International Journal of Advances in Applied Sciences (IJAAS)
Vol. 1, No. 2, June 2012, pp. 91~100
ISSN: 2252-8814

Prediction of Retention Index Based on Quantitative Structure-Retention Relationship

Mehdi Rahimi*, Hossein Farahbakhsh*, Nasrin Salehi, Mehdi Nekoei***
* Department of Chemistry, Faculty of Basic Sciences, Shahrood Branch, Islamic Azad University, Shahrood, Iran
**Department of Basic Sciences, Shahrood Branch, Islamic Azad University, Shahrood, Iran

Article Info

Article history:

Received Mar 23, 2012
Revised May 27, 2012
Accepted Jun 7, 2012

Keyword:

Zosimia absinthifolia
Essential oil
Retention indices
QSRR
MLR

ABSTRACT

In this paper, a quantitative structure-retention relationship (QSRR) model have developed for prediction of retention indices (RI) essential oils. First, the structure of the understudy molecules was drawn, using HyperChem software. Then the molecular descriptors which cover different information of molecular structures, were calculated by Dragon software. Subsequently, a suitable set of molecular descriptors that fulfill the best fitted models were selected using stepwise-multiple linear regression (SW-MLR) method. A simple model with low standard errors and high correlation coefficients was selected. The accuracy of the suggested model is illustrated using cross-validation, validation through an external test set and Y-randomization. The results illustrated that the linear techniques such as MLR combined with a successful variable selection procedure are capable to generate an efficient QSRR model for predicting the retention indices of different compounds. This model, with high statistical significance ($R^2_{calibration}$=0.99, $R^2_{prediction}$=0.981, Q^2_{LOO}=0.988, Q^2_{LGO}=0.984, REP(%)=3.827), could be used adequately for the prediction and description of the retention indices of other essential oil compounds.

Corresponding Author:

Nasrin Salehi,
Department of Basic Sciences,
Shahrood Branch, Islamic Azad University, Iran
Email: nssalehi@yahoo.com

1. INTRODUCTION

Essential oils, a new approach to prevent the proliferation of microorganism or protection of food from oxidation, are ubiquitously used as antibacterial [1]-[3], antifungal [3], [4], antioxidant [5] and made them useful as natural additives in the food industry. They are also used to control human diseases of microbial origin and to cure such diseases as atherosclerosis and cancer [6]. These essential oils have been used in the folk medicine for thousands of years as antimicrobial [7], [8]. Therefore, the assessment of gas chromatographic (GC) retention index (RI) of essential oils is a matter of great importance in the health of human being.

Zosimia absinthifolia is a herb that belongs to umbelliferae family and grows wild in iran. The plant materials were collected from Alshtar in North of Lorestan province at three stages including before flowering, full flowering and fruiting stages and subjected to hydrodistillation using a Cleavenger type apparatus for 3h [9].

GC and GC–MS are the main methods for the identification of these plant oils. Seeking quantitative relationship between the molecular structure and the gas chromatographic retention indices has been a basic task in chemistry. Correlations between the GC retention indices and the molecular structures can provide more profound insights into the interactions between the eluents and the stationary phases from a theoretical viewpoint. In addition, they can provide very important information about the effect of the chemical structures on the retention behavior and the possible mechanism of absorption and elution.

Quantitative structure–retention relationships (QSRR) represent statistical models, which quantify the relation between the structure of the molecule and the chromatographic retention indices of the compound, allowing the prediction of the retention indices of the novel compounds. QSRR on the retention indices have been reported for different types of organic compounds [10]-[14].

The application of these techniques usually requires variable selection for building well-fitted models. In this work, we employed the elimination selection-stepwise regression (ES-SWR) variable selection method. The result of this study was the development of a newlinear QSRR model containing 4 variables. The proposed methodology was validated using several strategies: cross-validation and external validation using division of the entire data set into training and test sets.

The aim of this work is to search for an efficient method to build an accurate quantitative relationship between the molecular structure and the retention indices of the *Zosimia absinthifolia* essential oils by SW-MLR.

