

ORIGINAL ARTICLE

# Prediction of the GC-MS Retention Indices for a Diverse Set of Terpenes as Constituent Components of Camu-camu (*Myrciaria dubia* (HBK) Mc Vaugh) Volatile Oil, Using Particle Swarm Optimization-Multiple Linear Regression (PSO-MLR)

Majid Mohammadhosseini

Department of Chemistry, College of Basic Sciences, Shahrood Branch, Islamic Azad University, Shahrood, Iran

(Received: 1 February 2014 Accepted: 9 April 2014)

## KEYWORDS

Particle swarm optimization-multiple linear regression (PSO-MLR)  
quantitative structure-retention relationship (QSRR)  
retention indices (RIs) prediction  
Essential oil  
Cross validation,  
External validation  
GC  
GC-MS

**ABSTRACT:** A reliable quantitative structure retention relationship (QSRR) study has been evaluated to predict the retention indices (RIs) of a broad spectrum of compounds, namely 118 non-linear, cyclic and heterocyclic terpenoids (both saturated and unsaturated), on an HP-5MS fused silica column. A principal component analysis showed that seven compounds lay outside of the main cluster. After elimination of the outliers, the data set was divided into training and test sets involving 80 and 28 compounds. The method was tested by application of the particle swarm optimization (PSO) method to find the most effective molecular descriptors, followed by multiple linear regressions (MLR). The PSO-MLR model was further confirmed through “leave one out cross validation” (LOO-CV) and “leave group out cross validation” (LGO-CV), as well as external validations. The promising statistical figures of merit associated with the proposed model ( $R^2_{\text{train}}=0.936$ ,  $Q^2_{\text{LOO}}=0.928$ ,  $Q^2_{\text{LGO}}=0.921$ ,  $F=376.4$ ) confirm its high ability to predict RIs with negligible relative errors of predictions (REP train=4.8%, REP test=6.0%).

## INTRODUCTION

Volatile essential oils (VEOs) are concentrated hydrophobic liquids comprising complex mixtures of pungent, low molecular weight organic compounds. VEOs are usually derived from odorous wild-growing plants by traditional and advanced

instrumental techniques such as hydro-distillation (HD), solvent extraction (SE), florasol extraction (FE), head space-solid phase microextraction (HS-SPME) as well as super-critical fluid extraction [1, 2]. Characterization of the chemical profiles found in the analysis of the oils has been extensively investigated because of the commercial benefits of VEOs in the flavor and

\* Corresponding author: majidmohammadhosseini@yahoo.com (M. Mohammadhosseini)

fragrance industry. VEOs also are important in aromatherapy (relaxant) and in pharmaceutical preparations, particularly for their therapeutic effects as sedatives, spasmolytics, antioxidants, antiviral and antibacterial agents [3]. *Camu-camu* (*Myrciaria dubia* (HBK) McVaugh) is a low-growing wild shrub found throughout the Amazon rainforest, mainly in swampy or flooded areas. It produces round, light orange colored fruits about the size of lemons, which contain a significant amount of vitamin C [4].

Undoubtedly, one of the most straightforward ways to determine the chemical profiles of VEOs is comparison of the Kovatz retention index of an unknown or target compound with reliable and authentic samples given in the literature. During the recent decades, this term has gained a growing interest due its high potential in identification of a wide spectrum of organic and natural products. The term RI is defined as the most popular dependent variable in quantitative structure–retention relationship (QSRR) studies because of its excellent reproducibility and accuracy. Relative retention times (RRTs) are also frequently used. In some cases, response factors are also predicted from molecular structure [5]. The RI criterion is independent of the chromatographic column conditions and/or problems caused, during the injection of volatile and thermally stable compounds, within the stationary phase pores, such as overloading, bleeding and trapping of the solutes.

There is a growing interest in the literature for prediction of diverse physicochemical properties of versatile series of organic compounds based upon quantitative structure-property and/or structure-activity relationships (QSPR or QSAR) [6-14]. Among the subsets of QSPR, QSRR is one of the most popular approaches. QSRR attempts to create a logical and reasonable connection between the structure and the retention behavior properties of compounds [15-17]. In the extracted numerical values, this connection is

discussed from diverse points of view. Furthermore, theoretically-based QSRR approaches are assumed to be a generally promising remedy for overcoming the lack of experimental data in complex chemical phenomena. In fact, each reliable QSRR model is used frequently to justify the molecular mechanism of chromatographic separation as well as the influence of polarity of the columns on retaining the solutes in mobile phase. The most important sequential steps engaged with each routine QSPR prediction involve: acquisition of a proper data set covering possible structural diversities within a defined group of substances, molecular geometry optimization, molecular descriptor generation, elimination of extra variables, feature selection, inspection on non-existence of significant correlation between the variables (bivariate correlation) used in the model development, external and cross validations, chance correlation and determination of the most effective contributions of each component in the final models. It is evident that the most critical step in each QSRR study is the feature selection step by which one can ensure that the constructed model is robust and has a high predictive capability. This step could be achieved using some powerful strategies like stepwise [12, 18], genetic algorithm [19-23] followed by multiple linear regressions (MLR), partial least squares (PLS), artificial neural network (ANN), support vector machine (SVM) or a combination of all [19, 23].

Despite the lower rate of growth in use of computational swarm intelligence (CSI), compared with other paradigms in artificial intelligence (AI), it is a challenging subject of interest. In recent decades, swarm-based algorithms have been of prime importance because of their promising ability to model local interactions among individuals under complicated conditions. Five factors in particular should be taken into consideration in each CSI computation, namely proximity, quality, diversity of response, stability and adaptability. These factors allow one, respectively, to

perform space and time computations, to respond to environmental quality factors, to produce a variety of different responses, to retain robust behaviors under mild environmental changes and to change behavior when it is dictated by external factors [24]. One of the more robust approaches within the CSI concept is particle swarm optimization (PSO), which has its origins in bird flocking models. In PSO, each individual is considered as a particle, following the best performing individual and moving towards the best conditions found by the individual itself. In terms of optimization, each particle moves towards two attractors, with the result that all particles converge on one solution. PSO can be considered to be one of the fastest variable selection methods, and it can be combined with regression methods [25].

The main objective of the present work was to build a reliable QSRR model and to evaluate its performance by comparison of calculated values with the retention indices of the constituent components of *Camu-camu* L. essential oil reported by Pino and colleagues [4]. It is proposed to use particle swarm optimization as one of the powerful selection techniques, in combination with multiple linear regressions for feature mapping.

A brief literature survey shows that despite the considerable progress made using this algorithm, very few papers have been published on prediction of diverse physicochemical parameters using PSO-MLR.

To the best of my knowledge, this is the first report concerning the use of PSO as a proper variable selection method for modeling the chromatographic retention behavior of the VEO components of *Camu-camu* L.

## MATERIALS AND METHODS

### *Computer hardware and software*

A Pentium IV personal computer (CPU at 3.06 GHz) with the Windows XP operating system was used. For the MLR analysis, the SPSS software package (version 14, SPSS, Inc.) statistical program was employed, and

more advanced calculations were performed in the MATLAB (Version 7.6, Mathworks, Inc., Natick, MA, USA) environment.

### *Instrument*

#### *Clevenger*

After harvesting the leaves of *Myrciaria dubia*, the water distilled oil was obtained from 100 g of leaves by simultaneous hydro-distillation-solvent extraction with 25 mL of pentane-diethyl ether (1:1, v/v) for 3 h. The extract was dried over anhydrous sodium sulfate and concentrated with a Kuderna-Danish apparatus coupled to a Vigreux column to eliminate the solvent.

#### *Gas chromatography*

An HP 6890 GC with a FID detector equipped with an HP-5MS fused silica capillary column film thickness (0.25  $\mu\text{m}$ ) was employed. The column temperature was programmed as follows: 70 °C isothermal for 2 min, 70–230 °C at 4 °C /min, then held for 10 min at 230 °C. Helium carrier gas was used at a flow-rate 1 mL/min. The injector and detector were maintained at 230 °C. The sample injection volume was 0.3  $\mu\text{L}$  with a split ratio of 1:10. Linear retention indices were calculated using n-paraffin standards.

