Quantitative Structure Property Relationship Modeling for Prediction of Retention index for a set of Essential Oils and Organic Substances of Plant and Animal Origin

J. B. Ghasemi*
Chemistry Department, K.N. Toosi University of Technology, Tehran, Iran
Z.Piravi-vanak
Institute of Standard and Industrial Research of Iran, Tehran, Iran
F. Khavarian
Chemistry Department, Razi University, Kermanshah, Iran

Abstract

Molecular similarity and quantitative structure property relationship (QSPR) analyses have been used to develop compact, robust, and definitive models for essential oils compounds. The QSPR models have been sought to provide an interpretation and characterization of retention index of essential oils. A training set of 66 structurally diverse compounds were selected to be representative of a parent set of 86 compounds and range in measured retention index. In order to evaluate the models, we chose another set with 20 molecules as a prediction set. Descriptors derived from semiempirical (AM1) molecular orbital calculations have been used to construct a QSPR for the retention index, RI, of a series of essential oils. Root mean square error of prediction (RMSEP), average relative error (*REP* %) and R^2 of prediction set for were about 0.011, 0.372 and 0.976, respectively.

***Keywords*:** Essential oils, gas chromatography, Retention index, QSPR, PLS, MLR

Introduction

The life force of a plant is called the essential oil, or 'essence.'[1] Essential Oils are highly concentrated substances the subtle, aromatic and volatile liquids extracted from the flowers, seeds, leaves, stems, bark and roots of herbs, bushes, shrubs and trees through distillation.[2] They are obtained from the plant in various ways, depending upon the nature of the part in which they occur-by compression, by distillation with steam, by dissolving the oils out (extraction) or absorbing them, and by pressure and maceration. There are basically 3 types of essential oils. Steam distilled oils and solvent extracted absolutes.[3] These aromatic plants and oils have been used for thousands of years dating back to ancient civilizations that used them to heal, enhance, soothe and excite the body and spirit[1]. Essential oils are natural mixtures of hydrocarbons (terpenes), oxygen- (alcohols, aldehydes, ketones, carboxylic acids, esthers, lactones) and sulfur-containing (sulfides, disulfides, trisulfides) organic substances of plant and animal origin. [4]

* Correspondence Author: Jahan.ghasemi@gmail.com

The use of essential oils is largely widespread in foods, flavours, deodorants, pharmaceuticals, drinks, cosmetics and medicine and embalming antiseptics especially with aromatherapy becoming increasingly popular.[5] They include whole industries (paint, petroleum, mining and manufacturing), food (processing and flavouring), drink (alcoholic and nonalcoholic flavourings), pharmaceutical products, perfumes and toiletries, hygiene products, and pesticides. The end uses of essential oils are determined by their chemical, physical, and sensory properties, which differ greatly from oil to oil. [6]

Jean Claude Lapraz and Paul Belaiche, found that essential oils have antibacterial, antifungal, antiviral, and antiseptic properties, and that the oils are powerful at oxygenating and carrying nutrients into cells. [3]

The quality and price of some oils are based on the percentage content of a single chemical component, so separation and measurement of individual components is very important. This is usually done using some form of chromatography; the most powerful is gas chromatography using capillary columns. Rigorous identification of components commonly employs a form of spectroscopy (mass, UV, IR, NMR) to indicate the molecular structure. [6]

The classical methods of chromatographic identification of compounds were based on calculation of retention indices by using different stationary phases. The aim of the work was to differentiate essential oils extracted from different plant species by identification their retention indexes. The method of identification was based on the calculation of new retention indices of essential oils compounds fractionated on a polar and non polar chromatographic column with temperature programming system. Several methods using relative retention indices were developed in order to reproduce the identification of compounds in gas chromatography. Generally, the retention values were expressed in relation to standards not present in material characteristics.[5]

A further benefit is that linear retention indices and retention time locked mass spectrometry libraries can be used as additional filters. This approach offers an accessible and powerful tool for characterizing complex mixtures of essential oils in a cost-efficient manner.[7]

C and GC–MS are the main methods for identification of these volatile plant oils. The compounds are identified by comparison of retention indices with those reported in the literature and by comparison of their mass spectra with libraries or with the published mass spectra data. Chromatographic retention for capillary column gas chromatography is the calculated quantity, which represents the interaction between the stationary phase and gas-phase solute molecule. This interaction can be related to the electronic, geometric and topological properties of the molecule. Mathematical modeling of these interactions helps chemists to find a model that can be used to obtain a deep understanding about the mechanism of interaction and to predict the retention indices of new or even unsynthesized compounds. QSPRs, mathematical equations relating chemical properties such as acidity, electrochemistry, reactivity and chromatographic behavior to a wide variety of structural, topological and electronic features of the molecules, have been widely used in the field of chromatographic sciences. Quantitative structure–retention relationships (QSRRs) represent statistical models which quantify the relation between the structure of the molecule and chromatographic retention indices of the compound, allowing the prediction of retention indices of novel compounds.[8] In this paper we report a QSPR model to predict the retention indices of some essential oils using molecular structural descriptors.

Materials and Experimental Detailes

Data set

The data sets of the retention indices were taken from the values reported in some articles.[9-11] Name and the retention indexes (*RI*) of the compounds that used in this study are listed in Table 1.