2. RESEARCH METHOD

2.1. Computer hardware and software

A Pentium IV personal computer (CPU at 3.06 GHz; ISIRAN Co., Tehran, Iran) with the Windows XP operating system was used. The geometry optimization was performed with HyperChem (Version 8.0 Hypercube Inc., Alberta, Canada). For the calculation of the molecular descriptors, the Dragon 2.1 software (Milano Chemometrics and QSAR Research Group, Milano, Italy) was used. The SPSS software (Version 14.0; SPSS Inc., Chicago, IL, USA) was employed for the simple MLR analysis. The other calculations were performed in the MATLAB (Version 7.0, Math works Inc., Natick, MA, USA).

2.2. Data set

The data set of the GC retention indices was taken from the values reported by Javidnia *et al* [15]. The data set was split into a training set and a test set. The test set of 15 compounds was selected randomly from the original 61 of essential oil components with the remaining compounds constituting the training set. The training set of 46 compounds, with RI values in the rang of 851-2500, was used to adjust the parameters of the model, and the test set of 15 compounds, with RI in the range of 886-2493, was used to evaluate its predictive ability.

2.2. Determination of molecular descriptors

The retention index in GC depends on the relative solubility of the solute in the mobile and stationary phases, which depend on the molecular structure and chemical properties of the solute. Differences between these properties govern retention behavior through the column. Molecular descriptors are defined as numerical characteristics associated with chemical structures. The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number applied to correlate physical properties.

The Dragon software was used to calculate the descriptors in this research and a total of 1481 molecular descriptors, from 18 different types of theoretical descriptor, were calculated for each molecule. Since the values of many descriptors are related to the bonds length and bonds angles etc., the chemical structure of every molecule must be optimized before calculating its molecular descriptors. For this reason, chemical structure of the 61 studied molecules were drawn with the Hyperchem software and saved with the HIN extension. To optimize the geometry of these molecules, the AM1 geometrical optimization was applied. After optimizing the chemical structures of all compounds, the molecular descriptors were calculated using Dragon. A wide variety of descriptors have been reported in the literature, having been used in the QSRR and QSAR analysis [16]-[21].

3. RESULTS AND ANALYSIS

After selection of the most important descriptors by stepwise, MLR was performed to build the linear model. To investigate the optimum number of descriptors to be used in a model for modeling RI, we have plotted a graph between numbers of descriptors against statistical parameters (R^2). Figure 1 shows the plot of R^2 as a function of the number of descriptors for the 1–10 parameter models.

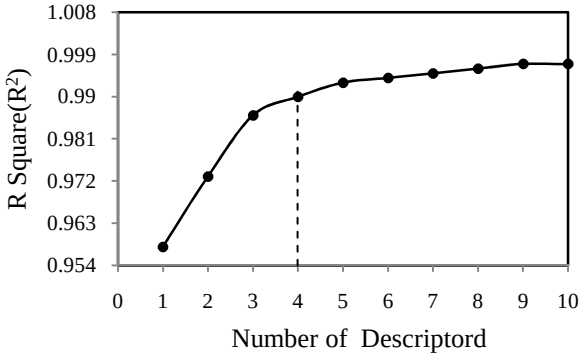

Figure 1. Influence of number of descriptors on R^2 of MLR model.

R^2 increased by increasing the number of descriptors. A perusal of Figure. 1 indicates that R^2 remain almost parallel to the X-axis (number of descriptors) after four parameters and higher order models. So we used the best correlation equation with four descriptors for the analysis. Good correlations with the experimental RI data were selected based on the squared correlation coefficient (R^2), Fisher criterion (F), and root mean square error (RMSE) of the regression. The following equation obtained by MLR method:

RI = 307.384(±31.570)
+ 30.558(±0.559) XMOD
+ 13.170(±1.557) PCD
- 460.439(±57.699) MATS2e
- 68.595(±17.076) GATS2e

From the above equation, it can be concluded that themost significant descriptors according to the SW-MLR algorithm are Modified Randic chi-1 index (XMOD), Difference of multiple path counts to path counts (PCD), Moran autocorrelation-lag2/weighted by atomic Sanderson electronegativities (MATS2e) and Geary autocorrelation-lag2/weighted by atomic Sanderson electronegativities (GATS2e). Table 1 presents the correlation matrix, where it is clear that the four selected descriptors are almost completely uncorrelated.

