OPEN ACCESS
Eurasian Journal of Analytical Chemistry
ISSN: 1306-3057
2017 12(7):1001-1014
DOI: 10.12973/ejac.2017.00228a

Application of Genetic Programming (GP) in Prediction of Gas Chromatographic Retention Time of some Pesticides

Mohammad Hossein Fatemi
University of Mazandaran, IRAN
Zahra Pahlevan Yali
University of Mazandaran, IRAN

Received 27 October 2016 ▪ Revised 11 July 2017 ▪ Accepted 14 July 2017

ABSTRACT
In this study, quantitative structure–retention relationship (QSRR) methodology was employed for modeling of gas chromatographic retention time for 74 pesticides. Stepwise multiple linear regression (SW-MLR) was used for the selection of most important descriptors. Multiple linear regression (MLR) and genetic programming (GP) were utilized to develop linear and symbolic regression equation models, respectively. Inspection to statistical parameters of developed MLR and GP models indicates symbolic regression equation via GP can be selected as the best fitted model. For this model, the square correlation coefficients (R^2) were 0.943 and 0.911, and the root-mean square errors (RMSE) were 2.56 and 2.77 for the training and test sets, respectively. The built GP model was assessed by leave one out cross-validation (Q^2_{cv} = 0.79, *SPRESS* = 2.57) as well as external validation. In addition, the result of sensitivity analysis of GP model suggest structural features and polarity are important factors responsible for gas-chromatographic retention time values of studied pesticides.

Keywords: quantitative structure–retention relationships, pesticide, retention time, multiple linear regression, genetic programming

INTRODUCTION

Pesticides are a group of organic compounds that are used in most sectors of the agricultural production on a large scale. These compounds prevent or reduce losses by pests and can improve quality and cosmetic appeal of the product [1, 2]. Pesticides can be classified based on functional groups in their molecular structure or their specific biological activity on target [3, 4]. Despite extensive use of these chemicals there are serious concerns about health risks arising for the general population from residues on food and drinking water [5, 6]. Most of these compounds have low rates of biodegradation and tendency to bioaccumulation that could make an environmental and human health risks [7-9].

Correspondence: Mohammed Hossein Fatemi, *Department of Chemistry, University of Mazandaran, Babolsar, Iran.*
mhfatemi@umz.ac.ir

Although the agricultural soil is the primary recipient of pesticide residues, water bodies that are adjacent to these areas are usually the ultimate recipient for these chemicals [10]. Pesticide residues frequently reported as common organic contaminants worldwide in surface and ground water [11-17]. Today, identification and quantification of pesticides in soil, air and water is very important field for chemist.

The traditional methods for determining residues of pesticides in environmental samples involve extraction that followed by gas chromatography (GC) analysis with nitrogen-phosphorus (NPD) or electron-capture detection (ECD) [18, 19]. The identification of separated pesticides can be done by comparison of their chromatographic retention with standard chemicals or from their MS spectra. Thought experimental determination of chromatographic retention of all pesticides is costly and time-consuming, therefore, developing of theoretical methods for estimation of this parameter for decreasing the time and costs of this experiment are very interesting and necessary. Among this method quantitative structure-retention relationship (QSRR) is a mathematical model based on the principle that the chromatographic retention of molecules is related to their structural features of molecules (molecular descriptors) which the obtained model can used to predict the retention for another compounds in the absence of experimental data [20-22].

There are some methods to related molecular descriptors to interested properties such as multiple linear regression (MLR), artificial neural network (ANN), partial least square (PLS), support vector machine (SVM) and genetic programming (GP). Among them GP is a new modeling method that mainly differs from other data driven models because it defines an explicit functional relationship between independent variables and desired property by optimizing forms and coefficients of equations simultaneously [23-25]. This is a symbolic regression tools that in compare to traditional regression method such as multiple linear regression there is no force to make about underlying relationship. Accordingly, this work attempt to offer a reliable QSRR model on gas-chromatographic retention time (t_{ry}) of some pesticides which were found in drinking water. Our goal was to create a simple, accurate and interpretable equation as QSRR model. Hence, MLR and GP methods were employed for developing linear and symbolic regression equations, respectively and compare prediction ability of them.

