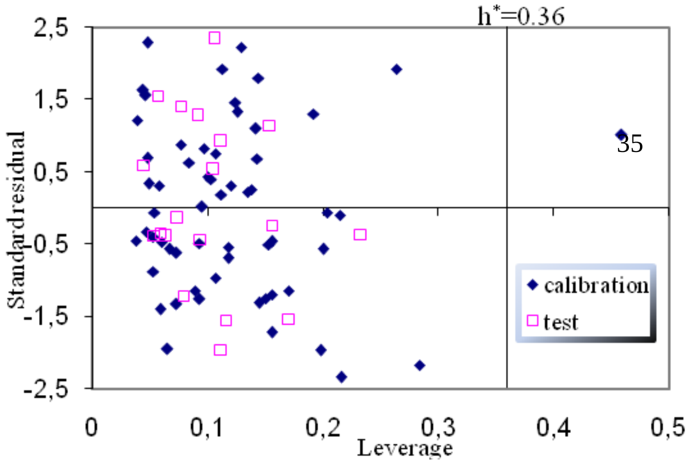

Figure 5. Williams plot of the MLR model for the entire data set.

4. CONCLUSION

In this paper, the QSPR method was applied to the prediction of the aqueous solubility of various type of pesticides. A six- parameter linear model was developed by hybrid GA/ MLR approach with R^2 of 88.95 and s of 0.52 log unit for the training set. The selected descriptors express many factors influencing aqueous solubility, to name: molecular size and shape, specific atomic properties, both macroscopic and microscopic effects and tendency of solutes to take part in the charge transfer interactions. Several validation techniques, including leave-one-out cross-validation and bootstrap, randomization tests, and validation through the test set, illustrated the reliability of the proposed model. All of the descriptors can be directly calculated from the molecular structure of the compound, thus the proposed model is predictive and could be used to estimate the solubility of pesticides. In this case, the applicability domain will serve as a valuable tool to filter out “dissimilar” chemical structures.

REFERENCES

[1] Price N R., Watkins R W., 2003. Quantitative structure-activity relationships (QSAR) in predicting the environmental safety of pesticides, *Pestic. Outlook*, Vol. 14, pp. 127- 129.

[2] Stevens J T., Breckenridge C B., 2001. Agricultural chemicals: regulation, risk assessment, and risk management, in Regulatory Toxicology. Ed. By S. C. Gad (Taylor & Francis Ltd., London,). 215 p.

[3] Mackay D., 2000. In Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences. Ed. By R. S. Boethling and D. Mackay (CRC Press LLC, Boca Raton) 205 pages.

[4] Lipinski C A., Lombardo F., Doming D W., Feeney P J., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliver. Rev*, Vol. 64, pp. 3- 25

[5] Jorgensen W L., Duffy E M., 2002. Prediction of drug solubility from structure. *Adv. Drug Deliver. Rev*, Vol. 54, pp. 355 – 366

[6] Kartizky A R., Kuanar M., Slavov S., Hall C. D., 2010. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem.Rev*, Vol. 110, pp. 5714 – 5789

[7] Skyner R E., McDonagh J L., Groom C R., van Mourik T., Mitchell J B O., 2015. A review of methods for the calculation of solution free energies and the modeling of systems in solution. *Phys. Chem. Chem. Phys*, Vol. 17, pp. 6174- 6191.

[8] Ruelle P., Kesselring U W., 1997. Aqueous solubility prediction of environmentally important chemicals from the mobile order thermodynamics. *Chemosphere*, Vol. 34, pp. 275- 298.

[9] Deeb O., Goodarzi M., 2010. Predicting the solubility of pesticides compounds in water using QSPR methods. *Molecular Physics*, Vol. 108, pp. 181- 192.

[10] Ran Y., He Y., Yang G., Johnson J L H., Yalkowsky S H., 2002. Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere,* Vol. 48, pp. 487- 509.

[11] Todeschini R., Consonni V., Mauri A., Pavan M., 2005. DRAGON Software – version 5.4-TALETE srl

[12] Hansen O C., 2004. Quantitative Structure-Activity Relationships (QSAR) and Pesticides. (Pesticides Research No. 94. The Danish Environmental Protection Agency,), http://www2.mst.dk/udgiv/publications/2004/87-7614-434-8/pdf/87-7614-435-6.pdf. (accessed 26-05-2014)

[13] Hyperchem^{TM}. Release 6.02 for windows. 2000. Molecular Modeling system

[14] Snee R D., 1977. Validation of Regression Models: Methods and Examples, *Technometrics*, Vol. 19, pp. 415-428

[15] Todeschni R., Ballabio D., Consonni V., Mauri A., Pavan M., 2009. MOBYDIGS – version 1.1 – Copyright TALETE srl (2004).