Table 1. Correlation matrix for the four selected descriptors.

	XMOD	PCD	MATS2e	GATS2e
XMOD	1			
PCD	0.240	1		
MATS2e	0.034	0.097	1	
GATS2e	0.264	0.055	-0.555	1

A brief explanation of the descriptors that were selected is as follows:
The first descriptor is XMOD, which is one of the topological descriptors. Its effect on the retention index was positive, which indicates that the retention index is directly related to this descriptor. The second descriptor of this model was difference of multiple path counts to path counts (PCD). It is one of the topological descriptors. Its effect on the retention index was positive. Another descriptors of this model was MATS2e and GATS2e that had a negative effect on the retention index. Which are 2D- Autocorrelations descriptors. A detailed description of the linear model based on compounds in the training set is summarized in Table 2.

Table 2. Selected descriptors of multiple linear regression.

No.	Symbols	Descriptor description	Group descriptor	Coefficient	MF	VIF
	Constant			307.384(±31.570)		
1	XMOD	Modified Randic chi-1 index	Topological	30.558(±0.559)	1.127	1.263
2	PCD	Difference of multiple path counts to path counts	Topological	13.170(±1.557)	0.022	1.101
3	MATS2e	Moran autocorrelation - lag2 / weighted by atomic Sanderson electronegativities	2D - Autocorrelations	-460.439(±57.699)	-0.038	1.519
4	GATS2e	Geary autocorrelation - lag 2 / weighted by atomic Sandersonelectronegativities	2D - Autocorrelations	-68.595(±17.076)	-0.111	1.609

The multi-collinearity between the above four descriptors were detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

$$VIF = \frac{1}{1 - r^2}$$

where r is the correlation coefficient of the multiple regression between the variables in the model. If VIF equals to 1, then no inter-correlation exists for each variable; if VIF falls into the range of 1–5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary. The corresponding VIF values of the four descriptors are shown in Table 2. As can be seen from this table, most of the variables had VIF values of less than 5, indicating that the obtained model has statistic significance.

To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor. This calculation was performed with the equation below:

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}}$$

Where MF_j represents the mean effect for the considered descriptor j, β_j is the coefficient of the descriptor j, d_{ij} stands for the value of the target descriptors for each molecule and, eventually, m is the descriptors number for the model. The MF value indicates the relative importance of a descriptor, compared with the other descriptors in the model. Its sign indicates the variation direction in the values of the activities as a result of the increase (or reduction) of the descriptor values [22]. The mean effect values are shown in Table 2.

Then the obtained model was used to predict the RI of compounds in the training and test sets. The statistical parameters for the training set were $R^2 = 0.99$ and $F = 981.772$. In addition, with the test set, the prediction results were obtained. The statistical parameters were $R^2 = 0.981$ and $F = 127.635$. The predicted versus experimental value based on SW-MLR was shown in Table 3 and Figure 2. Figure 2 shows the predicted versus experimental RI for all of the 61compounds studied, the training set and the test set.

Table 3. The data set and the corresponding observed and predicted RI values by SW-MLR for the training and test set.