MATERIALS AND METHODS

Data set

The experimental values of gas chromatographic retention time for 74 pesticides were found from reference [26]. These compounds including some pesticides were analyzed in drinking water by gas chromatography analysis with mass spectrometry (GC–MS) (DB-5 capillary column (30 m ×0.25 mm ×0.25 m; Agilent, USA)). The chemical name of database in the present work and their experimental values of t_R are listed in **Table 1**. The values of retention were varied from 0 to 50.36 minutes for structure 1 (*Teflubenzuron*) and 74

(*Deltamethrin*), respectively. The chemicals in the data set were divided into the training and test sets by Y-ranking method. In this way, the data were sorted according to their t_R values and then the training (60 compounds) and test (14 compounds) sets were chosen from the sorted lists with desired distances of each other. Training set was employed to model development and test set was used to evaluate the predictability of obtained model.

Table 1. The data set and corresponding experimental, MLR and GP predicted values of retention times (t_R) for studied pesticides

NO.	Compound name	t_R (min) Experimental	t_R (min) MLR	t_R (min) GP
1	Teflubenzuron	0	7.78	0.32
2	Diphenylamine	12.23	7.33	11.76
3[a]	Phorate	13.73	17.90	19.50
4	Thiometon	14.26	19.53	18.33
5	Dimethoate	14.65	16.79	15.95
6	Beta-HCH	15.56	14.78	15.51
7	Lindane HCH	15.74	14.50	14.90
8[a]	Quintozene	15.98	14.71	13.49
9	Diazinon	16.39	18.39	17.19
10	Disulfoton	16.75	19.45	20.11
11	Delta-HCH	17.13	14.78	15.51
12	Chlorothalonil	17.32	16.76	14.47
13[a]	Pirimicarb	17.93	21.09	16.93
14	Chlorpyrifos-methyl	19.31	20.29	19.89
15	Carbaryl	19.7	22.33	18.90
16	Alachlor	19.77	22.72	23.19
17	Metalaxyl	20.10	19.34	19.04
18[a]	Fenitrothion	21.10	21.03	23.10
19	Pirimiphos methyl	21.17	21.78	23.22
20	Dichlofluanid	21.62	19.29	23.80
21	Malathion	21.77	20.11	22.08
22	Aldrin-R	22.04	25.60	22.83
23[a]	Fenthion	22.39	25.37	25.03
24	Chlorpyrifos	22.52	21.05	25.58
25	Dicofol	22.73	30.29	30.17
26	Triadimefon	22.77	27.49	27.26
27	Cyprodinil	24.35	25.20	25.28
28[a]	Heptachlor-epoxide (Cis)	24.83	24.53	24.56
29	Penconazole	24.93	29.69	27.31
30	Heptachlor-epoxide(Trans)	25.15	23.30	23.51
31	Captan	25.44	22.12	21.56
32	Triadimenol	25.79	28.02	27.68
33[a]	Fipronil	25.8	26.52	25.18
34	Methidathion	26.71	25.64	26.86
35	o,p-DDE	26.99	29.84	30.53
36	Endosulfan-alpha	27.47	27.81	28.56

Table 1 (continued). The data set and corresponding experimental, MLR and GP predicted values of retention times (t_R) for studied pesticides

NO.	Compound name	t_R (min) Experimental	t_R (min) MLR	t_R (min) GP
37	Butachlor	27.82	29.82	31.19
38[a]	Fenamiphos	28.47	25.76	26.90
39	Imazalil	28.93	30.64	27.48
40	Profenofos	29.16	28.26	29.47
41	p,p-DDE	29.42	30.07	30.68
42	Carboxin	29.86	27.72	29.18
43[a]	Oxadiazon	29.86	26.49	29.34
44	o,p-DDD	30.09	29.14	30.27
45	Buprofezin	30.24	33.76	32.19
46	Endosulfan-beta	31.88	27.81	28.56
47	p,p-DDD	32.75	28.55	29.35
48[a]	o,p-DDT	32.76	30.27	33.10
49	Ethion	33.21	31.384	30.32
50	Triazophos	34.31	30.92	33.01
51	Benalaxyl	35.02	35.71	36.42
52	Edifenphos	35.14	35.47	34.73
53[a]	Propiconazole I	35.45	38.04	36.55
54	Fenhexamid	35.62	27.13	35.16
55	Propiconazole II	35.94	38.05	36.57
56	Tebuconazole	36.77	33.66	31.45
57	Iprodione	38.81	32.19	31.25
58[a]	Phosmet	39.05	35.98	34.20
59	Bifenthrin	39.57	38.69	39.19
60	Methoxychlor	39.73	38.64	39.11
61	Fenpropathrin	39.94	39.41	37.95
62	Azinphos-methyl	41.39	35.54	37.90
63[a]	Phosalone	41.41	37.00	37.44
64	Amitraz	42.29	41.26	45.76
65	Landa Cyhalothrin	42.68	42.13	44.13
66	Fenarimol	42.85	42.48	41.67
67	Bitertanol	44.34	43.82	43.52
68[a]	Permethrin I	44.56	48.93	47.65
69	Permethrin II	44.89	49.10	47.84
70	Prochloraz	45.26	47.85	44.40
71	Fenbuconazole	45.98	45.65	44.45
72	Cypermethrin-alpha	47.00	44.58	44.49
73	Esfenvalerate	48.60	51.73	52.09
74	Deltamethrin	50.36	49.43	49.10