No.	Compound	LOG RI	No.	Compound	LOG RI
1	Tricyclene	2.967	44	Methyl geranate	3.116
2	a-Pinene	2.969	45	Nonylbenzene	3.118
3	Camphene	2.972	46	Hexyl tiglate	3.124
4	6-Methyl-5-hepten-2-one	2.981	47	Neryl	3.127
5	Sabinene	2.983	48	(*Z*)-b-Damascenone	3.131
6	b-Pinene	2.984	49	(*Z*)-3-Hexenyl	3.132
7	3-Octanol	2.99	50	Eugenol	3.132
8	b-Myrcene	2.991	51	Geranyl	3.134
9	Myrcene	2.996	52	b-Bourbonene	3.137
10	a-Terpinene	3.002	53	Dodecyl	3.137
11	Limonene	3.008	54	4-Nonylphenol	3.138
12	β-Phellandrene	3.009	55	2-nonyl-phenol	3.139
13	p-Cymene	3.011	56	Alpha-Copaene	3.139
14	(Z)- b-Ocimene	3.013	57	b-Elemene	3.139
15	1,8-Cineole	3.014	58	cis-Jasmone	3.144
16	(E)- b-Ocimene	3.016	59	n-Tetradecane	3.146
17	γ-Terpinene	3.019	60	(-)-Caryophyllene	3.146
18	trans-Sabinene	3.02	61	Myristic	3.155
19	Terpinolene	3.031	62	a-Humulene	3.156
20	cis-Sabinene	3.032	63	Benzyl	3.156
21	Linalool	3.033	64	Alloaromadendrene	3.158
22	1-Octen-3-yl	3.039	65	Hexahydrofarnesyl	3.158
23	trans-p-2-menthen-1-ol	3.041	66	Dibutylphthalate	3.159
24	n-Undecane	3.041	67	Hexadecanoic	3.162
25	neo-Alloocimene	3.047	68	Di-iso-octyl	3.163
26	pinocarvone	3.052	69	Seychellene	3.164
27	Terpinen-4-ol	3.062	70	Phytol	3.165
28	Myrtenal	3.064	71	acoradiene	3.166
29	a-Terpineol	3.066	72	Germacrene D	3.166
30	Borneol	3.066	73	Bicyclogermacrene	3.168
31	Lavandulol	3.067	74	Germacrene A	3.171
32	Verbenone	3.068	75	Neryl isobutanoate	3.173
33	p-Cymen-8-ol	3.07	76	(+)-zigma-Cadinene	3.176
34	Cumine	3.08	77	alpha-Bisabolene	3.177
35	Neral	3.085	78	cis-Nerolidol	3.186
36	Geraniol	3.091	79	Spathulenol	3.189
37	Thymol	3.092	80	Geranyl n-butyrate	3.194
38	Geranial	3.094	81	trans-Nerolidol	3.194
39	Bornyl	3.101	82	10-epi- g-Eudesmol	3.204
40	Piperitenone	3.102	83	cis-Methyl	3.217
41	Perilla	3.103	84	Heptadecane	3.23
42	(+)-p-Menth-1-en-9-ol	3.104	85	(*E,E*)-Farnesol	3.237
43	Methyl	3.11	86	Eicosane	3.301

Table 1. Essential oils and their retention index value

Analytical gas chromatography was carried out using a Thermoquest 2000 GC system with a DB-1 capillary column (30 m m,0.25 mm; 0.25 mm film thickness). The carrier gas (Mobile Phases), ionization energy performed, injector temperature, flow rate and temperature programming for three capillary columns are reported in Table 2.

		A	B	C
1	Mobile Phases	Helium	Helium	Helium
2	Flow rate	1.5 mL/min	1 mL/min	1 mL/min
3	Column	DB-1 capillary column	CP-Sil 8 CB	SGE-BPX5MS fused silica
4	Temperature programming	50-260°C with a 2.5°C/min rate		50-150°Cat 3 °C/min rate
5	Detector	FID	MS	DSQ/A1300,(E.I Quadrapole)
6	Data processing system	Thermoquest 2000 GC	HP5971mass spectrometer	Thermofinnigan Trace GC/Trace
7	Ionizationenergy performed	at 70 eV	at 70 eV	at 70 eV
8	Injector temperature		250 ºC	220 °C

[A]Ref. 9, [B]Ref. 10, [C]Ref. 11

Table 2. Details of used three capillary column chromatography

Computer Hardware and Software

All calculations were done by the following software's:
ChemDraw Ultra version 9.0 (ChemOffice 2005, CambridgeSoft Corporation) software was used for drawing the molecular structures.[12] Conversion of 2D structures into 3D structures was performed with chem3D of ChemOffice Ultra version 9.0 and optimizations of molecular structures were done by the same software with MM2 and MOPAC by theory method AM1 with minimum RMS Gradient 0.100. ChemOffice linked to Excel for generating structural descriptors. SPSS ver. 11.5 software was used for variable selection and other calculations were done in the MATLAB (version 7.0, Mathworks, Inc.) environment.

Calculation of Descriptors

Due to the diversity of the molecules studied in this work, 44 different descriptors are chosen and listed in Table 3. These parameters encode different aspects of the molecular structure, and consist of twenty three steric (1-23), fifteen thermodynamic (24-38), and six electronic (39-44) descriptors. To avoid from the standard error in geometry optimization of each molecule during optimizing process we optimize them with MOPAC and MM2 techniques several times from different starting point geometry, until root mean square (RMS) gradient values becomes smaller than 0.001 kcal mol^{-1}. Then the conformation with the lowest energy of each molecule was considered for calculation of the electronic properties.

No.	Descriptors	Notation	Group
1	Balaban Index	Bindx	Steric
2	Cluster Count	ClsC	Steric
3	Connolly Accessible Area	SAS	Steric
4	Connolly Molecular Area	MS	Steric
	Connolly Solvent-Excluded		
5	Volume	SEV	Steric
6	Diameter	Diam	Steric
7	Exact Mass	Mass	Steric
8	Molecular Topological Index	Tindx	Steric
9	Molecular Weight	MW	Steric
10	Number Of Rotatable Bonds	NRBo	Steric
11	Ovality	Ovality	Steric
12	Polar Surface Area	PSAr	Steric
13	Principal Moment of Inertia – X	PMIX	Steric
14	Principal Moment of Inertia – Y	PMIY	Steric
15	Principal Moment of Inertia –Z	PMIZ	Steric
16	Radius	Rad	Steric
17	Shape Attribute	ShpA	Steric
18	Shape Coefficient	ShpC	Steric
19	Sum Of Degrees	Sdeg	Steric
20	Sum Of Valence Degrees	SVDe	Steric
21	Total Connectivity	Tcon	Steric
22	Total Valence Connectivity	TVCon	Steric
23	Wiener Index	Windx	Steric
24	Boiling Point	BP	Thermodynamic
25	Critical Pressure	Pc	Thermodynamic
26	Critical Temperature	Tc	Thermodynamic
27	Critical Volume	Vc	Thermodynamic
28	Heat of Formation	HF	Thermodynamic
29	Henry's Law Constant	H	Thermodynamic
30	Ideal Gas Thermal Capacity	Cp	Thermodynamic
31	LogP	LogP	Thermodynamic
32	Melting Point	MP	Thermodynamic
33	Molar Refractivity	MR	Thermodynamic
34	Molar Refractivity	MR1	Thermodynamic
35	Partition Coefficient (Octanol/Water)	CLogP	Thermodynamic
36	Standard Gibbs Free Energy	G	Thermodynamic
37	Vapor Pressure	VP	Thermodynamic
38	Water Solubility	Sol	Thermodynamic
39	DipoleLength	DPLL	Electronic
40	ElectronicEnergy	ElcE	Electronic
41	HOMO Energy	Homo	Electronic
42	LUMO Energy	Lumo	Electronic
43	Repulsion Energy	NRE	Electronic
44	Total Energy	TotE	Electronic