No.	Compound	RI (Exp)	SW-MLR	D[a]	E(%)[b]
Training set					
1	(E)-2-Hexenal	851	886.7	35.7	4.2
2	Heptanal	904	980.7	76.7	8.48
3	α-Pinene	935	964	29	3.1
4	Benzaldehyde	961	1019.2	58.2	6.06
5	Sabinene	975	986.5	11.5	1.18
6	β-Pinene	978	967.6	-10.4	-1.06
7	Myrcene	993	1020.6	27.6	2.78
8	δ-2-Carene	1002	974.9	-27.1	-2.7
9	Octanal	1003	1073.8	70.8	7.06
10	ρ-Cymene	1026	1015	-11	-1.07
11	β-Phellandrene	1030	1023	-7	-0.68
12	(Z)-β-Ocimene	1038	1014	-24	-2.31
13	γ-Terpinene	1059	1007.4	-51.6	-4.87
14	Octanol	1071	1054.5	-16.5	-1.54
15	Linalool	1103	1090.2	-12.8	-1.16
16	trans-ρ-Menth-2-en-1-ol	1143	1115.6	-27.4	-2.4
17	(E)-2-Nonenal	1163	1168.3	5.3	0.46
18	Borneol	1170	1132.5	-37.5	-3.21
19	α-Terpineol	1190	1110.5	-79.5	-6.68
20	Octyl acetate	1216	1213.5	-2.5	-0.21
21	Methyl thymol	1238	1259.8	21.8	1.76
22	Piperitone	1256	1176.8	-79.2	-6.31
23	Geranial	1266	1205.8	-60.2	-4.76
24	Bornyl acetate	1289	1333.9	44.9	3.48
25	trans-Pinocarvyl acetate	1301	1341.7	40.7	3.13
26	δ-Elemene	1339	1392	53	3.96
27	Citronellyl acetate	1356	1348.8	-7.2	-0.53
28	Octyl butyrate	1398	1407.6	9.6	0.69
29	β-Caryophyllene	1421	1446.4	25.4	1.79
30	Octyl-2-methylbutyrate	1441	1482.3	41.3	2.87
31	Neryl propionate	1460	1455.1	-4.9	-0.34
32	Germacrene D	1483	1497.5	14.5	0.98
33	Bicyclogermacrene	1498	1477.4	-20.6	-1.38
34	Citronellyl butyrate	1533	1544.2	11.2	0.73
35	Geranyl butyrate	1566	1551.8	-14.2	-0.91
36	Caryophyllene oxide	1584	1555.8	-28.2	-1.78
37	Geranyl-2-methylbutyrate	1605	1629.5	24.5	1.53
38	Citronellylvalerate	1626	1638.3	12.3	0.76
39	Caryophylla-4(14), 8(15)-dien-5β-ol	1645	1575.7	-69.3	-4.21
40	Geranyltiglate	1705	1630.6	-74.4	-4.36
41	(E)-Sesquilavandulyl acetate	1739	1793.2	54.2	3.12
42	(E,E)-Farnesyl acetate	1849	1833.3	-15.7	-0.85
43	Geranyllinalool	2025	2035	10	0.49

[a] experimental RI - predicted RI
[b] Relative error

Table 1. Continued

No.	Compound	RI (Exp)	SW-MLR	D[a]	E(%)[b]
Training set					
44	Osthole	2144	2144.4	0.4	0.02
45	Tricosane	2300	2309.7	9.7	0.42
46	Pentacosane	2500	2492.7	-7.3	-0.29
Test set					
1	Camphene	950	947.3	-2.7	-0.28
2	3-Octanone	984	1021	37	3.76
3	δ-3-Carene	1012	974.6	-37.4	-3.7
4	(E)-β-Ocimene	1049	1014	-35	-3.34
5	cis-ρ-Menth-2-en-1-ol	1125	1115.6	-9.4	-0.84
6	ρ-Cymen-8-ol	1187	1123.3	-63.7	-5.37
7	Neral	1241	1205.8	-35.2	-2.84
8	Lavandulyl acetate	1294	1326	32	2.47
9	Neryl acetate	1365	1351.3	-13.7	-1
10	α-Humulene	1457	1481.5	24.5	1.68
11	Geranylisobutyrate	1517	1526.5	9.5	0.63
12	Octylhexanoate	1591	1594.1	3.1	0.19
13	Geranylvalerate	1658	1648.3	-9.7	-0.59
14	Hexadecanoic acid	1978	1802.3	-175.7	-8.88
15	Tetracosane	2400	2401.2	1.2	0.05

[a] experimental RI - predicted RI
[b] Relative error

The residuals (experimental RI-predicted RI) versus experimental RI value, obtained by the SW-MLR modeling, shown in Figure 3. The distribution of the residuals on both sides of the zero line indicates there is no systematic error in the SW-MLR model.