Diversity analysis

Diversity analysis was performed to make sure the structures in the training and test sets are representative of both data set [27, 28]. Distance score between two different

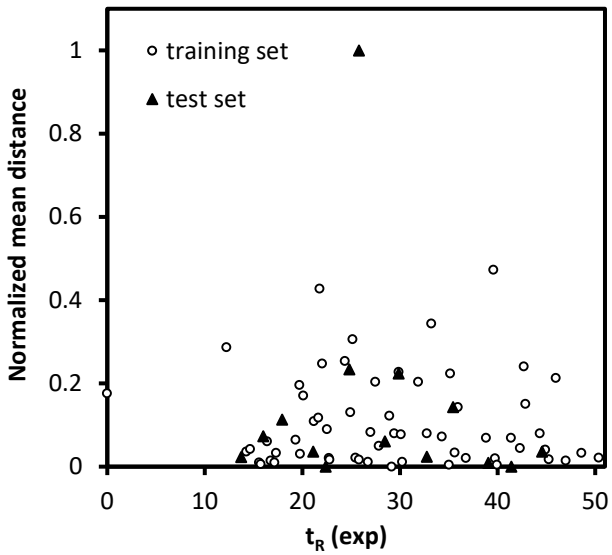

Figure 1. The results of diversity analysis

compounds of p_i and q_j (d_{ij}) can be measured by euclidean distance norm in variable space from following equation (1):

$$d_{ij} = \|p_i - q_j\| = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2} \quad (1)$$

In above equation, k is the number of variable in descriptor matrix X, and m is the number of compounds. The x_{ik} and x_{jk} parameters are the k^{th} descriptors of i and j compounds, respectively. In the following the mean distance of one molecule to others ($\overline{d_i}$) was calculated as follow:

$$\overline{d_i} = \frac{\sum_{j=1}^{m} d_{ij}}{m-1} \quad i = 1,2,\dots,m \quad (2)$$

Then the mean distances among molecules in descriptor space were normalized within the range of 0-1 and plotted against the values of the t_R (**Figure 1**). Inspection to this figure illuminates that the structures of molecules are diverse in training and test sets and represent of the whole data set.

Descriptors calculation and selection

The purpose of QSRR model is to quantitatively correlate the structural variation of studied chemicals to their retention time by applying theoretical molecular descriptors. In the first step of QSRR modeling the structures of these chemicals were drawn and optimized by semiemperical AM1 methods using Hyperchem program (version 7) [29]. Then, molecular descriptors for each molecule were calculated by PaDEL (version 2.11) [30], Codessa (version

2.0) [31] and Dragon (version. 3.0) [32] softwares. Then all calculated descriptors were screened for detecting constant or near constant descriptors to removing them. Thought, some generated descriptors for each molecule, encoded similar information, therefore, it was desirable to eliminate those that show high correlation (R > 0.90) with each other. Between these two descriptors, the one that was lower correlation with the desired properties was removed. At the end of this step stepwise multiple linear regression (SW-MLR) was used to select of the most relevant descriptors from remaining ones. At this step seven descriptors were selected from 992 remaining descriptors, which their name and definitions were shown in **Table 2**. These descriptors were used for developing linear and symbolic regression equations, respectively by MLR and GP method for QSRR modeling.

The variation inflation factors (VIF) were calculated to check any multi-collinearity among the seven selected descriptors by following equation:

$$VIF = \frac{1}{1 - R^2} \tag{3}$$

In the above equation, R is correlation coefficient of multiple regression between each descriptor and the other descriptors in the predictive model. If VIF value is being equal to 1, it indicates that does not exit intercorrelation for each descriptor in model; if it falls between 1.0 and 5.0, it reflects that the model is acceptable, and if VIF becomes larger than 10.0, the model is unstable due to high colliniearity among selected descriptors and should be recheck [33]. The calculated values of VIF among selected descriptors are shown in **Table 2**, which indicate that these descriptors are independent.