Table 3 . Descriptors categories

Selection of Descriptors

All descriptors with zero and/or constant values for all the molecules in the data set were eliminated. The correlation matrix was calculated for all of descriptors, one of the two descriptors which has the pair wise correlation above 0.94 (R^2 >0.94) and it has a large correlation with the other descriptors was eliminated. By the correlation matrix 15 descriptors was eliminated from primary ones. The stepwise regression method was used as the variable selection method to select the suitable descriptors among 29 theoretical descriptors generated by Chemoffice software. The number of variables retained in the model is based on the levels of significance assumed for inclusion and exclusion of variables from the model.[13, 14] The descriptors that were remained including boiling point (BP) standard Gibbs free energy (G) and lowest unoccupied molecular orbital (LUMO). The numerical values of these descriptors are reported in Table 4.

No.	Compands	BP	Lumo	G
1	a-Pinene	430.425	1.265	149.85
2	b-Pinene	423.978	1.325	182.6
3	Myrcene	429.399	0.492	272.12
4	p-Cymene	451.514	0.527	133.66
5	gama-Terpinene	442.539	1.239	103.7
6	Linalool	477.226	1.186	58.85
7	Terpinen-4-ol	485.022	1.369	-66.65
8	a-Terpineol	483.155	0.804	-20.5
9	p-Cymen-8-ol	501.948	0.502	2.12
10	Nonanoic	535.53	1.296	-319.01
11	Perilla	514.367	1.232	20.57
12	Decanoic	551.922	1.295	-310.59
13	2-nonyl-phenol	515.255	1.275	242.55
14	(-)-Caryophyllene	536.655	1.487	161.93
15	Alloaromadendrene	517.222	1.418	257.93
16	Acoradiene	526.696	1.29	234.94
17	Germacrene A	543.026	1.04	171.42
18	alpha-Bisabolene	545.014	1.064	236.49
19	trans-Nerolidol	565.089	1.166	172.62
20	Heptadecane	565.834	3.29	92.26

Table 4. The numerical values of descriptors for the prediction set.

Ordinary Least Squares Regression modeling

The calculated descriptors were collected in a data matrix X with a dimension of ($m \times n$), where m and n are the number of molecules and the number of descriptors, respectively. The linear relationship between *RI* and calculated descriptors was obtained through multiple linear regression analysis using more molecules as the calibration samples. The stepwise selection and elimination of variable procedure of SPSS software was employed to select the most relevant set of descriptors. It should be noted that through a typical stepwise regression run, SPSS produces many models ranking them based on calibration correlation coefficient where some of them may be over-fitted.[8] In order to test the final model performances, 20 molecules out of 86 molecules were selected as external test set molecules. These samples were selected based on the both property and descriptors spaces. To do so, the

data matrix joining descriptors and retention index was subjected to principal component analysis (PCA). The first three principal components, explained 76.3% of variances. Distribution of this value between these three PC are in this manner; PC_1=53.69, PC2=12.87 and PC3=9.72.

Results and discussion

Retention in GC is the result of competitive solubility of the solute between the mobile and the stationary phases. The molecular structure and chemical properties of the solute determine the type and the extent of the interaction of the solute with these phases. The differences between these properties govern the retention behavior through the column[8]. Aim of the work is the development of a mathematical model that uses molecular descriptors, x_j with j = 1 . . . p, as input variables (features) and is capable of producing an output, I, that is a good estimation of the corresponding experimental retention index, I. A linear model is given by:

$$I = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p \quad (1)$$

with b_j being the regression coefficient for descriptor j, and b_0 the intercept. The regression methods compared are, multiple linear regression, (MLR: ordinary least squares regression), and partial least squares regression. (PLS)[15] PLS decreases the number of independent variables in a special way. This technique constructs a set of linear combinations of the input variables for regression and has been developed primarily for prediction. [16]

Van den Dool and Kratz proposed a generalization of the retention index system including linear temperature-programmed gas chromatography as follows:

$$I_x = 100\ [\frac{t_x - t_n}{t_{n+1} - t_n} + n\ ] \quad (2)$$

where I_x is the temperature-programmed retention index, t_n, t_{n+1} and t_x the retention time (in minute) of the two n-alkanes containing n and n + 1 carbons and of the compound of interest, respectively. This relation show direct relationship between retention index and retention time and number of carbons too (n) . [5]

Evaluation of regression models

Model development and validation is a critical problem in QSPR studies. Because of the great effort required to measure the RI of a large number of compounds, a variety of methods have been proposed to estimate or predict the RI, either directly from physical properties or from quantitative structure–property relationship (QSPR) models.

Calibration model is built by using known property data, which in some instances need to select the most relevant set of descriptors from the pool of calculated descriptors. The calibration model is then evaluated for prediction and generalization. A successful calibration model must have an ability to predict not only the property of calibration molecules (internal validation) but also of the external sources (external validation). To do so, the data are generally splitted into two sets including calibration set (or training) and prediction set (or validation).