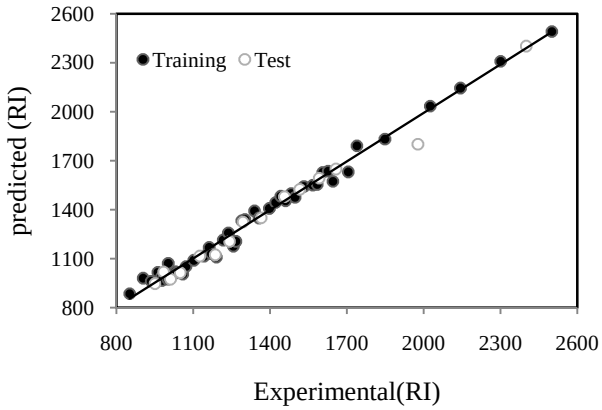

Figure 2. The predicted RI values by the MLR modeling vs. the experimental RI values.

The standardized regression coefficient revealed the significance of individual descriptors, displayed in the regression models in Figure 4.

The results illustrated once more that the linear MLR technique combined with a successful variable selection procedure is adequate to generate an efficient QSRR model for predicting the RI of compounds.

For a more exhaustive testing of the predictive power of the model, validation of the model was also carried out using the LOO and the LGO cross-validation techniques on the training set of compounds. For LOO cross-validation, a data point is removed from the set, and the model is recalculated. The predicted RI for that point is then compared with its actual value. This is repeated until each data point has been omitted once. For LGO, 20% of the data points are removed from the dataset and the model was refitted, the predicted values for those points were then compared with the experimental values. Again, this is repeated until each data point has been omitted once. The results produced by the LOO(Q^2_{LOO} = 0.988) and the LGO (Q^2_{LGO} = 0.984) cross-validation tests illustrated the quality of the obtained model. Figure 5 and Figure 6 show values of the Q^2_{LOO} and Q^2_{LGO}.

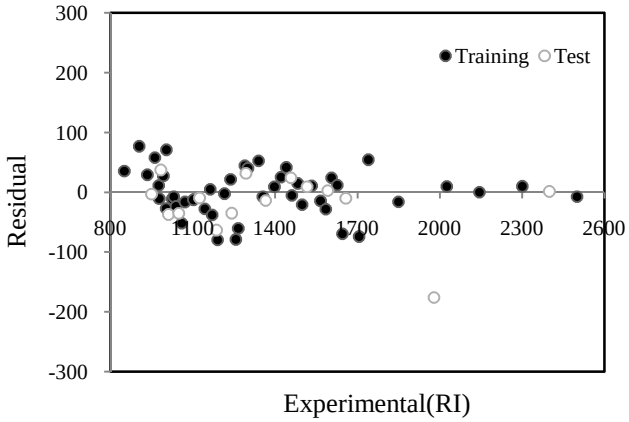

Figure 3. Plot of the residuals against the experimental values of the retention indices.

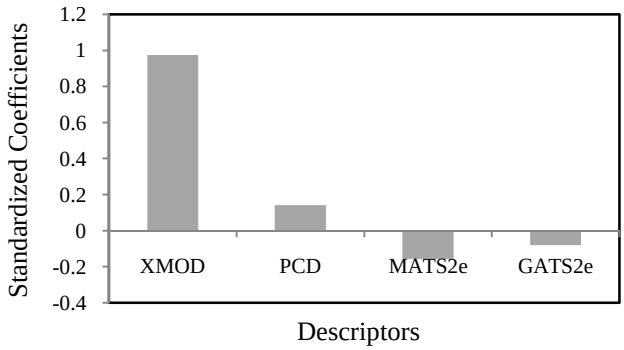

Figure 4. Dependence of standardized regression coefficients on the descriptor used in SW-MLR.