Genetic programming (GP)

Conventional regression techniques optimize the coefficients for a pre-specified form of the model. According to this, prediction ability of obtained polynomial models (based on MLR) often is limited. To overcome this drawback, symbolic regression via genetic programming (GP) recently gained as new model fitting tools [24]. Symbolic regression attempts to uncover the intrinsic relationships of the data set and optimize both form and coefficients of model with searching the space of mathematical expressions.

GP basically is a symbolic regression modeling tool with the capability of generating mathematical equation which developed by *Koza* [25]. This method is based on principle of Darwinian's theory and done by evolutionary algorithms [23]. In GP the individuals (mathematical functions) represented as a tree structure with nodes and terminals that evolutionary process is working over them. The nodes correspond to various mathematical operators with adjustable weights. The weight defines probability of operator for choosing in function (greater probability with increasing weight) while the terminals of branches are typically variables (descriptors) or constants values. For evolutionary process firstly random initial population of individual (mathematical equation) were generated by defining population size and choosing selection method [25, 34]. Several kinds of selection methods can

Table 2. Specification of multiple linear regression model

No.	Descriptor name	Notation	Coefficient	Se	VIF
		Constant	0.58	0.00	-
1	solvation connectivity index of order 1	X1sol	0.22	±0.01	1.64
2	Balaban centric index	BAC	-0.07	±0.01	1.48
3	2D-autocorrelation of lag 2 weighted by ionization potential	AATSC2i	0.09	±0.01	2.32
4	Geary autocorrelation of lag 5 weighted by intrinsic state	GATS5s	0.03	±0.01	1.41
5	Maximum number of hydrogen atom	hmax	-0.08	±0.01	2.42
6	Radial Distribution Function - 040 / weighted by atomic mass	RDF040 m	-0.04	±0.00	1.22
7	Broto-Moreau autocorrelation of lag 8 weighted by gasteiger charge	ATSC8c	0.03	±0.01	1.24

be used for choosing an individual from population. Skrgic selection (SS) is one of these methods that based on probability graphs of selection [35]. In this method individual with maximum fitness in population most probably selected for later breeding. In the next step, new generations of equation created by mutation, crossover and reproduction. Then the fitness of each equation is evaluated using the fitness function. These steps are repeated until a desired function is achieved then the obtained model was evaluated by different methods [25, 34]. Simultaneous optimization of forms and coefficients of equations increase predictive ability of the model. Also, in comparison to MLR there is no force to make linear or nonlinear relationship between independent variables and retention time [24].

RESULT AND DISCUSSION

Modeling

The present study investigates the use of MLR and GP for developing QSRR model as mathematical regression equation to predict the gas chromatographic retention times of some pesticides. Multiple linear regression is one of the earliest and commonest methods for generation linear equation as QSRR model [36, 37]. MLR model is a mathematical equation which quantitatively related the selected descriptors as independent variables and studied retention times. The equation and statistical parameters of developed seven descriptors by MLR are shown in **Table 2**.

The selected descriptors were also used to develop GP model. As noted earlier, GP is a symbolic regression tools that capable to generate an interpretable mathematical equation as best fitted model. The GPdotNET (version 3.0) [38] software was used for genetically developing GP model. In GP model, evolutionary process was employed for optimization of functional form and the coefficients for predicted GP model by choosing suitable values of various control parameters (**Table 3**). The quality of the predictive equation can be improved by changing values of these parameters.

Table 3. The optimal values of control parameters for GP model

Control parameters	Optimal values
Function set	+ , −, * , / , (1/x)
Weights of function set	3, 3, 2, 2, 1
Population size	2000
Selection method	Skrgic selection (SS)
Crossover	0.9
Mutation	0.4
Fitness function	RMSE

Table 4. The statistical parameters for MLR and GP models

Model	R^2 training	R^2 test	RMSE training	RMSE test
MLR	0.915	0.898	.313	2.95
GP	0.943	0.911	2.56	2.77

The fitness of each solution (mathematical equation) is evaluated and monitored using RMSE (root mean square error) as a fitness function. The best generated GP equation was shown in the following equation (4):

$$t_R(\min) = \left(\frac{\left(\left(4.37 - \left(\frac{BAC - ATSC8c}{4.37 - GATS5s}\right)\right) - \left(\frac{BAC}{(4.37 + 7.73) + \left(\frac{7.73}{ATS8c}\right)}\right)\right) + \left(X1sol + \left(\frac{X1sol + (AATSC2i + X1sol)}{\left(\frac{BAC}{X1sol}\right) + (4.37 - 0.27)}\right)\right)}{\left(\frac{((RDF040m + hmax) * (4.37 - BAC)) - ((AATSC2i + X1sol) + (X1sol * hmax))}{\left(\frac{X1sol}{GATS5s + RDF040m}\right) + ((hmax + X1sol) + 7.73)}\right) + 7.73} \right) \quad (4)$$

Then predictively and robustness of developed MLR and GP models were evaluated by several validation methods. The resulted statistical parameters of these tests are shown in **Table 4**. By comparison of these parameters it was concluded that GP model was superior over MLR.