Sometimes an extra data set, named external test set, is also used. The calibration and prediction sets are used in the model development steps and the overall prediction ability of the model is evaluated by application to predict the property of the external test set. Cross-validation is another tool to evaluate the model performance and generalization. Therefore, QSPR models are derived from the initial partitioning of compounds, and consequently data splitting influences the performances of the developed model.[8]

Root mean square of errors (RMSEs), that calculated for the prediction sets, are reported to indicate the predicted accuracy of models, which is calculated by the root square

of the sum of squared errors in prediction divided by their total number. The orthogonality of the descriptors in the model was established through variance inflation factor (VIF)[17, 18]. The VIF is defined as $1 / (1 - r_i^2)$ where r_i is the multiple correlation coefficient for the ith variable regressed on the p -1 others, p is being the number of variables contributed to the model. VIF value larger than 5 indicates that the information of the descriptors may be hidden by the correlation of the other descriptors. [19]

The predictive power of the regression model developed on the selected training set is estimated on the predicted values of prediction set chemicals, by the internal Q^2 that is defined:

$$Q_{\text{int}}^{\ 2} = 1 - \frac{\sum_{i-1}^{pred} (y_i - \hat{y}_i)^2}{\sum_{i-1}^{pred} (y_i - \hat{y}_{tr})^2}$$

where y_i and $\hat{y}_i$ are the measured and predicted (over the prediction set) values of the dependent variable, respectively. $\hat{y}_{tr}$ is the averaged value of the dependent variable for the training set; the summations cover all the compounds in the prediction set[20].

To Comparison between calculated log RI values for external prediction set with experimental data we calculate their residuals and REP in order to signification stability of models (MLR, PLS) that are shown in

No.	Exp.	MLR Model			PLS Model		
	(log RI)	Pred. (log RI)	Residual	REP (%)	Pred. (log RI)	Residual	REP (%)
1	2.969	3.014	0.045	1.516	2.992	0.023	0.789
2	2.984	3.01	0.025	0.871	2.988	0.004	0.132
3	2.996	3.02	0.024	0.801	2.997	0.001	0.026
4	3.011	3.033	0.022	0.731	3.012	0	0.013
5	3.019	3.024	0.005	0.166	3.004	-0.015	-0.485
6	3.033	3.061	0.028	0.923	3.051	0.017	0.568
7	3.062	3.059	-0.003	-0.098	3.047	-0.015	-0.488
8	3.066	3.058	-0.008	-0.261	3.042	-0.024	-0.778
9	3.07	3.081	0.011	0.358	3.07	0	0
10	3.087	3.095	0.008	0.259	3.088	0.001	0.025
11	3.103	3.103	0	0	3.104	0.001	0.023
12	3.104	3.116	0.011	0.387	3.115	0.01	0.324
13	3.139	3.126	-0.013	-0.414	3.138	-0.001	-0.022
14	3.146	3.145	-0.001	-0.032	3.163	0.017	0.525
15	3.158	3.131	-0.028	-0.855	3.145	-0.013	-0.411
16	3.166	3.139	-0.026	-0.853	3.155	-0.011	-0.339
17	3.171	3.151	-0.02	-0.631	3.167	-0.004	-0.117

Table 5continued…

18	3.177	3.16	-0.017	-0.535	3.18	0.003	0.089
19	3.194	3.179	-0.016	-0.47	3.204	0.009	0.291
20	3.23	3.185	-0.045	-1.393	3.225	-0.004	-0.136

Table 5. Experimental and calculated log RI values for external prediction set.

Retention index in GC is the result of competitive solubility of the solute between the mobile and the stationary phases. The quality of retention index is depending on the molecular structure and chemical properties of the solute and the differences between these properties govern the retention behavior through the column. Molecular descriptors define the molecular structure and physicochemical properties of molecules by a single number. A wide variety of descriptors have been reported for using in QSAR/QSPR analyses. The electronic descriptors such as *E*HOMO, *E*LUMO and dipole moments have been derived from AM1 calculations. This type of descriptors will help to identify the specific interactions between polar stationary phase and different fragments of the molecules. More than 44 descriptors were calculated for each molecule. After preprocessing of the data and elimination of constant or collinear variables, 29 descriptors were remained and used for future analyses. The co-linearity threshold in QSAR/QSPR studies is usually considered 0.9, i.e. descriptors with $R^2 > 0.94$ are selected as collinear. The model with the highest Q^2 was selected as optimum. The prediction ability of the models was measured by relative error of prediction (REP) using Eq. (4).[8]

$$REP \quad (\%) \quad = 100 \quad \times [\frac{\sum_{j=1}^{N} (\hat{y}_j - y_j)^2}{\sum_{j=1}^{N} (y_j)^2}] \qquad (4)$$

MLR Analysis

MLR method provides equation linking the structural features to the property of the compounds for predicting the property of interest is in the form of the following equation:

$$P_j = P_o + \sum_{i=1}^{n} c_i X_{ij} + e_i \qquad (5)$$

where P_j is the predicted value of the property for a given compound j, P_0 is the intercept coefficient, c_i are the descriptor coefficients, X_{ij} are the descriptor values and n is the number of descriptors. P_0 and P_j are determined by using the least-squares method the residual error terms (ei) are the differences between the predicted and observed logRI.

In Figures 1a and 1b, we draw predicted log RI *versus* experimental log RI obtained by the MLR and PLS modeling. These plots show clearly that distribution of data is normal without systematic error because they distribute from the straight trend quite randomly. Figures 2a and 2b draw the residuals as a function of the experimental data (log RI).

.