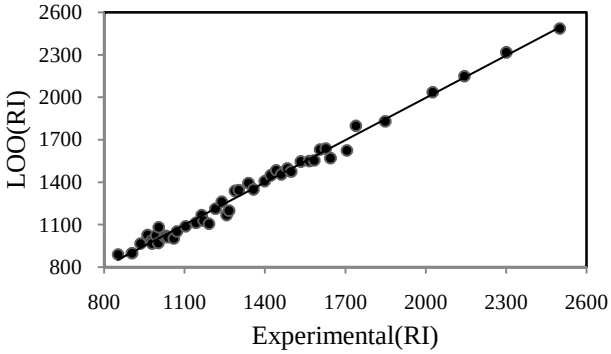

Figure 5. The predicted LOO (RI) values by the cross-validation modeling vs. the experimental RI values.

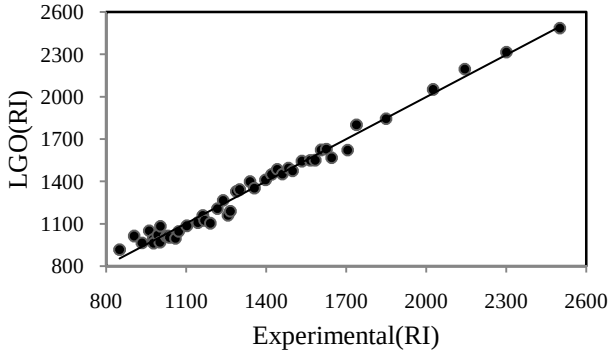

Figure 6. The predicted LGO (RI) values by the cross-validation modeling vs. the experimental RI values.

The model was further validated by applying Y-randomization. Several random shuffles of the Y vector (RI) were performed and the low R^2 and Q^2 values that were obtained showing that the good results in the original model use not due to a chance correlation or structural dependency of the training set. The results of the Y-randomization test are presented in Table 4. The proposed method, due to the high predictive ability and simplicity could be a useful aid to the costly and time consuming experiments for determining the RI of other compound.

Table 4. R^2 and Q^2 values after several Y-randomization tests.

Iteration	R^2	Q^2
1	0.067	0
2	0.092	0.001
3	0.025	0.101
4	0.008	0.086
5	0.066	0.002
6	0.103	0.004
7	0.081	0.012
8	0.029	0.003
9	0.012	0.129
10	0.121	0.001

4. CONCLUSION

In this paper a simple QSRR model was presented for prediction RI of the essential oils. This model is a multivariate linear model, which has four variables (molecular descriptors). These four molecular descriptors were selected using SW–MLR technique. These variables are calculated based on the chemical structure molecules. The validation procedures (cross-validation, separation of data into independent training and validation sets) illustrated the accuracy and robustness of the produced QSRR model not only by calculating its fitness on sets of training data, but also by testing the predictive ability of the model. The QSRR model with simply calculated molecular descriptors could be employed to estimate the retention index for new compounds.

ACKNOWLEDGEMENTS

The authors acknowledge the kind financial support provided by the Research Council ofIslamic Azad University of Shahrood for the preparation of this study.

REFERENCES

[1] S. Burt, "Essential oils: their antibacterial properties and potential applications in foods," *International Journal of Food Microbiology*, vol. 94, pp. 223-253, 2004.

[2] S. T. Chang, P. F. Chen and S. C. Chang, "Antibacterial activity of leaf essential oils and their constituents from Cinnamomumosmophloeum," *Journal of Ethnopharmacology*, vol. 77, pp. 123-127, 2001.

[3] D. Kalemba and A. Kunicka, "Antibacterial and antifungal properties of essential oils," *Current Medicinal Chemistry*, vol. 10, pp. 813-829, 2003.

[4] C. L. Wilson, J. M. Solar, A. El Ghaouth and M. E. Wisniewski, "Rapid evaluation of plant extracts and essential oils for antifungal activity against Botrytis cinerea," *Plant Disease*, vol. 81, pp. 204-210, 1997.

[5] M. Burits and F. Bucar, "Antioxidant activity of Nigella sativa essential oil," *Phytotherapy Research*, vol. 14, pp. 323-328, 2000.

[6] P. H. Warnke, E. Sherry, P. A. J. Russo, Y. Acil, J. Wiltfang, S. Sivananthan, M. Sprengel, J. C. Roldàn, S. Schubert and J. P. Bredee, "Antibacterial essential oils in malodorous cancer patients: Clinical observations in 30 patients," *Phytomedicine*, vol. 13, pp. 463-467, 2006.