Therefore, further investigation was focused on GP model. The GP model was used to predict the retention time of training and test sets. These calculated values were shown in **Table 1**. The plot of the GP predicted versus experimental values of t_R for the training and test sets was shown in **Figure 2**, which indicates good correlation among them ($R^2_{training} = 0.943$ and $R^2_{test} = 0.911$). Moreover, the residuals of these predicted were plotted against the experimental values of the retention time (**Figure 3**). Random propagation of the residuals on both sides of zero line indicates there is no systematic error in developed GP model.

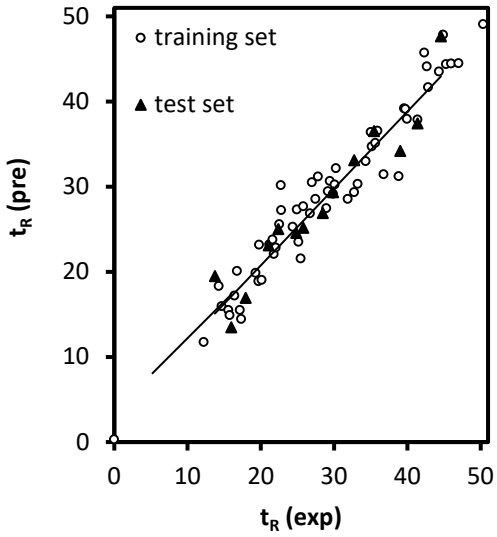

Figure 2. Comparison between predicted and experimental values of retention time by GP model

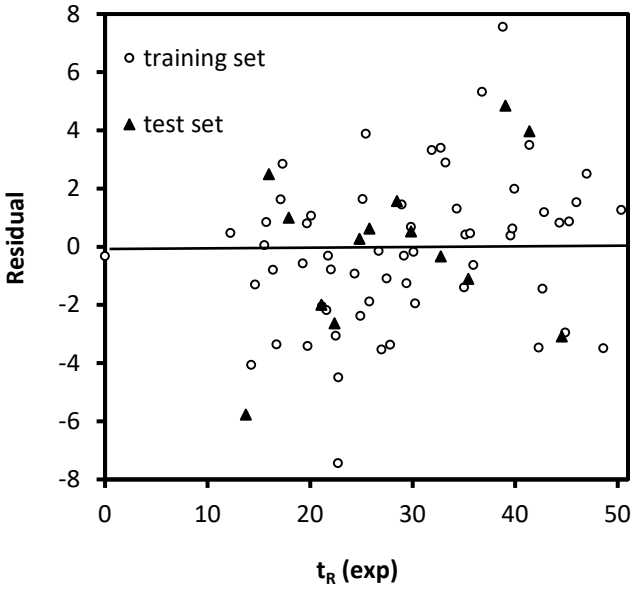

Figure 3. Residuals of predicted *vs.* the experimental values of retention times

Models validation

Validation of the obtained model is a very important step in QSPR (quantitative structure-property relationship) studies. Leave one out (*LOO)* cross validation procedure is one of method that applied for evaluation of a QSPR model. In *LOO* method, the data each of

molecule in the training set is removed and the model was expanded on the remained molecules then the resulting model was employed to predict the intended properties of removing molecule. This procedure is repeated for all molecules of training set and then the cross validated correlation coefficient (Q_{cv}^2) and standard deviation based on predicted residual sum of square ($SPRESS$) calculated by following equations:

$$Q_{cv}^2 = 1 - \frac{\sum(y_i - y_{0i})^2}{\sum(y_{0i} - y_{mean})^2} \quad (5)$$

$$SPRESS = \sqrt{\frac{\sum(y_i - y_{0i})^2}{n - k - 1}} \quad (6)$$

In above equations, y_i, y_{0i} and y_{mean} are the predicted, experimental and mean values of experimental property, respectively; n is the number of compounds in the training set and k is the number of descriptors in the model [39]. The calculated values of Q_{cv}^2 and $SPRESS$ for leave one out cross validation test for MLR model were 0.78 and 2.94 and for GP model were 0.79 and 2.75, respectively, which demonstrate the robustness of these models.