#### *Gas chromatography-mass spectrometry*

An HP 6890 series II equipped with a mass selective detector HP-5973N was used. Similar capillary column and temperature programming were utilized as in the GC-FID technique. Mass spectra were recorded in the electron-impact mode at 70 eV.

#### *Data handling and descriptor generation*

The chemical structure of each component in the selected dataset was drawn using Hyperchem 8.0 (Hypercube, Inc) software package. The semi-empirical Austin Model 1 (AM1) Hamiltonian method was applied to optimize the corresponding 3D molecular structures [26].

The geometry optimization was done using the Polak–Ribiere algorithm until the root mean square gradient was 0.001 Kcal/mol. Geometry optimization was run

multiple times over a variety of starting points for each molecule, and the lowest energy conformation was utilized for the calculation of electronic properties. Full optimization of all bond lengths and angles was performed, Regardless of any symmetry constraint. All calculations were accomplished at the restricted Hartree-Fock level without any configuration interaction.

Dragon 2.1 software (Milano Chemometrics and QSAR Research Group, Milano, Italy) was used to seek the descriptors for model construction. Accordingly, entire optimized geometries were transferred into the Dragon program package. These descriptors can be classified into 18 groups, namely, the classes constitutional, topological, geometrical, charge, GETAWAY (Geometry, Topology and Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), molecular walk count, BCUT, 2D-autocorrelation, aromaticity index, Randic molecular profile, radial distribution function, functional group and atom-centered fragment. The molecular descriptor is defined as a numerical characteristic associated with a distinct chemical structure. As a matter of fact, it is the final result of a logical and mathematical procedure transforming chemical information encoded within a symbolic representation of a molecule into a useful number applied to correlate physical properties [27].

By means of a perfect and well documented interpretation of the utilized descriptors in the regression model, one can improve one's insights on effective parameters governing the trends of separation through chromatographic columns, and perceive which interactions play the key role in the retention processes.

#### *Preliminary treatments and principal component analysis*

In a pre-evaluation step, all the descriptors are checked to ensure that respective numerical values of each molecular descriptor are available for each structure. In

addition, significant variations in the corresponding values of each variable should be observed and unavailable or incomplete columns of variables should be discarded. Accordingly, a majority (about 90%) of zero values were excluded because they were unable to maintain any significant discrimination. Moreover, columns containing more than ninety percent of repeated values and/or constant or near constant patterns were removed from the initial matrix. The next step involved inspection of probable correlation between variables through pair correlation comparisons. This originates from the fact that in the final proposed models the descriptors used should be completely independent of each other. In accordance with to the non-redundant descriptors (NRD) rule, in order to decrease the redundancy existing in the descriptors data matrix, the correlations of descriptors were examined with each other and with the RIs of the molecules, and descriptors which showed high interrelation (i.e.,  $R > 0.90$ ) were detected [28]. For each cluster of the descriptors with close correlation coefficients, just one of them was kept for construction of the final QSRR model and the rest were deleted. In other words, descriptors that are highly correlated ( $R^2 \geq 0.95$ ) encode similar information and one of them should be eliminated. The remaining descriptors were gathered in an  $n \times m$  data matrix (D), where  $n = 108$  and  $m = 173$  are the numbers of compounds and descriptors, respectively. A column vector (y) was constructed using the RI data.

Principal component analysis (PCA) is a popular new criterion in data analysis for classification. PCA can clarify several underlying components and help to explain the extensive variance seen in the majority of data [29, 30]. The main aim in each PCA evaluation is to characterize each object in the input matrix (rows) without analysis of any variable (columns). Instead, the data is projected in a much smaller subset of new variables or principal component scores. These new variables (factors) are linear combinations of the initial

ones, highlighting the variance within a dataset that can remove the redundancies. As a matter of fact, successive principal components ranked in decreasing order of eigenvalues imply diminishing amounts of variance [31]. The principal components (PCs) are orthogonal or independent and are scaled so that their variances are equal to unity. Also, they are arranged so that the variance explained by the first PC is maximal, the variance explained by the second PC is smaller, and so on, with the last variance being the smallest [32]. Performance of PCA on the whole data of 115 compounds and descriptors and then plotting the first and second principals reveal that compounds 71, 107, 111, 112, 113, 114 and 15 are outliers (see Fig. 1). In Fig. 2, the molecular structures of the seven molecules out of the main cluster are shown.

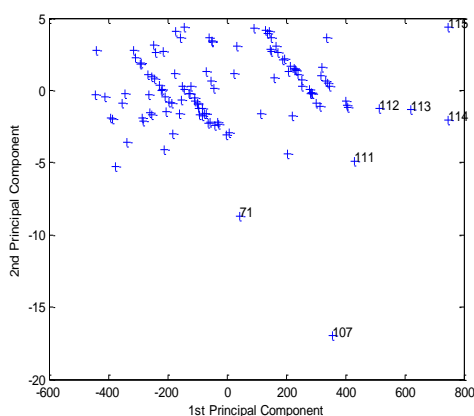


Figure 1. First and second principal components for the factor spaces of the descriptors and RI data

Fig. 1 shows that aforementioned molecules behave differently from other molecules with respect to both molecular structure (descriptors) and retention behavior (RI). Therefore, these molecules are discarded from the initial dataset in subsequent analysis. According to the distribution pattern of the data in factor spaces (Fig. 1), the training and prediction molecules were selected homogeneously, so that molecules in different zones are included in both subsets. The training set (80 compounds) was used to generate the best linear models while the test set (28 compounds) was used to evaluate its predictive ability. The chemical structures of the entire molecules, drawn using ChemDraw Ultra (ChemBioOffice 2008) package, are shown in Tables 1 and 2, along with their corresponding retention indices. Subsequently, the developed PSO algorithm was applied to the training set to find the most effective molecular descriptors justifying the RI parameter

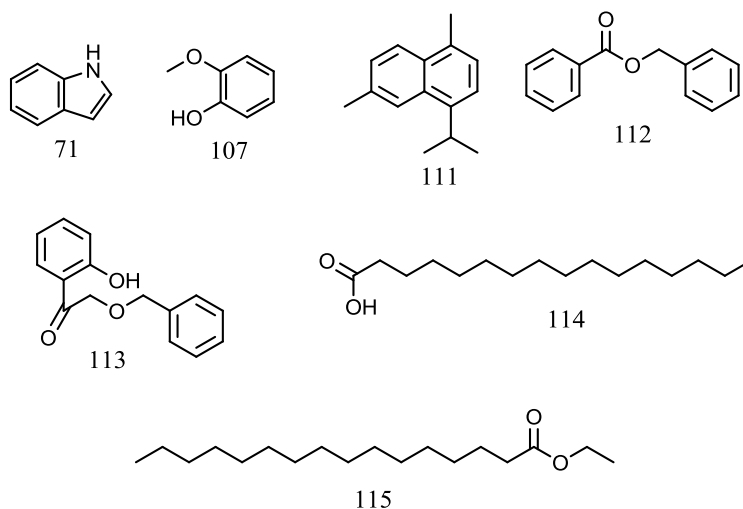


Figure 2. Structure of the outliers in the main dataset

## RESULTS AND DISCUSSION

### PSO-MLR model

Particle swarm optimization was developed by Kennedy and Eberhart in 1995 as a stochastic optimization algorithm based upon social simulation models [33].