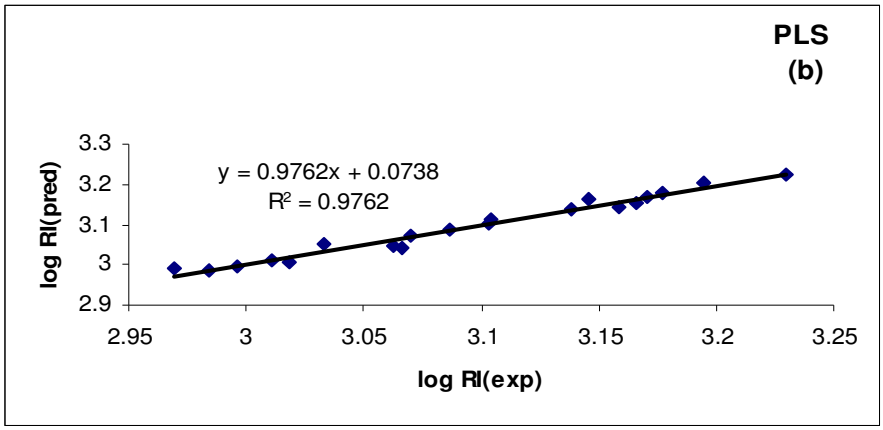


Fig.1

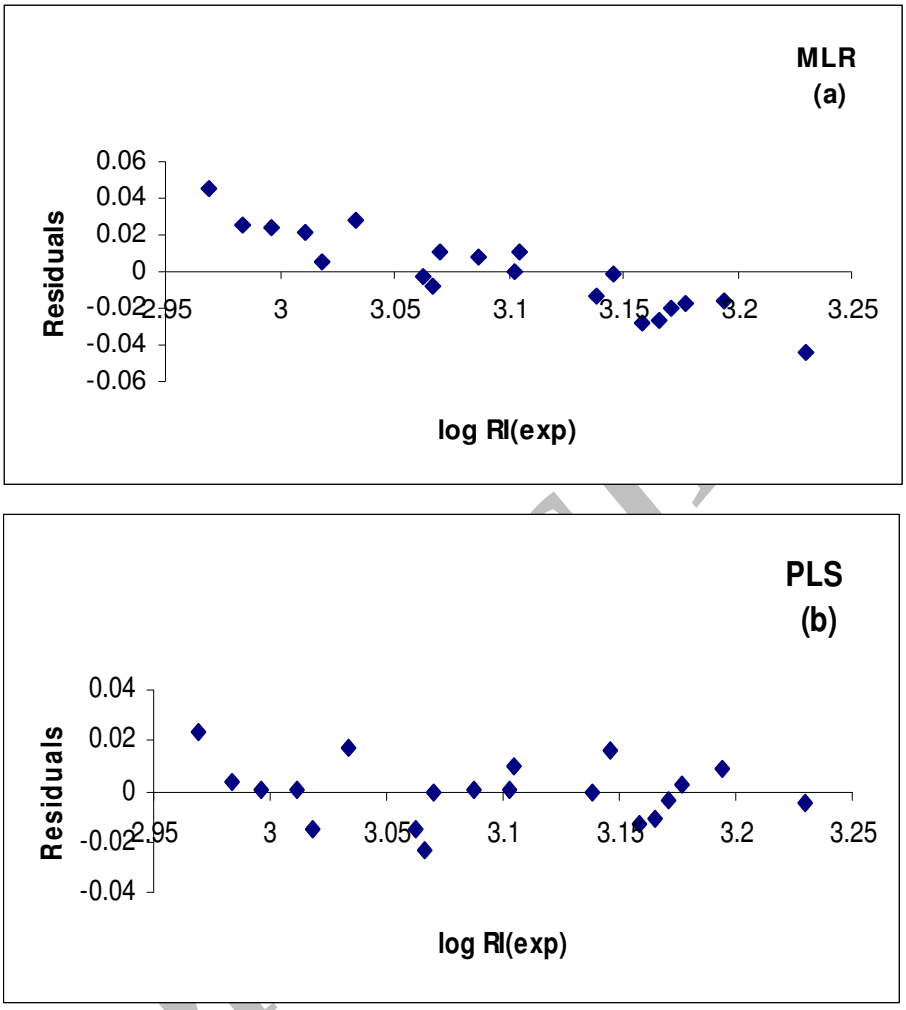
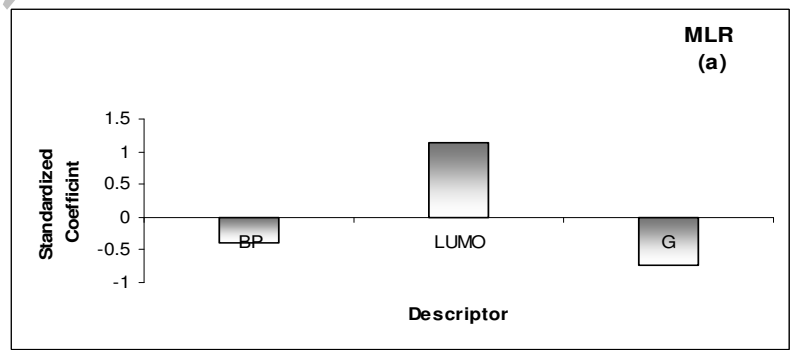


Fig. 2.

The agreement observed between the predicted experimental values in Fig. 1a and b and the random distribution of residuals about zero mean in Figures 2a and 2b confirms the good predictive ability of MLR and PLS modeling. Figures 3a and 3b show the standardized regression coefficient reveals the significance of an individual descriptor presented in the regression model. This plot shows the strength of selected descriptors found in a model. The greater the absolute value of a coefficient, the greater the weight of the variable in the model**.**



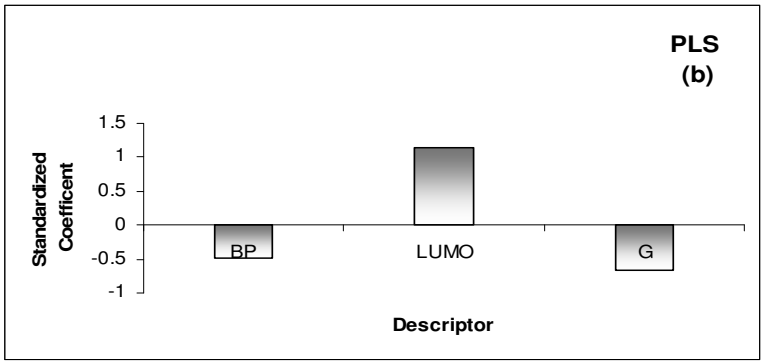


Fig. 3.