Applicability domain

Applicability domain (AD) analysis was employed to determine model is capable for prediction the property of new compounds with unavailable experimental data that defined as the response and chemical structure spaces which reliable model can be predicted [40]. In this study, leverage approach and William plot were employed to indicate the applicability domain of model. The leverage or hat value is calculated as:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (7)$$

where h_i is the leverage of the compound i in the descriptor space, x_i is the descriptor raw vector, X is matrix of descriptor and superscript T refers to the transpose of the vector and matrix, respectively. The warning leverage h^* is constant at $\frac{3(p+1)}{n}$ that p is the descriptors number in the model and n is the compounds number in training set. The standardized residuals were plotted against the leverage values to display the AD of a model for training and test sets that limited by ±3 times of standardized residuals (outlier of response) and warning h^*, respectively [40]. The molecules with standardized residuals out of this range and or leverage greater than $h^* = 0.40$ are considered as out of AD of models. As can be seen in **Figure 4** one molecule 54 (*Fenhexamid*) from training set are identified as outliers according to its h^* value. The anomalous behavior of this compound could be originated from its natures and molecular structures.

Descriptors interpretation

Sensitivity analysis (SA) was carried out on the GP model to determine the relative importance of descriptors in the QSRR model. In this method, the differences between the RMSE of the complete model and obtained root mean square error when the value of i^{th}

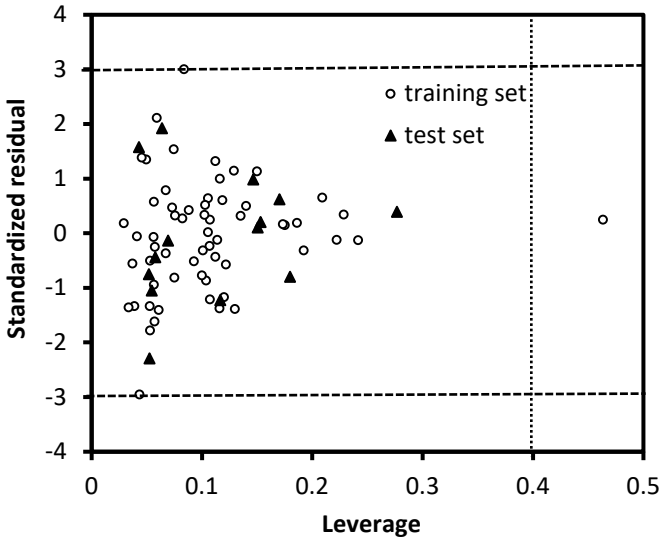

Figure 4. William plot for GP model (h^*=0.4)

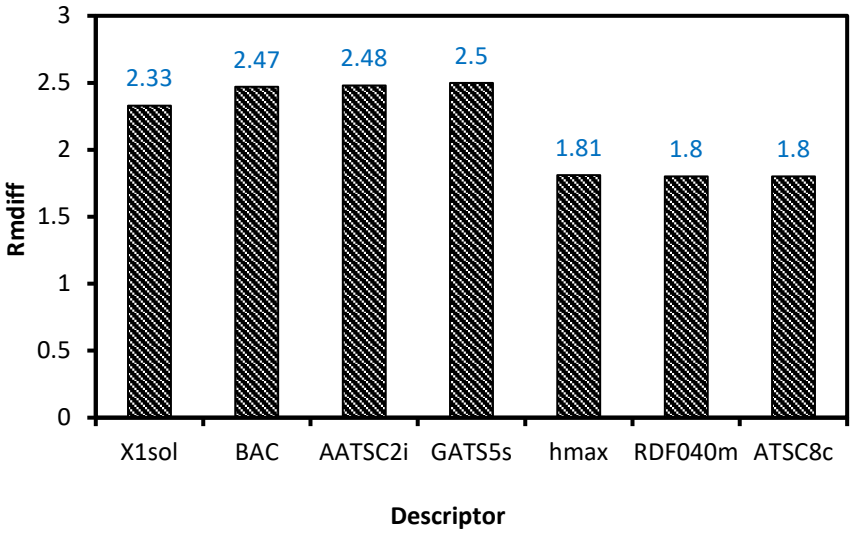

Figure 5. Sensitivity analysis plot for GP model

descriptor was set at zero were calculated (RMSE) and shown as $Rmdiff_i$. Each descriptor which causes greater $Rmdiff_i$ value is more important [41]. The results of sensitivity analysis on the GP model was shown in **Figure 5** which indicated that the importance orders of descriptors are; *GATS5s* ~ *AATSC2i* ~ *BAC* > *X1sol* > *hmax* > *RDF040m* ~ *ATSC8c*.