[16] Leardi R., Boggia R., Tarrile M., 1992. Genetic Algorithm as a Strategy for Feature Selection, *J. Chemom*, Vol. 6, pp. 267 – 281

[17] Xu J., Zhang H., Wang Lei., Liang G., Wang Luoxin., Shen X., Xu W., 2010. QSPR study of absorption maxima of organic dye- sensitized solar cells based on 3D descriptors. *Spectrochimica Acta Part A*, Vol. 76, pp. 239-247.

[18] Todeschini R., Maiocchi A., Consonni V., 1999. The K Correlation Index: Theory Development and its Application in Chemometrics. *Chemom,* Int. Lab. Syst, Vol.46, pp. 13 – 29

[19] Tropsha A., Gramatica P., Gombar V K., 2003. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR Comb. Sc*i, Vol. 22, pp. 69 – 77

[20] Shen M., Béguin C., Golbraikh A., Stables J P., Kohn H., Tropsha A., 2004. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds, *J. Med. Chem*, Vol. 47, pp. 2356 – 2364

[21] Weisberg S., 2005. Applied Linear Regression, 3rd edn. (John Wiley and sons, Inc., New Jersey,)
[22] SCAN- Software for Chemometric Analysis- 1995. version 1.1- for Windows, Minitab USA.

[23] Clare B W., 1994. Frontier orbital energies in quantitative structure- activity relationships: a comparison of quantum chemical methods, *Theor. Chim. Acta*, Vol. 87, pp. 415 – 430

[24] Huang Q G., Kong I., Wang L S., 1996. Applications of Frontier molecular orbital energies in QSAR studies, *Bull. Environ. Contam.* Toxicol, Vol.56, pp. 758 – 765.

[25] Gasteiger J., Sadowski J., Schuur J., Selzer P., Steinhauer L., Steinhauer V., 1996. Chemical information in 3D space, *J. Chem. Inf. Comput. Sci*, Vol. 36, pp. 1030 – 1037.

[26] Schuur J., 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*, Vol. 36, pp. 334 – 344.

[27] Todeschini R., Lasagni M., Marengo E., 1994. New Molecular descriptors for 2D- and 3D – structures, *Theory. J. Chemom*, Vol. 8, pp. 263 – 272.

[28] Todeschini R., Gramatica P., Marengo E., Provenzani R., 1995. Weighted holistic invariant molecular descriptors. Part. 2. Theory development and applications on modeling physico- chemical properties of polyaromatic hydrocarbons (PAH), *Chemom. Intell. Lab. Syst,* Vol. 27, pp. 221 – 229.

[29] Consonni V., Todeschini R., Pavan M., 2002. Structure/ response correlations and similarity/ diversity analysis by GETAWAY descriptors. Part I. Theory of the novel 3D molecular descriptors, *J. Chem. Inf. Comput. Sci*, Vol.42, pp. 682 – 692.

[30] Consonni V., Todeschini R., Pavan M., Gramatica P., 2002. Structure/ responsecorrelations and similarity/ diversity analysis by GETAWAY descriptors. Part II. Application of the novel 3D molecular descriptors in QSAR/ QSPR studies, *J. Chem. Inf. Comput. Sci*, Vol. 42, pp. 693 – 705.

[31] Ghose A K., Crippen G M., 1986. Atomic physico-chemical parameters for three- dimensional- structure-directed quantitative structure- activity relationships. I. Partition coefficients as a measure of hydrophobicity, *J. Comput. Chem*, Vol. 7, pp. 565 – 577.

[32] Viswanadhan V N., Reddy M R., Bacquet R J., Erion M D., 1993. Assessment of methods used for predicting lipophilicity : application to nucleopides and nucleoside bases, *J. Comput. Chem*., Vol. 14, pp. 1019 – 1026.

[33] Todeschini R., Consonni V. , 2009. Molecular Descriptors for Chemoinformatics Volumes I & II. (WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009).

[34] Ramsey F. L., Schafer D. W., 2002. The Statistical Sleuth: A Course in Methods of Data Analysis, 2nd edn. (Wadsworth group, USA).

[35] Zheng F., Bayram E., Sumithran S P., Ayers J T., Zhen C G., Schmitt J D., Dwoskim L P., Crooks P A., 2006. Bioorg. Med. Chem, Vol. 14, pp. 3017 – 3037.
[36] Guha R., Jurs P C., 2005. Interpreting Computational Neural Network QSAR Models: A Measure of Descriptor Importance, *J. Chem. Inf. Model*, Vol. 45, pp. 800 – 806.

[37] Koopmans T C., 1933. Ordering of wave functions and eigenenergies to the individual electrons of an atom, *Physica*, Vol. 1, pp. 104-113.