Correlation between these variables and response as correlation matrix of measured data are given in Table 6. It is found that BP has a high correlation to the response with the correlation coefficient equal to 0.88. Applying MLR to the data, gave a model with R^2= 0.973 REP%=0.15661 SEP=0.02223 and Q^2_{int}=0.915. The corresponding figures for PLS were R^2=0.976 REP%= 0.0832 SEP= 0.0118 and Q^2_{int}=0.976. The model parameter value, standardized coefficient and mean effect of MLR model are presented in Table 7. The mathematical model that generated by multiple linear regression (MLR) is :

	LOG RI	*BP*	*Lumo*	*G*
LOG RI	1			
BP	0.882	1		
Lumo	0.295	0.117	1	
G	0.014	-0.347	0.007	1

Table 6. Correlation matrix for the dependence of log

.

	Model parameters		Standardized coefficients	VIF
Source	Value	Standard error	Value	value
Intercept	2.4687	0.0409		
BP	0.00121	7.70E-05	0.9957	1.3095
LUMO	0.0064	0.00045	0.0803	1.0531
G	9.89E-05	2.78E-05	0.2223	1.2934

Table 7. Model parameters value and standardized coefficients and mean effect for MLR model

Log (RI) = 2.4686 **+** 0.00121 **BP** + 0.0063 **LUMO** + 9.894e-05 **G** (6)

The agreement between experimental and predicted values, high correlation coefficient, low RMSEP in both model (PLS, MLR), and random distribution of residuals about zero confirms the good predictive ability of MLR and PLS modeling. All of these statistical parameters and their values are present in Table 8.

Parameter	MLR	PLS
RMSEP	0.022	0.012
SEP	0.022	0.012
REP%	0.157	0.083
R^2_{pred}	0.973	0.976
Q_{int}^2	0.916	0.976
N LV[1]	_	2
N DS[2]	3	3

[1]Number of latent variables.

[2]Number of descriptors.

Table 8. Statistical parameters of MLR and PLS models

In Figure 4 we plot PREES (The optimum number of factors was concluded as the first local minima) versus PC number. This figure shows number of principle component through the sharply deviation seen in curve.

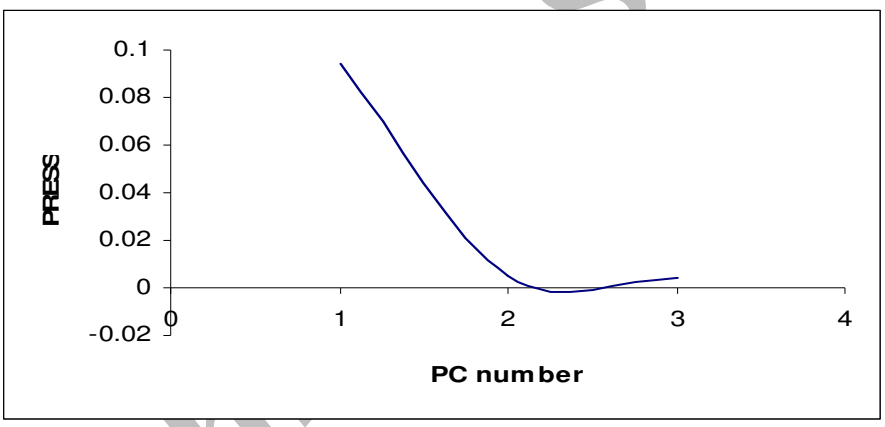


Fig. 4. PRESS vs. number of factors for the PLS model.

Interpretation of Descriptors

Retention is a phenomenon that is primarily dependent on the interactions between the solute and the stationary phase molecules. The forces associated with these interactions can be related to the geometric and topological structures and electronic environments of the molecule. Quantitative structure–property relationships (QSPRs) have been demonstrated to be a powerful tool for the investigation of the chromatographic parameters. QSPRs have been used to obtain simple models to explain and predict the chromatographic behavior of various classes of compounds. [13] The lowest unoccupied molecular orbital energy (LUMO) is electronic descriptor. In particular, electronic parameters are considered important in the establishment of QSAR models and are helpful to quantify different types of intermolecular and intramolecular interactions, as these interactions are usually responsible for properties of chemical and biological systems. The transfer of a pair of electrons from the HOMO to the LUMO is, by definition, a reaction between a Lewis acid and a Lewis base. Thus, the parameter LUMO is a measure of the ability of a molecule to interact with the π and n-electron pairs of the other molecules. The reduction in energy in molecular orbital is the driving force for chemical bond formation[21] According to Frontier Orbital Theory, the shapes and symmetries of the highest-occupied and lowest unoccupied molecular orbitals

(HOMO and LUMO) are crucial in determining the chemical reactivity of a species and the stereochemical and regiochemical outcome of a chemical reaction. The energies of the highest-occupied and the lowest-unoccupied molecular orbitals (HOMO/LUMO energies) are frequently used quantum chemical descriptors. As a consequence, the derived QSAR models will include information regarding the nature of the intermolecular forces involved in determining the biological activity of the compounds in question. HOMO energy in particular has been identified as being of significant value to QSAR studies.[22] Molecules with low LUMO energy values are more able to accept electrons than molecules with high LUMO energy values. The LUMO energy value is increased with the presence of electron donating groups (EDGs).this remark also agrees with Kemnitzer et al. who recommended the introduction of EDGs such as NMe_2, NH_2, NHEt, and OMe. [23] The boiling point (BP) of a compound is predetermined by the intermolecular interactions in the liquid and by the difference in the molecular internal partition function in the gas phase and in the liquid at the boiling temperature. Therefore, it should be directly related to the chemical structure of the molecule. Various rules and formulas were proposed early on to correlate boiling points of homologous hydrocarbons with the number of carbon atoms or molecular weight. [24] The normal boiling points of liquids reflect the strength of the intermolecular forces (among other forces present) that hold them together. The stronger the intermolecular forces, the more tightly the atoms will be held together and, therefore, the higher the normal boiling point. The boiling point can be directly correlated to the chemical structure of a molecule. Quantitative structure–property relationships (QSPR) and quantitative structure–activity relationships (QSAR) methodology has been reported quite extensively in the literature to predict many physicochemical properties, such as vapor pressures, chromatographic retention and capillary electrophoretic mobilities, aqueous solubility and boiling points[25]. The column temperature is an important factor and variable that must be controlled with a precision about 1/10 C°. Optimum column temperature is depending on sample boiling point and degree of desired separation and resolution. In general, optimum separation is relevant to minimum temperature. In spite of this, use low temperature makes increasing in time elution. So we need more time to complete the analysis. [26]