Among selected descriptors, *GATS5s, AATSC2i* and *BAC* are the most important descriptors that effects on the chromatographic retention time of studied pesticides. *GATS5s and AATSC2i* belonged to *2D autocorrelations* type descriptors which calculated between atoms separated by 5 and 2 chemical bonds adjusted for the intrinsic state and ionization potential,

respectively [30, 42, 43]. *Balaban centric index* (*BAC*) is *centric index* type descriptor that reflect the topology of molecules in tree structure which view from the center and measure molecular branching [44, 45]. The next descriptor appearing in the model is *X1sol*. This descriptor is a *connectivity indices type* which counts solvation connectivity index of order 1 [45]. *hmax* is the next descriptor which belong to *Atom type electrotopological state* descriptor and defined as maximum number of hydrogen atom [30]. *RDF040m* and *ATSC8c* are the last descriptors appearing in the model with the same effects on the retention time. *RDF040m* is one of the *burden-CAS-university of Texas* (*BCUT*) descriptors that submitted as signal *40*, weighted by atomic masses [45, 46] while *ATSC8c* belonged to *2D autocorrelations* type descriptors which *Broto-Moreau* autocorrelation of lag 8 weighted by gasteiger charge [30, 43]. Inspection to these descriptors indicates that the topological structure and polarity of pesticides play significant role on their gas-chromatographic retention times.

CONCLUSION

The present study investigates the use of MLR and GP methods for developing linear and symbolic equations as QSRR models for prediction of gas chromatographic retention time of some pesticides. Comparison of statistical parameters of developed models indicated that symbolic regression equation via GP can be selected as best fitted model. The superiority of GP model indicates that there is some nonlinear relationship between selected descriptors and retention time of these chemicals. The sensitivity analysis of GP equation indicated that the structural features and polarity of studied pesticides are important factors responsible for the GC retention of these chemicals on DB-5 capillary column.

REFERENCES

1. Cooper, J., & Dobson, H. (2007). The benefits of pesticides to mankind and the environment. *Crop Protection, 26*(9), 1337-1348.
2. Oerke, E.-C., & Dehne, H.-W. (2004). Safeguarding production—losses in major crops and the role of crop protection. *Crop protection, 23*(4), 275-285.
3. Van der Hoff, G. R., & van Zoonen, P. (1999). Trace analysis of pesticides by gas chromatography. *Journal of Chromatography A, 843*(1), 301-322.
4. Ahmed, F. E. (2001). Analyses of pesticides and their metabolites in foods and drinks. *TrAC Trends in Analytical Chemistry, 20*(11), 649-661.
5. Davis, J. R., Brownson, R. C., & Garcia, R. (1992). Family pesticide use in the home, garden, orchard, and yard. *Archives of environmental contamination and toxicology, 22*(3), 260-266.
6. Jaga, K., & Dharmani, C. (2003). Sources of exposure to and public health implications of organophosphate pesticides. *Revista panamericana de salud pública, 14*(3), 171-185.
7. Wang, N., et al. (2012). Simultaneous determination of pesticides, polycyclic aromatic hydrocarbons, polychlorinated biphenyls and phthalate esters in human adipose tissue by gas chromatography–tandem mass spectrometry. *Journal of Chromatography B, 898*, 38-52.