[7] K. Fisher and C. Phillips, "Potential antimicrobial uses of essential oils in food: Is citrus the answer?," *Trends in Food Science & Technology*, vol. 19, pp. 156-164, 2008.

[8] T. Yangui, M. Bouaziz, A. Dhouib and S. Sayadi, "Potential use of Tunisian Pituranthos chloranthus essential oils as a natural disinfectant," *Letters in Applied Microbiology*, vol. 48, pp. 112-117, 2009.

[9] H. Amiri, "Quantitative And Qualitative Changes Of Essential Oil Of ZosimiaAbsinthifolia (Vent.) Link. In Different Phenological Stages," *Iranian Journal of Medicinal and Aromatic Plants*, vol. 24, pp. 217-224, 2008.

[10] M. Jalali-Heravi and M. H. Fatemi, "Artificial neural network modeling of Kováts retention indices for noncyclic and monocyclic terpenes," *Journal of Chromatography A*, vol. 915, pp. 177-183, 2001.

[11] Z. Garakani-Nejad, M. Karlovits, W. Demuth, T. Stimpfl, W. Vycudilik, M. Jalali-Heravi and K. Varmuza, "Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds," *Journal of Chromatography A*, vol. 1028, pp. 287-295, 2004.

[12] J. Acevedo-Martinez, J. C. Escalona-Arranz, A. Villar-Rojas, F. Tellez-Palmero, R. Perez-Roses, L. Gonzalez and R. Carrasco-Velar, "Quantitative study of the structure–retention index relationship in the imine family," *Journal of Chromatography A*, vol. 1102, pp. 238-244, 2006.

[13] P. Tulasamma and K. S. Reddy, "Quantitative structure and retention relationships for gas chromatographic data: Application to alkyl pyridines on apolar and polar phases," *Journal of Molecular Graphics and Modelling*, vol. 25, pp. 507-513, 2006.

[14] K. Heberger and T. Kowalska, "Quantitative structure–retention relationships: VI. Thermodynamics of Kováts retention index–boiling point correlations for alkylbenzenes in gas chromatography," *Chemometrics and Intelligent Laboratory Systems*, vol. 47, pp. 205-217, 1999.

[15] K. Javidnia, R. Miri, M. Soltani and A. R. Khosravi, "Constituents of the Oil of Zosimiaabsinthifolia (Vent.) Link. from Iran," *Journal of Essential Oil Research*, vol. 20, pp. 114-116, 2008.

[16] R. Todeschini and V. Consonni, "Handbook of molecular descriptors," *Wiley-VCH*, Weinheim, 2000.

[17] L. B. Kier and L. H. Hall, "Molecular Connectivity in Structure-Activity Analysis," *RSP-Wiley*, Chichetster, UK, 1986.

[18] E. V. Konstantinova, "The Discrimination Ability of Some Topological and Information Distance Indices for Graphs of Unbranched Hexagonal Systems," *Journal of chemical information and computer sciences*, vol. 36, pp. 54-57, 1996.

[19] G. Rucker and C. Rucker, "Counts of all walks as atomic and molecular descriptors," *Journal of chemical information and computer sciences*, vol. 33, pp. 683-695, 1993.

[20] J. Galvez, R. Garcia, M. T. Salabert and R. Soler, "Charge Indexes. New Topological Descriptors," *Journal of chemical information and computer sciences*, vol. 34, pp. 520-525, 1994.

[21] P. Broto, G. Moreau and C. Vandicke, "Molecular structures: Perception, autocorrelation descriptor and SAR studies," *European Journal of Medicinal Chemistry*, vol. 19, pp. 66-70,1984.

[22] M. Adimi, M. Salimi, M. Nekoei, E. Pourbasheer and A. Beheshti, "A quantitative structure–activity relationship study on histamine receptor antagonists using the genetic algorithm–multi-parameter linear regression method," *Journal of the Serbian Chemical Society*, vol.77, pp. 639-650, 2012.