The basic algorithm of PSO has the following nomenclature:

$x_z^i$  Particle position

$v_z^i$  Particle velocity

$w_{ij}$  Inertia weight

$p_z^i$  Best “remembered” individual particle position

$p_z^g$  Best “remembered” swarm position

$c_1, c_2$  Cognitive and social parameters

$r_1, r_2$  Random numbers between 0 and 1

Table 1. Names, RIs and molecular structures of the compounds used in the training set

No.	Compound	RI <sup>a</sup>	Structure	No.	Compound	RI	Structure	No	Compound	RI	Structure
1	hexanal	802		15	6-methyl-5-hepten-2-one	986		29	2-nonanone	1092	
2	ethyl butyrate	804		16	6-methyl-5-hepten-2-ol	992		30	nonanal	1101	
3	2-furfural	836		17	ethyl hexanoate	998		31	isoamyl isovalerate	1103	
4	(Z)-3-hexenol	859		18	$\alpha$ -phellandrene	1003		32	$\alpha$ -fenchol	1117	
5	Hexanol	871		19	$\alpha$ -terpinene	1017		33	$\alpha$ -campholenal	1126	
6	2-heptanone	892		20	p-cymene	1025		34	cis-p-menth-2,8-dien-1-ol	1138	
7	(E,E)-2,4-hexadienal	910		21	limonene	1029		35	trans-pinocarveol	1139	
8	$\alpha$ -thujene	930		22	benzyl-alcohol	1033		36	4-keto-isophorone	1148	
9	$\alpha$ -pinene	939		23	(Z)- $\beta$ -ocimene	1037		37	camphene hydrate	1150	
10	camphene	954		24	2-phenylacetaldehyde	1042		38	(E,Z)-2,6-nonadienal	1155	
11	thuja-2,4-(10)-diene	958		25	$\gamma$ -terpinene	1060		39	pinocarvone	1165	
12	(E)-2-heptenal	960		26	cis-sabinene hydrate	1070		40	borneol	1169	
13	$\beta$ -pinene	979		27	terpinolene	1087		41	trans-linalool oxide (pyranoid)	1175	
14	1-octen-3-ol	982		28	cis-linalool oxide (furanoid)	1089		42	p-cymen-8-ol	1183	

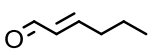
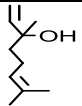
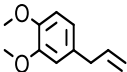
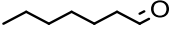
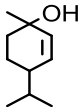
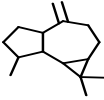

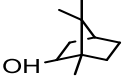
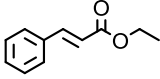
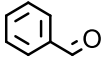
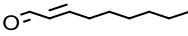
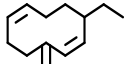
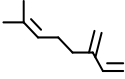

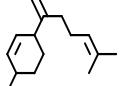
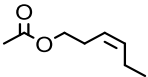

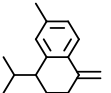
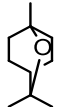
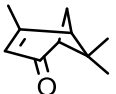
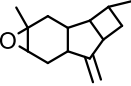
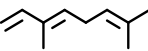
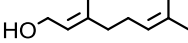
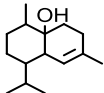

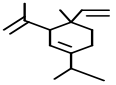
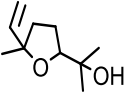
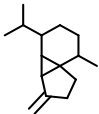
Table 1. (Continued)

No.	Compound	RI	Structure	No.	Compound	RI	Structure	No.	Compound	RI	Structure
43	(Z)-3-hexenyl butyrate	1186		57	benzyl isovalerate	1394		71	trans-calamenene	1529	
44	$\alpha$ -terpineol	1189		58	ethyl decanoate	1396		72	trans-cadina-1(2),4-diene	1535	
45	hexyl butyrate	1195		59	$\alpha$ -gurjunene	1410		73	$\alpha$ -calacorene	1546	
46	ethyl octanoate	1197		60	$\beta$ -caryophyllene	1419		74	(E)-nerolidol	1563	
47	decanal	1202		61	$\beta$ -gurjunene	1434		75	(Z)-3-hexenyl benzoate	1567	
48	trans-pulegol	1215		62	(E)-isoeugenol	1451		76	spathulenol	1578	
49	trans-carveol	1218		63	$\alpha$ -humulene	1455		77	globulol	1586	
50	carvone	1243		64	allo-aromadendrene	1460		78	viridiflorol	1593	
51	ethyl salicylate	1270		65	cis-cadina-1,(6),4-diene	1472		79	$\alpha$ -muurolol	1650	
52	(Z)-3-hexenyl isovalerate	1281		66	trans-cadina-1,(6),4-diene	1477		80	$\alpha$ -cadinol	1654	
53	eugenol	1359		67	$\gamma$ -muurolene	1480					
54	$\alpha$ -copaene	1377		68	bicyclogermacrene	1500					
55	(Z)-3-hexenyl hexanoate	1382		69	$\alpha$ -muurolene	1500					
56	$\beta$ -elemene	1391		70	$\delta$ -cadinene	1523					

<sup>a</sup> Retention index



Table 2. Names, RIs and molecular structures of the compounds used in the testing set.

No.	Compound	RI	Structure	No.	Compound	RI	Structure	No.	Compound	RI	Structure
1	(E)-2-hexenal	855		11	linalool	1097		21	methyl eugenol	1404	
2	heptanal	902		12	cis-p-menth-2-en-1-ol	1122		22	aromadendrene	1441	
3	$\alpha$ -fenchene	953		13	camphor	1146		23	ethyl-(E)-cinnamate	1467	
4	benzaldehyde	962		14	(E)-2-nonenal	1162		24	germacrene D	1487	
5	myrcene	991		15	terpinen-4-ol	1177		25	(E)- $\gamma$ -bisabolene	1531	
6	(Z)-3-hexenyl acetate	1005		16	methyl salicylate	1192		26	$\beta$ -calacorene	1560	
7	1,8-cineol	1031		17	verbenone	1205		27	caryophyllene oxide	1583	
8	(E)- $\beta$ -ocimene	1050		18	geraniol	1253		28	cubenol	1647	
9	acetophenone	1065		19	$\delta$ -elemene	1338					
10	trans-linalool oxide (furanoid)	1073		20	$\beta$ -cubebene	1388					

Calculation of the velocity is done using the following equation:

$$v_{z+1}^i = w_{ij} v_z^i + c_1 r_1 (p_z^i - x_z^i) + c_2 r_2 (p_z^g - x_z^i), \quad (\text{eq. 1})$$

Whereas the position of the individual particles is determined as follows:

$$x_{z+1}^i = x_z^i + v_{z+1}^i. \quad (\text{eq. 2})$$

The basic PSO algorithm is defined as follows [34]:

(1) Initialize

(a) Set constants  $z_{\max}$ ,  $c_1$ ,  $c_2$

(b) Randomly initialize particle position  $x_0^i \in D$  in  $R^n$  for  $i=1, \dots, p$

(c) Randomly initialize particle:

velocities  $0 \leq v_o^i \leq v_i^{\max}$  for  $i=1, \dots, p$

(d) Set  $Z = 1$

(2) Optimize

(a) Evaluate function value using design space coordinates

(b) If  $f_z^i \leq f_{best}^i$  then  $f_{best}^i = f_z^i$ ,  $p_z^i = x_z^i$ .

(c) If  $f_z^i \leq f_{best}^g$  then  $f_{best}^g = f_z^i$ ,  $p_z^g = x_z^i$ .

(d) If stopping condition is satisfied then go to 3.

(e) Update all particle velocities  $v_z^i$  for  $i = 1, \dots, p$

- (f) Update all particle positions  $x_z^i$  for  $i = 1, \dots, p$   
 (g) Increment  $z$ .  
 (h) Go to 2(a).  
 (3) Terminate.