It seems that retention index and boiling point have a direct relationship with each other. As we see in the Table 5 boiling point have a high correlation with retention index (Figure 5). It means that boiling point has an important and major effect on the retention index and retention index will increase if boiling points increase. b-Pinene is cyclic and myrcene has a linear structure. Compounds with a linear structure because of their more available surface for interaction with each other have a high boiling point and retention index's too in comparison with cyclic ones. We have an increasing of boiling point with increasing in molecular weight and number of carbon atoms. Whatever molecules with high molecular weight and big volume, the retention index increase respectively. Another parameter that has an important effect on boiling point is polarizability and functional groups. p-cymen-8-ol has a –OH group and polarized in comparison with p-cymene. Because of strong interaction and H-bonding between these kinds of molecules they have high boiling point and therefore high retention index. The strictures of these two molecules are show in Figure 6. In comparison between p-Cymene and perilla alcohol molecules we see that in perilla alcohol because of it's –OH functional group, molecule is polar and if interaction with stationary phase was more important it must be exit faster than p-cymene because the stationary phase (DB-1) is non polar. [27] But in reality we see that perilla alcohol exit later and has a high retention index because it has a high boiling point and intermolecular interaction is more important rather than interaction with stationary phases (Figure 7). So we can conclude that boiling point has a more effect and importance in comparison with kind of stationary phases and their interactions that is according to reality. Boiling point (BP) is perhaps the best one from among

the physicochemical properties in describing chromatographic retention. [16] In general the most important parameter in GC is boiling point.

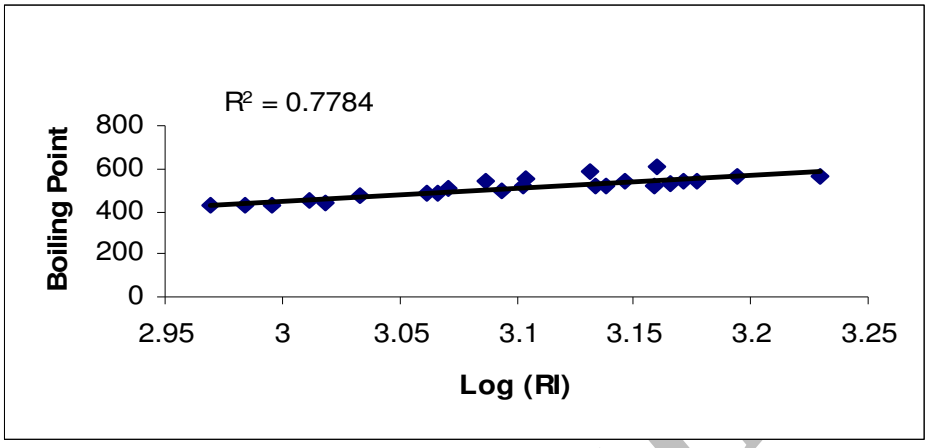
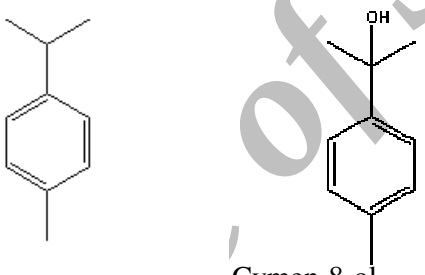


Fig. 5. Correlation between Log (RI) and Boling Point.



Cymen-8-ol.

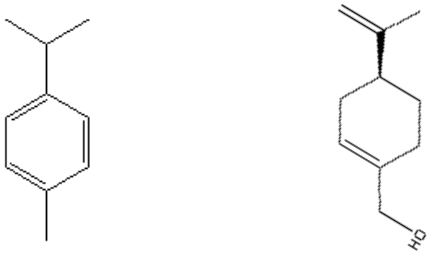


Fig. 7 p-Cymene Perilla alcohol

The empirical characterization of the temperature dependence of the retention index has been reported in a number of studies [28-33]. These studies incorporate thermodynamic models of retention into the retention index equation. Since the retention index system is entirely thermodynamic in origin, small departures from temperature independent behavior are typically explained through the temperature dependence of the individual enthalpic and entropic terms which make up the retention index equation. Although basic understandings of the governing principles of retention exist for GC, there is an incomplete understanding of how changes in physical conditions affect relative retention. For the case of partitioning of

non-polar solutes between a vapor phase and a non-polar solvent, it is well known that the Gibbs free energies for the transfer from the vapor phase to the liquid phase increase with increasing temperature, but may do so at different rates.

The partition coefficients for the gas-to-liquid transfer of the analytes can be determined directly from the ratios of the average analyte number densities in the coexisting phases:

$$K_{trans} = \frac{(\rho_{liquid})}{(\rho_{gas})} = \exp[-\Delta G_{trans} / RT] \qquad (7)$$

where R and T are the molar gas constant and absolute temperature, respectively. The number densities are mechanical properties, and hence the partition coefficients and Gibbs free energies of transfer can be determined more precisely following the Gibbs ensemble route than using standard free energy evaluation techniques.

The Gibbs free energy of transfer can be separated into enthalpic and entropic contributions. The enthalpy of transfer itself contains two terms: the internal energy of transfer and the pressure-volume term.