8. Boxall, R. (2001). Post-harvest losses to insects—a world overview. *International Biodeterioration & Biodegradation, 48*(1), 137-152.
9. Watts, C. et al. (1989). Pesticides: analytical requirements for compliance with EEC directives. *Water Pollut. Res*, 16-34.
10. Pereira, J. L. et al. (2009). Toxicity evaluation of three pesticides on non-target aquatic and soil organisms: commercial formulation versus active ingredient. *Ecotoxicology, 18*(4), 455-463.
11. Hallberg, G. R. (1989). Pesticides pollution of groundwater in the humid United States. *Agriculture, ecosystems & environment, 2*6(3), 299-367.
12. Leistra, M., & Boesten, J. (1989). Pesticide contamination of groundwater in western Europe. *Agriculture, ecosystems & environment, 26*(3), 369-389.
13. Klaine, S. et al. (1988). Characterization of agricultural nonpoint pollution: Pesticide migration in a west Tennessee watershed. *Environmental toxicology and chemistry, 7*(8), 609-614.
14. Aharonson, N. (1987). Potential contamination of ground water by pesticides. *Pure and applied chemistry, 59*(10), 1419-1446.
15. Gilliom, R. J. et al. (2006). Pesticides in the nation's streams and ground water, 1992-2001. *Geological Survey (US)*.
16. Rathore, H. S., & Nollet, L. M. (2012). Pesticides: Evaluation of environmental pollution: CRC Press.
17. Ali, U. et al. (2014). Organochlorine pesticides (OCPs) in South Asian region: a review. *Science of the Total Environment, 476*, 705-717.
18. Junk, G., & Richard, J. (1988). Organics in water: solid phase extraction on a small scale. *Analytical Chemistry, 60*(5), 451-454.
19. Zaugg, S. D. et al. (1995). Methods of analysis by the US Geological Survey National Water Quality Laboratory; determination of pesticides in water by C-18 solid-phase extraction and capillary-column gas chromatography/mass spectrometry with selected-ion monitoring, US Geological Survey: Open-File Reports Section/ESIC [distributor].
20. Kaliszan, R. (1987). Quantitative structure-chromatographic retention relationships.
21. Kaliszan, R. (2007). QSRR: quantitative structure-(chromatographic) retention relationships. *Chemical reviews, 107*(7), 3212-3246.
22. Dearden, J., Cronin, M., & Kaiser, K. (2009). How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). SAR and QSAR in Environmental *Research, 20*(3-4), 241-266.
23. Koza, J. R. (1990). Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems. Stanford University, Department of Computer Science.
24. Koza, J. R. (1992). Genetic programming: on the programming of computers by means of natural selection. Vol. 1. MIT press.
25. Koza, J. R. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and Computing, 4*(2), 87-112.

26. Ghasemi, S., Karami, H., & Khanezar, H. (2014). Hydrothermal synthesis of lead dioxide/multiwall carbon nanotube nanocomposite and its application in removal of some organic water pollutants. *Journal of Materials Science,* 49(3), 1014-1024.
27. Rouvray, D. H. (1992). Definition and role of similarity concepts in the chemical and physical sciences. *Journal of Chemical Information and Computer Sciences, 32*(6), 580-586.
28. Maldonado, A. G. et al. (2006). Molecular similarity and diversity in chemoinformatics: from theory to applications. *Molecular diversity,* 10(1), 39-79.
29. HyperChem, H. (2002). Release 7 for windows, HyperCube, Ed.
30. Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry, 3*2(7), 1466-1474.
31. Katritzky, A., V. Lobanov, & Karelson, M. (1995). CODESSA: training manual. University of Florida, Gainesville, FL.
32. Todeschini, R. et al. (2003). DRAGON-Software for the calculation of molecular descriptors. Web version, 3.
33. Stine, R. A., (1995). Graphical interpretation of variance inflation factors. The American Statistician, 49(1), 53-56.
34. Koza, J. R. (1990). Concept formation and decision tree induction using the genetic programming paradigm. in International Conference on Parallel Problem Solving from Nature. Springer.
35. Hrnjica, B. (2011). Skrgic Selection in GPdotNET.
36. Kramer, R. (1998). Chemometric techniques for quantitative analysis. CRC Press.
37. Vandeginste, B. G. et al. (1998). Handbook of chemometrics and qualimetrics. Elsevier.
38. Hrnjica, B. (2013). GPdotNET - artificial intelligence tool.
39. Wold, S., Eriksson, L., & Clementi, S. (1995). Statistical validation of QSAR results. *Chemometric methods in molecular design*, 309-338.
40. Atkinson, A. C. (1985). Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis. [519.536 A875].
41. Miller, D. (1974). Model validation through sensitivity analysis. in Proceedings of the 1974 Summer Computer Simulation Conference.
42. Geary, R. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician, 5*, 115–146. *Detection of Disease Clustering, 389*.
43. Moreau, G., & Broto, P. (1980). The auto-correlation of a topological-structure-a new Molecular Descriptor, Gauthier-Villars 120 Blvd Saint-Germain, 75280 Paris Cedex 06, France, p. 359-360.
44. Balaban, A. (1979). Chemical Graphs, 34. 5 New Topological Indexes for the Barcnhing of Tree-Like Graphs. *Theoretica Chimica Acta, 53*(4), 355-375.
45. Marleau, G., Hébert, A., & Roy, R. (2000). A user guide for DRAGON. Version DRAGON_000331 release 3.04. Report IGE-174 Rev, 5.
46. Stanton, D. T. (1999). Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *Journal of chemical information and computer sciences, 3*9(1), 11-20.