The algorithm is based on the population of search points moving stochastically in the search space. The main inspiration and criteria defining this powerful approach are derived from concepts and rules that govern socially organized populations in nature, such as flocks of birds, schools of fish and herds of animals. As a matter of fact, wonderful potential for self-organization and impressive behaviors of each group lead to well ordered collective behaviors that cannot be described simply by aggregating the behavior of each team member. The common principle in all of these templates is that an elite member dictates the best and the most straightforward pathway to the target. Since its development, PSO has gained growing attention in engineering fields due to its ability to provide solutions efficiently, requiring only minimal implementation efforts. In PSO, the best situation ever encountered by each individual, namely its experience, is saved in memory followed by its merger to part or the whole population, and ends by biasing its movement towards the most promising detected regions. The communication scheme is determined by a fixed or adaptive social network serving a crucial role on the convergence properties of the algorithm. A training set with eighty compounds was employed to establish a reliable model. The compounds selected in this set are logical representatives of the entire dataset according to the PCA plot, after elimination of outliers. This study gave rise to distinguishing three independent molecular descriptors contributing to the best linear model. The

adopted descriptors, which have appreciable correlations with Kovatz indices (KI), belong to topological (PCR and SOK) and 2D autocorrelation (MATS4e) classes, and appear in the form:

$$\text{KI} = 474.479 + 300.303(\pm 78.310)\text{MATS4e} - 4.026(\pm 9.149)\text{PCR} + 28.441(\pm 0.886)\text{SOK} \quad (\text{eq. 3})$$

$N_{\text{train}}=80; R=0.968; R^2_{\text{train}}=0.937; \text{adjusted } R^2=0.934; \text{standard error of estimate}=59.39; F\text{-statistical}=376.4; P \text{ value} < 10^{-4}$

Table 3 lists all the types of molecular descriptors available from Dragon (2.1) and their numbers after performance of NRD step, together with a brief description of the implemented ones in the linear model. Meanwhile, bivariate analyses on each pair of the utilized descriptors in the proposed model gave rise to the following correlation coefficients (r):

- i) MATS4e and PCR: 0.139
- ii) MATS4e and SOK: 0.105
- iii) PCR and SOK: 0.232

These negligible relationships between the selected variables reveal their independent identities in prediction of RIs. In addition, to pre-processing and a supplemental evaluation, the descriptors were standardized by centering, prior to scaling to unit variance. This was mainly to give all variables an equal chance to influence the model, regardless of their original size or variance [16]. The values of this statistical attempt for MATS4e, PCR and SOK were 0.112, -0.013 and 0.953, respectively. Accordingly, the highest value found was related to SOK, confirming its superior influence over MATS4e and PCR in the model.

Table 3. Descriptors implemented in the PSO-MLR model

No.	Descriptor type	No. of calculated descriptors	No. of descriptors remaining <sup>a</sup>	Name of descriptors the PSO-MLR model	Coefficient
1	Constitutional	47	-	-	-
2	Topological	262	10	PCR <sup>b</sup> and SOK <sup>c</sup>	-4.026 +28.441
3	Molecular walk counts	21	1	-	-
4	BCUT	64	4	-	-
5	Galvez topol. charge indices	21	-	-	-
6	2DAutocorrelation	96	41	MATS4e <sup>d</sup>	+300.303
7	Charge	14	3	-	-
8	Aromaticity indices	4	-	-	-
9	Randic molecular profiles	41	1	-	-
10	Geometrical	58	6	-	-
11	RDF	150	-	-	-
12	3D-MoRSE	160	53	-	-
13	WHIM	99	15	-	-
14	GETAWAY	197	38	-	-
15	Functional	121	-	-	-
16	Atom-centred fragments	120	-	-	-
17	Empirical	3	1	-	-
18	Properties	3	-	-	-
19	Sum	1481	173	3	-

<sup>a</sup> After NRD stage [27]<sup>b</sup> PCR: Ratio of multiple path counts to path counts<sup>c</sup> SOK: Kier symmetry index topological descriptors<sup>d</sup> MATS4e: Moran autocorrelation-lag 4/weighted by atomic Sanderson electronegativities

Tables 4 and 5 show the numerical values, deviations and relative errors computed for corresponding compounds both in the training and in the test sets. The closeness of the predicted RIs with experimental values indicates that this is a valid strategy.

#### Model validation and statistical parameters

Leave-one-out (*LOO*), leave-group-out (*LGO*) cross-validations and external validation (*EV*) were conducted to validate the predictive power, consistency and reliability of the constructed model. The former two approaches (*LOO* and *LGO*) were applied only to the training set, while the *EV* was applied to both the training and the test sets. For *LOO* cross-validation (eq. 4), a data point is removed from the set, and the model is rebuilt with the remaining compounds. The predicted property for the discarded compound is then compared with its actual value. This is repeated until each data point is omitted once. For *LGO* (eq. 5), 20% of the data points are

removed from the dataset, and the model is refitted. The predicted values for those points are then compared with their corresponding experimental values. Again, this is repeated until each data point has been omitted once.

In the two equations above, PRESS and SSY are the prediction error sum of the squares and sum of the squares of the deviations of the experimental values from their mean, respectively. PRESS is a standard index to measure the accuracy of a modeling method based on the cross-validation technique, and can be defined as eq. 6. The cross-validation correlation coefficient ( $Q^2$ ) is 0.921 for *LGO* and 0.928 for *LOO*. This confirms that the obtained regression model has a good internal and external predictive power. In Fig. 3 the integrated plot of predicted values for both sets (training and test) together with calculated values for *LOO* and *LGO* cross validations is shown.

Table 4. The observed and predicted RI values by PSO-MLR for the training set of *Myrciaria dubia*, differences as well as their relative errors

Num.	Exp. RI	Calc. RI <sup>a</sup>	D <sup>b</sup>	RE(%) <sup>c</sup>	Num.	Exp. RI	Calc. RI	D	RE(%)
1	802	841.8	39.8	5.0	30	1101	1056.4	-44.6	-4.1
2	804	867.1	63.1	7.8	31	1103	1206.2	103.2	9.4
3	836	962.1	126.1	15.1	32	1117	1076.5	-40.5	-3.6
4	859	848.8	-10.2	-1.2	33	1126	1168.7	42.7	3.8
5	871	752.1	-118.9	-13.7	34	1138	1187.6	49.6	4.4
6	892	912.8	20.8	2.3	35	1139	1160.6	21.6	1.9
7	910	901.9	-8.1	-0.9	36	1148	1197.9	49.9	4.3
8	930	982.7	52.7	5.7	37	1150	1076.2	-73.8	-6.4
9	939	1004.8	65.8	7.0	38	1155	1122.3	-32.7	-2.8
10	954	964.9	10.9	1.1	39	1165	1195.4	30.4	2.6
11	958	1014.2	56.2	5.9	40	1169	1110.5	-58.5	-5
12	960	952.8	-7.2	-0.8	41	1175	1197.8	22.8	1.9
13	979	1014.1	35.1	3.6	42	1183	1156.2	-26.8	-2.3
14	982	959.7	-22.3	-2.3	43	1186	1226.2	40.2	3.4
15	986	993.9	7.9	0.8	44	1189	1130.2	-58.8	-4.9
16	992	969	-23	-2.3	45	1195	1201.1	6.1	0.5
17	998	1048.7	50.7	5.1	46	1197	1211	14	1.2
18	1003	1025.6	22.6	2.3	47	1202	1129.8	-72.2	-6
19	1017	1008.6	-8.4	-0.8	48	1215	1142.8	-72.2	-5.9
20	1025	994.4	-30.6	-3.0	49	1218	1167.8	-50.2	-4.1
21	1029	1042	13	1.3	50	1243	1198.4	-44.6	-3.6
22	1033	938.9	-94.1	-9.1	51	1270	1480.7	210.7	16.6
23	1037	1041.3	4.3	0.4	52	1281	1324.2	43.2	3.4
24	1042	1094.6	52.6	5.0	53	1359	1372.2	13.2	1.0
25	1060	1008.9	-51.1	-4.8	54	1377	1443.8	66.8	4.9
26	1070	1121.3	51.3	4.8	55	1382	1394.4	12.4	0.9
27	1087	998.2	-88.8	-8.2	56	1391	1467	76	5.5
28	1089	1211.8	122.8	11.3	57	1394	1461.2	67.2	4.8
29	1092	1056.4	-35.6	-3.3	58	1396	1370.6	-25.4	-1.8