Since the enthalpy of transfer is calculated independently from the partition coefficients, the entropy of transfer at every specific temperature can be determined from the usual thermodynamic relation:

$$\Delta G\,(T) = \Delta H - T\,\Delta S \qquad (8)$$

In chromatography experiments, absolute free energies are rarely measured, since the phase ratio is very difficult to determine, and varies significantly from column to column and also with temperature. Relative free energies, though, do not depend on the phase ratio, and are very reproducible. The calculation of the enthalpic and entropic contributions to the Gibbs free energies of transfer allows us to rationalize the observed temperature dependence for the retention indices of the compounds. Using the standard thermodynamic equation for the Gibbs free energy given in Eq. (8), and assuming that the enthalpy and entropy are constant (see below), the temperature dependence of the Gibbs free energy is larger for molecules with larger entropies of transfer. [34]

Conclusion

In this study we used multiple linear regression (MLR) and PLS with leave-one-out cross-validation techniques to model and predict retention index of a large set of essential oils like a-pinene, camphene, sabinene. Both methods resulted in useful models with good generalization and prediction ability to predict the retention indices of a separate test set compounds. The method presented here enables an automatic estimation of retention indices from the molecular structure, using a model derived from about 66 relevant compounds. The identification of unknowns in GC–MS analyses can be supported by excluding hit list structures which give predicted retention indices very different from the experimental values; thus the identification of unknown RI is facilitated. This model is making up four descriptor including boiling point (BP) standard Gibbs free energy (G) and lowest unoccupied molecular orbital energy (LUMO). The linear model produced by MLR and PLS methods could reproduce more than 97% of variances in the retention data with prediction error as low as 0.0832%.

References:

1. Essential Oils History, http://www.artesianspas.com/main/aromahistory.asp
2. Hauck, D.W., History of Essential Oils, http://www.crucible.org/oils-history.htm

3. Essential Oils Defined. What are essential oils? http:// www. healthysecrets. com/ herbal_ oils/essentialoils.html
4. Voda, K., Boh, B., Vrtacnik, M., *Int. Biodeterior. Biodegrad*, **51**, 51 (2003).
5. H´erent, M.F., Bie, V.D., and Tilquin, B., *J. Pharm. Biomed. Anal*., **43**, 886 (2007).
6. Essential oils and their production http://www.crop.cri.nz/home/products-services/ publications/ broadsheets/039essentialoils.pdf
7. Namara, K.M., Howell, J., Huang, Y., Jr, A. R., *J. Chromatogr. A*., **1164**, 281 (2007).
8. Hemmateenejad, B., Javadnia, K., and Elyasi, M., *Anal. Chem. Acta.*, **592**, 72 (2007).
9. Cakir, A., Kordali, S., Kilic, H., and Kaya, E., *Biochem. Syst. Ecol.*, **33** (3), 245 (2005).
10. Judþentienë, A. and Mockutë, D., *Chemija*, **15**, 64 (2004).
11. Monsef-Esfahani, H. R., Karamkhani, F., Nickavar, B., Abdi, K., and Faramarzi, M. A.., *Chem. Nat. Compd.,* **43** (1) (2007).
12. ChemOffice 2005, CambridgeSoft Corporation, http://www.cambridgesoft.com/
13. Jalali-Heravi, M., and Fatemi, M.H., *J. Chromatogr. A.*, **915**, 177 (2001).
14. Ghasemi, J., and Saaidpour, S., J. Surf. Det., to be published.
15. Garkani-Nejad, Z., Karlovits, M., Demuth, W., Stimpfl, T., Vycudilik, W., Jalali-Heravi, M. and Varmuza, K., *J. Chromatogr. A.,* **1028**, 287 (2004).
16. Farkas, O., Héberger, K., and Zenkevich, I. G.**,** *Chomom. Intell. Lab. Syst*., **72** (2), 184 (2004).
17. Chatterjee, S., Hadi, A., and Price, B., *Regression Analysis by Examples*, third ed. Wiley-VCH, New York (2000).
18. Shapiro, S., and Guggenheim, B., *Quant. Struct. Act. Relat*., **17**, 327 (1998).
19. Cho, D.H., Lee, S.K., Kim, B.T., and No, K. T, *Bull. Kor. Chem. Soc.*, **22**, 388 (2001).
20. Ghasemi, J., Abdolmaleki, A., Asadpour, S., and Shiri, F., *QSAR & Comb. Sci., DOI*: 10.1002/qsar.200730022.
21. Héberger, K., and Kowalska, T. , *Chomom. Intell. Lab. Syst*., **47** (2), 205 (1999).
22. Melagraki, G., Afantitis, A., Sarimveis, H., and Koutentis, P.A., *J. Markopoulos, O. Igglessi-Markopoulou, Bioorg. Med. Chem.*, 15, 7237 (2007).
23. Afantitis, A., Melagraki, G., Sarimveis, H., and Koutentis, P.A., *J. Markopoulos, O. Igglessi-Markopoulou, Bioorg. Med. Chem.*, **14**, 6686 (2006).
24. Katritzky, A. R., Mu, L., and Lobanov, V. S., *J. Phys. Chem*., **100**, 10400 (1996).
25. Lia, Q., Chen, X., and Hu, Z., *Chomom. Intell. Lab. Syst*., **72**, 93 (2004).
26. Douglas A., Skoog, *Fundamentals of Analytical Chemistry*, Sixth edition, Saunders College Publishing (1992).
27. Bulska, E., Emteborg, H., Baxter, D.C., Frech, W., Ellingsen, D. and Thomassen, Y., *Analyst*, **117**, 657 (1992).
28. Budahegyi, M. V., Lombosi, E. R., and Lombosi, T. S., Mészáros, S. Y.,. Nyiredy, S. Z, Tarján, G., Timár, I., and Takács, J. M., *J. Chromatogr.,* **271**, 213(1983).
29. Takács, J., Rockenbauer, M., and Olácsi, I., *J. Chromatogr.* 42, 19 (1969).
30. Takács, J., *J. Chomatogr,* **799** A, 213(1998).
31. Gonzalez, F.R., and Nardillo, A.M., *J. Chromatogr,* **842**(A), 29 (1999).
32. Tudor, E., *J. Chromatogr* A*,* **858**, 65 (1999).
33. Tudor, E., *J. Chromatogr* A*,* **859** , 49 (1999).
34. Wick, C. D., Siepmann, J. I., Klotz, W. L., and Schure, M. R., *J. Chromatogr* A*,* **954**, 181 (2002).