Table 4. (Continued)

Num.	Exp. RI.	Calc. RI	D	RE(%)	Num.	Exp. RI.	Calc. RI	D	RE(%)
59	1410	1460.3	50.3	3.6	70	1523	1445.5	-77.5	-5.1
60	1419	1476	57	4.0	71	1529	1491.3	-37.7	-2.5
61	1434	1433.2	-0.8	-0.1	72	1535	1454.5	-80.5	-5.2
62	1451	1359.6	-91.4	-6.3	73	1546	1527.3	-18.7	-1.2
63	1455	1486.9	31.9	2.2	74	1563	1540.6	-22.4	-1.4
64	1460	1466.3	6.3	0.4	75	1567	1573	6	0.4
65	1472	1450.7	-21.3	-1.4	76	1578	1595.6	17.6	1.1
66	1477	1450.7	-26.3	-1.8	77	1586	1557.3	-28.7	-1.8
67	1480	1456.9	-23.1	-1.6	78	1593	1557.3	-35.7	-2.2
68	1500	1482.1	-17.9	-1.2	79	1650	1549	-101	-6.1
69	1500	1452.1	-47.9	-3.2	80	1654	1549	-105	-6.3

<sup>a</sup> Particle swarm optimization multiple linear regression<sup>b</sup> Deviation<sup>c</sup> Relative error (%)Table 5. The observed and predicted RI values PSO-MLR for the test set for *Myrciaria dubia* as well as deviations and relative errors

Num.	Exp. RI.	Calc. RI <sup>a</sup>	D <sup>b</sup>	RE(%) <sup>c</sup>	Num.	Exp. RI.	Calc. RI	D	RE(%)
1	855	881.6	26.6	3.1	15	1177	1083.9	-93.1	-7.9
2	902	912.8	10.8	1.2	16	1192	1384.7	192.7	16.2
3	953	994.3	41.3	4.3	17	1205	1222.3	17.3	1.4
4	962	963	1.0	0.1	18	1253	1161.6	-91.4	-7.3
5	991	1050.9	59.9	6.0	19	1338	1446.8	108.8	8.1
6	1005	1101.2	96.2	9.6	20	1388	1442.2	54.2	3.9
7	1031	1075.7	44.7	4.3	21	1404	1456.9	52.9	3.8
8	1050	1041.3	-8.7	-0.8	22	1441	1466.3	25.3	1.8
9	1065	1034	-31	-2.9	23	1467	1418.9	-48.1	-3.3
10	1073	1211.8	138.8	12.9	24	1487	1466.7	-20.3	-1.4
11	1097	1122.6	25.6	2.3	25	1531	1463.8	-67.2	-4.4
12	1122	1132.8	10.8	1.0	26	1560	1535.3	-24.7	-1.6
13	1146	1139.2	-6.8	-0.6	27	1583	1697.2	114.2	7.2
14	1162	1096.2	-65.8	-5.7	28	1647	1525.9	-121.1	-7.4

<sup>a</sup> Particle swarm optimization multiple linear regression<sup>b</sup> Deviation<sup>c</sup> Relative error (%)

$$Q_{LOO}^2 = \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^n (y_{exp} - y_{pred})^2}{\sum_{i=1}^n (y_{exp} - \bar{y})^2} \quad (\text{eq. 4})$$

$$Q_{LGO}^2 = \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{test} (y_{exp} - y_{pred})^2}{\sum_{i=1}^{test} (y_{exp} - \bar{y}_{train})^2} \quad (\text{eq. 5})$$

$$PRESS = \sum_{i=1}^n (y_{pred} - y_{exp})^2 \quad (\text{eq. 6})$$

The good agreement between the experimental and predicted indices confirms the considerable potential of the proposed PSO-MLR model for a wide range of natural compounds. Fig. 4 lists the residuals (experimental RI - predicted RI) versus experimental RI values, obtained by PSO-MLR modeling. The normal distribution of the residuals on both sides of the zero line indicates there is no systematic error in the model.

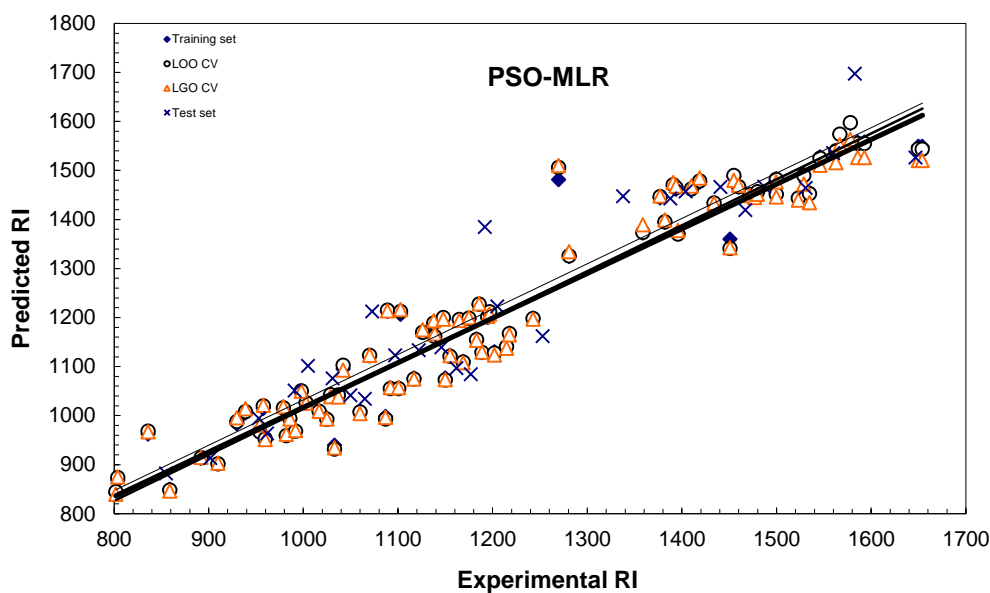
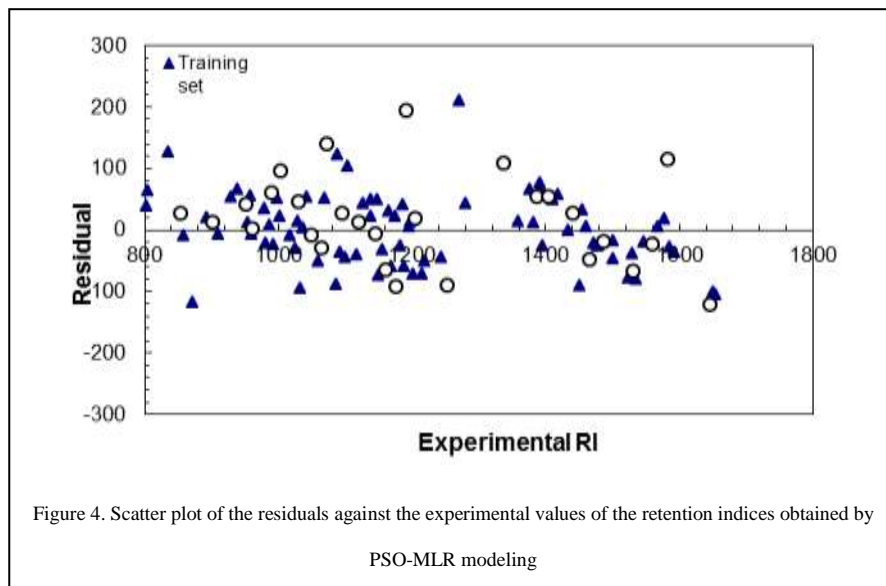


Figure 3. The cross validations results and predicted RIs vs. the experimental values for training and test sets by the PSO-MLR modeling

As it is customary in all QSRR studies, six common statistical parameters were used to appraise the quality of the constructed models. These parameters are square correlation coefficient ( $R^2$  or determination coefficient), relative error of prediction (REP), root mean square error of prediction (RMSEP), standard error of prediction (SEP), relative standard error of prediction (RSEP) and

mean absolute error (MAE). The first parameter is  $R^2$  (eq.7), which indicates the quality of fitness or divergence of the points from a straight line, can be calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{pred} - \bar{y})^2}{\sum_{i=1}^n (y_{exp} - \bar{y})^2} \quad (\text{eq.7})$$



The second and third statistical characters are REP and RMSEP. REP (eq. 8) represents the predictive ability of each individual, with low values being desired, while RMSEP (eq. 9) relates the mean discrepancy between predicted and experimental values. It can be interpreted as the average prediction error having the same units as the original response values. The third and fourth parameters are SEP and RSEP, which can be determined from equations 10 and 11, respectively. SEP and RSEP are general methods used to evaluate the predictive applicability of a regression model. The final figure of merit is MAE which is a statistical term dealing with the average distance of predicted values from their exact ones; it can be determined using eq. 12.

$$REP(\%) = \frac{100}{\bar{y}} \left[ \frac{1}{n} \sum_{i=1}^n (y_{pred} - y_{exp})^2 \right]^{0.5} \quad (\text{eq. 8})$$

$$RMSEP = \left[ \frac{1}{n} \sum_{i=1}^n (y_{pred} - y_{exp})^2 \right]^{0.5} \quad (\text{eq. 9})$$

$$SEP = \left[ \frac{\sum_{i=1}^n (y_{pred} - y_{exp})^2}{n - 1} \right]^{0.5} \quad (\text{eq. 10})$$

$$SEP = \left[ \frac{\sum_{i=1}^n (y_{pred} - y_{exp})^2}{n - 1} \right]^{0.5} \quad (\text{eq. 10})$$

$$RSEP(\%) = 100 \left[ \frac{\sum_{i=1}^n (y_{pred} - y_{exp})^2}{\sum_{i=1}^n (y_{exp})^2} \right]^{0.5} \quad (\text{eq. 11})$$

$$MAE(\%) = \frac{100}{n} \left[ \sum_{i=1}^n |(y_{pred} - y_{exp})| \right]^{0.5} \quad (\text{eq. 12})$$

In all the above equations,  $y_{pred}$ ,  $y_{exp}$ ,  $\bar{y}$ ,  $\bar{y}_{train}$  and  $n$  represent the predicted retention index, the experimental value of the RIs, the mean of experimental RI in the

predicted (test) set, the mean RI in the training set and number of samples in the training or test sets,

respectively. All of the statistical results both for training and for test sets are given in Table 6.

Table 6. Statistical parameters related to PSO-MLR model

Num	Parameter	Adopted set	Value
1	$R^2$ <sup>a</sup>	Training set	0.936
		Test set	0.898
2	REP (%) <sup>b</sup>	Training set	4.8
		Test set	6.0
3	RMSEP <sup>c</sup>	Training set	57.9
		Test set	73.5
4	SEP <sup>d</sup>	Training set	58.3
		Test set	74.8
5	RSEP <sup>e</sup>	Training set	4.7
		Test set	5.9
6	MAE <sup>f</sup>	Training set	75.7
		Test set	142.8

<sup>a</sup> Determination coefficient; <sup>b</sup> Relative error of prediction; <sup>c</sup> Root mean square error of prediction; <sup>d</sup> Standard error of prediction; <sup>e</sup> Relative standard error of prediction; <sup>f</sup> Mean absolute error

#### *Y-randomization test*

To insure of either non-existence of chance correlation between the implemented variables in the model or model robustness, *Y*-randomization test is a proper tool [35]. Accordingly, the dependent variable vector (RI) is randomly shuffled or scrambled and new QSRR models are explored using the original independent feature matrix. This scrambling destroys any relationship between the descriptors and the dependent feature. It is expected that over the several repetitions, the constructed models have low  $R^2$  and  $Q^2$  values. In Table

7, some of the statistical parameters obtained by *Y*-randomization have been demonstrated for 10 iterations. The negligible values of  $R^2$  and  $Q^2$ , compared with the original models, argue for the validity of the computations. Moreover, a simple comparison of the outputs obtained by *Y*-randomization with those obtained using the constructed PSO-MLR, demonstrates that the results obtained using the proposed model are based upon rationale relationships rather than just chance correlations.

Table 7. PRESS, CV and statistical values after several *Y*-randomization tests

PRESS <sup>a</sup>	S <sub>PRESS</sub> <sup>b</sup>	SST <sup>c</sup>	$R^2$ CV <sup>d</sup>	PRESS/SST	REP	RMSEP	SEP	RSEP	$R^2$ LOO	$R^2$ LGO
4452408	242.042	482837.7	0.011213	9.221334	367828.7	234.4525	235.9133	19.02941	4.47exp-10	0.005142
4550598	244.6963	458003.8	0.004195	9.935721	375940.5	237.0237	238.5005	19.2381	0.010141	0.000147
4563679	245.0478	257334.4	0.000244	17.73443	377021.2	237.3641	238.843	19.26573	0.029028	0.024657
4875626	253.2844	463731.9	0.002411	10.51389	402792.3	245.3424	246.8711	19.9133	0.001339	0.000485
5023715	257.1021	288473.6	0.043087	17.41481	415026.4	249.0405	250.5922	20.21345	0.056903	0.0395
4621214	246.5876	265801.1	0.001495	17.38599	381774.4	238.8557	240.3439	19.38679	0.013495	0.009571
5103269	259.1298	154333.7	0.173084	33.06645	421598.6	251.0046	252.5685	20.37287	0.130406	0.331035
4413296	240.9765	200496.1	0.001049	22.01188	364597.6	233.4205	234.8749	18.94565	0.005529	0.01004
4532224	244.2018	279355.4	8.29E-05	16.22387	374422.6	236.5447	238.0185	19.19922	0.00486	0.001844
4637226	247.0144	291588.1	0.001049	15.90334	383097.2	239.2691	240.7599	19.42035	3.51exp-5	0.001055

<sup>a</sup> Prediction error sum of the squares

<sup>b</sup> Uncertainty of prediction

<sup>c</sup> The total sum of squares

<sup>d</sup> Cross-validated square correlation coefficient

#### *Comparison of the constructed model with similar reports by others*

During the recent years, several QSRR models have been developed for prediction the numerical values of RIs owing to a wide variety of organic compounds encountered in diverse conditions [15, 23, 36].



Regardless of discrepancies in feature selection steps, some points are common to all of them. Table 8 lists brief characteristics of the most popular, published ways

to model RIs of some of the VEOs. A simple inspection of the data tabulated in this Table leads to the observations below.

Table 8. Comparison of the proposed models in modeling of chromatographic retention behavior using different QSRR approaches

Essential oil sample	Descriptors classes	Technique	S.P. <sup>1</sup>	Ref.
<i>Pistacia lentiscus</i> L.	Topological, total charge, WHIM, GETAWAY, 2D autocorrelations	SW-MLR <sup>a</sup> , GA-MLR <sup>b</sup>	BP-1, BP-20	17
<i>Rosemary</i> and <i>sage</i>	Topological, constitutional, electronic, quantum mechanical	GA-MLR, GA-PLS <sup>c</sup> , L-M ANN <sup>d</sup>	DB-5	19
<i>Bidens pilosa</i> Linn. var. <i>Radiata</i>	Topological, 2D autocorrelations, GETAWAY, 3D-MoRSE, properties, WHIM, atom centred-fragment	GA-MLR	DB-5MS column	20
<i>Citrus sudachi</i>	Topological, constitutional, RDF, GETAWAY, empirical, properties	MLR, PLS <sup>e</sup> , P-PLS <sup>f</sup> , SVM <sup>g</sup>	HP5	22
<i>Ylang-Ylang</i>	Topological, geometric, electronic	MLR <sup>h</sup>	DB-1, DB-wax	27
six <i>Stachys</i> species	Topological, constitutional, atom centred-fragment, electronic, quantum mechanical	GA-MLR, GA-PLS, GA-KPLS <sup>i</sup> , L-M ANN <sup>j</sup>	Innowax (GC); HP-1 (GC-MS)	35
Camu-camu ( <i>Myrciaria dubia</i> (HBK)Mcvaugh)	Topological, 2D autocorrelations descriptors	PSO-MLR <sup>k</sup>	HP-5MS	This work

<sup>a</sup> Stepwise multiple linear regression; <sup>b</sup> Genetic algorithm-multiple linear regression; <sup>c</sup> Genetic algorithm-partial least squares; <sup>d</sup> Levenberg-Marquardt artificial neural network; <sup>e</sup> Partial least squares; <sup>f</sup> Poly partial least squares; <sup>g</sup> Support vector machine; <sup>h</sup> Multiple linear regression; <sup>i</sup> Genetic algorithm-kernel partial least squares; <sup>j</sup> Artificial neural network; <sup>k</sup> Particle swarm optimization-multiple linear regression; <sup>1</sup> Stationary phase

1. Topological indices are used in all the interpreted models. Topological descriptors include valence and non-valence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule, encoding information about the size, composition and the degree of branching of a molecule. The significant influence of this molecular descriptor type is reasonable since the size of the solutes penalizes the quality of physicochemical equilibria through the chromatographic capillary columns. In the proposed model, two molecular descriptors from topological class contribute, namely S0K and PCR. The interesting point here is a clear dissimilarity between the signs of these variables; this is a somewhat challenging issue.

S0K, which is defined as the Kier symmetry index, has a positive sign showing its reinforcement nature against the model, while another descriptor possesses a negative character and a converse relation with RI. Symmetry is a critical parameter affecting the physicochemical equilibria between each solute and its adjacent stationary phase of low polarity. Symmetry and non-polar identity of a distinct species are in close and direct relationship. In other words, more symmetry results in an enhanced trend in non-polar nature for a substance,

and according to the “like dissolves like” rule, non-polar solutes in the mobile phase have much more time to maintain interactions while polar compounds with low degree of symmetry are exited from the chromatographic columns.

PCR is measure of the ratio of multiple path counts to path counts. This molecular descriptor essentially depicts the eddy diffusion term or multi-flow paths (A) in the van Deemeter equation (eq. 10) in GC

$$H = A + \frac{B}{U} + C_s U + C_m U \quad (\text{eq. 10})$$

Eddy diffusion is one of the most important factors responsible for band broadening especially in packed columns [37], and is induced by the uneven diameter of the stationary phase or support capillary columns, which automatically results in unexpected mobile phase flow-rate through the solid bed. Some solute molecules are thus displaced more quickly than the average, whereas others are retarded. This contribution is related to the variety of channels available for any solute molecule throughout the elution process. These channels are defined by the interstitial volume between the beads of the column package, so they correspond to a variety of

shapes and flow velocities. This produces a distribution in elution time which is classically considered as Gaussian and weakly depends on flow rate. As a rule of thumb, the theoretical plate height corresponding to this effect can be considered as being equal to the bead diameter of the packing for well-packed columns [38]. In fact, two cases are effective in the definition of A, which are the quality of column packing and the mean particle size of each solute. In designing the new columns, uniformity of packing and low distribution of particle sizes of the stationary phases diminish the unfavorable pathways, and also improve the quality of chromatographic peaks in each determination. The agreement of the constructed model with those previously mentioned implies its high potential for prediction of RIs of a variety of natural compounds found in the volatile oils.

2. The charges of ionic species have considerable effects on RIs especially in ion chromatography.

3. WHIM descriptors are molecular descriptors, based on statistical indices, and are calculated based on the projections of the atoms along principal axes [27]. They are built in such a way as to capture relevant molecular 3-dimensional information regarding to the molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. These indices are calculated from (x, y, z)-coordinates of a molecule within different weighting schemes in a straightforward manner and represent a very general approach to describing molecules in a unitary conceptual framework. A detailed description of their chemical meaning and of the WHIM theory is reported elsewhere [39].

4. GETAWAY descriptors are mainly based on a leverage matrix (molecular influence matrix) and can be easily calculated from the spatial coordinates of the atoms in a molecule or by summing atom weights viewed by a divergent angular scattering function. They are relatively new descriptors and were highly developed by Consonni and co-workers [40]. Since

these descriptors are associated with atomic masses and van der Waals characteristics, their restricted behavior can be understood. Experimentally, larger solutes in chromatography exit sooner along the columns.

5. 2D autocorrelations are spatial autocorrelations calculated based on an H-depleted molecular graph weighted by atomic physico-chemical properties. 2D autocorrelations are molecular descriptors which describe how a considered property is distributed along a topological molecular structure. Molecular property includes a set of heterogeneous molecular descriptors describing physicochemical and biological properties as well as some molecular characteristics obtained by published models [27]. In the proposed PSO-MLR model, one 2D autocorrelation is present (MATS 4e) with positive sign which means an appreciable in RI enhancement with its increase.

Finally, the contribution of the various descriptors applied to MLR models is shown in Fig. 5. According to Fig. 5, topological and 2D autocorrelations have the most importance in the models.

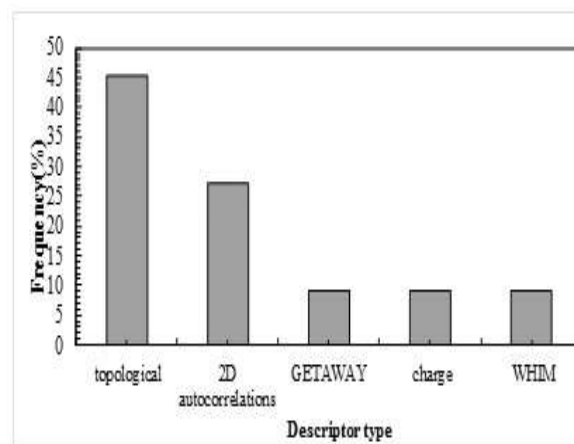


Figure 5. Total frequencies of the predictors in the proposed models

## CONCLUSION

This study is the first to focus mainly on particle swarm optimization (PSO) combined with multiple linear regression (MLR) to predict retention indices (RIs). Principal component analysis revealed the seven compounds were out of the main cluster and should be discarded. The training (calibration) and testing (predicted) sets included 80 and 21 compounds, respectively. The linearly constructed model involved only three molecular descriptors, of which two (PCR and S0K) belong to the topological group and one (MATS4e) was from 2D autocorrelation class. According to standardization, the most contribution was due to the S0K variable. The promising output of this QSRR based model implies that the proposed strategy can be effectively utilized for simulating the RIs of VEOs constituents passing through chromatographic columns. The proposed model can also provide insights for predicting these key parameters for unknown compounds occurring in a wide variety of VEOs. Work is continuing in our laboratory to improve this strategy and to increase its usefulness.

## ACKNOWLEDGEMENT

The author expresses his special thanks to the Office of Research Affairs of the Islamic Azad University of Shahrood for financial support and the time dedicated to this work.

## REFERENCES

1. [http://en.wikipedia.org/wiki/Essential\\_oil](http://en.wikipedia.org/wiki/Essential_oil).
2. Akhlaghi H., Nekoei M., Mohammadhosseini M., Motavalizadehkakhky A., 2012. Chemical composition of the volatile oils from the flowers, stems and leaves of *Prangos latiloba* Korov. using the head space solid phase microextraction method prior to analysis by gas

chromatography-mass spectrometry. Journal of Essential Oil Bearing Plants. 15, 328-335.

3. Eminagaoglu O., Tepe B., Yumrutas O., Akpulat H. A., Daferera D., Polissiou M., Sokmen A., 2007. The *in vitro* antioxidative properties of the essential oils and methanol extracts of *Satureja spicigera* (K. Koch.) Boiss. and *Satureja cuneifolia* ten. Food Chemistry. 100 (1): 339-343.
4. Pino J.A., Quijano C.E.C., 2008. Volatile constituents of *Camu-camu* (*Myrciaria dubia* (HBK) McVaugh) leaves. Journal of Essential Oil Research. 20 (3): 205-207.
5. Heberger K., 2007. Quantitative structure-(chromatographic) retention relationships. Journal of Chromatography A. 1158 (1-2): 273-305.
6. Goudarzi N., Goodarzi M., Mohammadhosseini M. M., Nekoei M., 2009. QSPR models for prediction of half wave potentials of some chlorinated organic compounds using SR-PLS and GA-PLS methods. Molecular Physics. 107 (17): 1739-1744.
7. Hansch C., Leo A. 1995. Exploring QSAR Fundamentals: Applications in Chemistry and Biology. American Chemical Society. Washington DC.
8. Khajeh A., Modarress H., 2012. Quantitative structure-property relationship prediction of liquid heat capacity at 298.15 K for organic compounds. Industrial & Engineering Chemistry Research. 51 (17): 6251-6255.
9. Mirkhani S.A., Gharagheizi F., 2012. Predictive quantitative structure-property relationship model for the estimation of ionic liquid viscosity. Industrial & Engineering Chemistry Research. 51 (5): 2470-2477.
10. Mohammadhosseini M., Nekoei M., 2012. Quantitative structure-electrochemistry relationship (QSER) study for prediction of half-wave reduction potentials of some chlorinated organic compounds by GA-MLR. Asian Journal of Chemistry. 25, 349-352.
11. Nekoei M., Salimi M., Dolatabadi M., Mohammadhosseini M., 2011. Prediction of

- antileukemia activity of berbamine derivatives by genetic algorithm-multiple linear regression. *Monatshefte fur Chemie*. 142 (9): 943-948.
12. Nekoei M., Salimi M., Dolatabadi M., Mohammadhosseini M., 2011. A quantitative structure-activity relationship study of tetrabutylphosphonium bromide analogs as muscarinic acetylcholine receptors agonists. *Journal of the Serbian Chemical Society*. 76 (8): 1117-1127.
13. Senior S.A., Nassar A.M., 2012. Determination of  $pK_a$  for substituted benzoic acids in mixed solvent using density functional theory and QSPR. *International Journal of Quantum Chemistry*. 112, 683-694.
14. Toropova A.P., Toropov A.A., Benfenati E., Gini G., 2012. QSAR models for toxicity of organic substances to daphnia magna built up by using the CORAL freeware. *Chemical Biology & Drug Design*, 79 (3): 332-338.
15. Fragkaki A.G., Tsantili-Kakoulidou A., Angelis Y.S., Koupparis M., Georgakopoulos C., 2009. Gas chromatographic quantitative structure-retention relationships of trimethylsilylated anabolic androgenic steroids by multiple linear regression and partial least squares. *Journal of Chromatography A*, 1216 (47): 8404-8420.
16. Ghasemi J., Asadpour S., Abdolmaleki A., 2007. Prediction of gas chromatography/electron capture detector retention times of chlorinated pesticides, herbicides, and organohalides by multivariate chemometrics methods. *Analytica Chimica Acta*. 588 (2): 200-206.
17. Mohammadhosseini M., Deeb O., Alavi- Gharabagh A., Nekoei M., 2012. Exploring novel QSRRs for simulation of gas chromatographic retention indices of diverse sets of terpenoids in *Pistacia lentiscus* L. essential oil using stepwise and genetic algorithm multiple linear regressions. *Analytical Chemistry Letters*. 2, 80-102.
18. Mohammadhosseini M., Zamani H.A., Akhlaghi H., Nekoei M., 2011. Hydrodistilled volatile oil constituents of the aerial parts of *Prangos serpentina* (Rech.f., Aell. Esfand). Herznstadt and Heyn from Iran and quantitative structure-retention relationship simulation. *Journal of Essential Oil Bearing Plants*. 14 (5): 559-573.
19. Chen H.F., 2008. Quantitative predictions of gas chromatography retention indexes with support vector machines, radial basis neural networks and multiple linear regression. *Analytica Chimica Acta*. 609 (1): 24-36.
20. Mohammadhosseini M., 2012. Chemical profile and antibacterial activity in hydrodistilled oil from aerial parts of *Prangos ferulacea* (L.) Lindl. and prediction of gas chromatographic retention indices by using genetic algorithm multiple linear regressions. *Asian Journal of Chemistry*. 24, 3814-3820.
21. Noorizadeh H., Farmany A., 2010. qsrr models to predict retention indices of cyclic compounds of essential oils. *Chromatographia*. 72 (5-6):563-569.
22. Riahi S., Ganjali M. R., Pourbasheer E., Norouzi P., 2008. QSRR study of GC retention indices of essential-oil compounds by multiple linear regression with a genetic algorithm. *Chromatographia*. 67 (11-12): 917-922.
23. Riahi S., Pourbasheer E., Ganjali M. R., Norouzi P., 2009. Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine. *Journal of Hazardous Materials*. 166 (2-3): 853-859.
24. Parsopoulos K.E., Vrahatis M.N. 2010. Particle Swarm Optimization and Intelligence: Advances and Applications. Information Science Publishing (IGI Global)
25. Mohammadhosseini M., 2013. Novel PSO-MLR algorithm to predict the chromatographic retention behaviors of natural compounds. *Analytical Chemistry Letters*. 3 (4): 226-248.

26. Dewar M.J.S., Zeoblish E.G., Healy E.F., Stewart J. J., 1985. Development and use of quantum mechanical molecular models.76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society*. 107, 3902-3909.
27. Todeschini R., Consonni V. 2000. *Handbook of Molecular Descriptors*. Wiley-VCH. Weinheim, Germany.
28. Olivero J., Garcia T., Payares P., Vivas R., Diaz D., Daza E., Geerliger P., 1997. Molecular structure and gas chromatographic retention behavior of the components of ylang–ylang oil. *Journal of Pharmaceutical Sciences*. 86, 625-630.
29. Heberger K., 1999. Evaluation of polarity indicators and stationary phases by principal component analysis in gas-liquid chromatography. *Chemometrics and Intelligent Laboratory Systems*. 47, 41-49.
30. Wold S., Esbensen K., Geladi P., 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*. 2, 37-52.
31. Heberger K., Milczewska K., Voelkel A., 2005. Principal component analysis of polymer-solvent and filler-solvent interactions by inverse gas chromatography. *Colloids and Surfaces a-Physicochemical and Engineering Aspects*. 260 (1-3): 29-37.
32. Heberger K., Gorgeny M., 1999. Principal component analysis of Kovats indices for carbonyl compounds in capillary gas chromatography. *Journal of Chromatography A*. 845, 21-31.
33. Eberhart R.C., Kennedy J., presented in part at the Proceedings of the 6th Symposium on Micro Machine and Human Science, Nagoya, Japan, 1995.
34. Castillo O., Melin P., 2012. Optimization of type-2 fuzzy systems based on bio-inspired methods:A concise review. *Information Sciences*. 205, 1-19.
35. 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 & Combinatorial Science*. 22 (1): 69-77.
36. Noorizadeh H., Farmany A., Khosravi A., 2010. Investigation of retention behaviors of essential oils by using QSRR. *Journal of the Chinese Chemical Society*. 57 (5A): 982-991.
37. Skoog D., Holler F.J., Nieman T.A. 1998. *Principles of Instrumental Analysis*, Fifth Ed.
38. Cazes J. 2004. *Encyclopedia of Chromatography*, Update Supplement.
39. Todeschini R., Gramatica P., 1997. SD-Modeling and prediction by WHIM descriptors. Part 5: Theory development and chemical meaning of the WHIM descriptors. *Molecular Informatics*. 16, 113-119.
40. Consonni V., Todeschini R., Pavan M., 2002. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences*. 42 (3): 682-692.