Research Article

Advanced Journal of Chemistry-Section A, 2018, 1(1), 12-31 http://ajchem-a.com

Prediction of two-dimensional gas chromatography time-of-flight mass spectrometry retention times of 160 pesticides and 25 environmental organic pollutants in grape by multivariate chemometrics methods

Issa Amini[a], Kaushik Pal[b], Sharmin Esmaeilpoor[a,*], Aydi Abdelkarim[c]

[a]Department of Chemistry, Payame Noor University, Tehran, PO BOX 19395-4697, Iran.

[b]Department of Nanotechnology, Bharath University,BIHER Research Park, Chennai, Tamil Nadu 600073, India

[c]Department of Chemical and Materials Engineering, College of Engineering, National College of Chemical Industry, Nancy, Polytechnic Institute of Lorraine, France Frankfurt Am Main Area, Germany.

*E-mail address: sharminesmaeilpoor@yahoo.com, Corresponding author: Tel.: + 989188413709

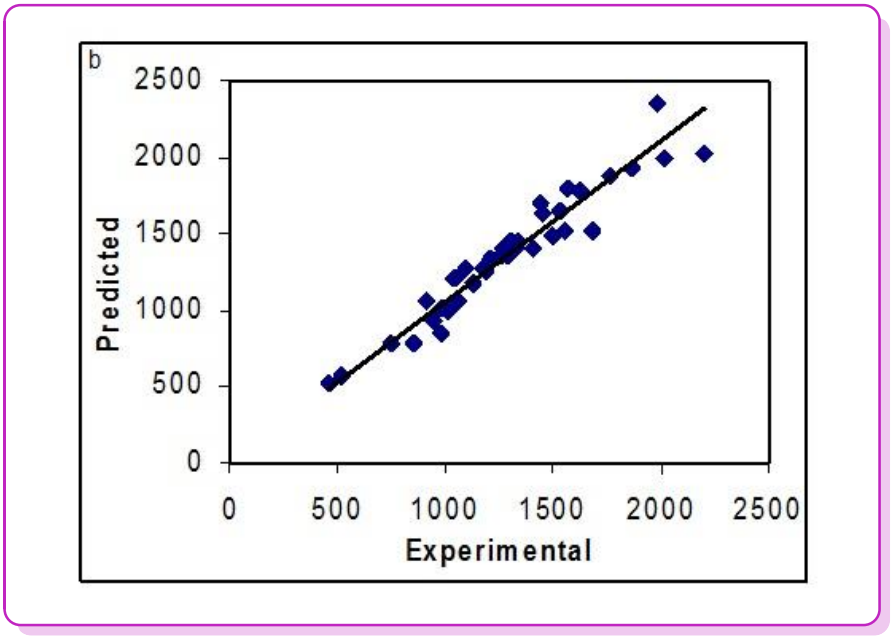
Received: 28 September 2018, Revised: 20 October 2018, Accepted: 10 November 2018

A B S T R A C T

A quantitative structure–retention relation (QSRR) study was conducted on the retention times of 160 pesticides and 25 environmental organic pollutants in wine and grape. The genetic algorithm was used as descriptor selection and model development method. Modeling of the relationship between the selected molecular descriptors and retention time was achieved by linear (partial least square; PLS) and nonlinear (kernel PLS: KPLS and Levenberg-Marquardt artificial neural network; L-M ANN) methods. The QSRR models were validated by cross-validation as well as application of the models to predict the retention of external set compounds, which did not have contribution in model development steps. Linear and nonlinear methods resulted in accurate prediction whereas more accurate results were obtained by L-M ANN model. The best model obtained from L-M ANN showed a good R^2 value (determination coefficient between observed and predicted values) for all compounds, which was superior to those of other statistical models. This is the first research on the QSRR of the compounds in wine and grape against the retention time using the GA-KPLS and L-M ANN.

Keywords: Grape, Wine, Pesticide residue, organic pollutants, GC×GC–TOFMS, Quantitative structure–retention relationships, Kernel partial least square, Levenberg-Marquardt artificial neural network.

GRAPHICAL ABSTRACT



1. Introduction

Wine has become a significant beverage in many nations around the world, which has relatively low alcohol content (13–16°), accepted and consumed by all ages. Wine not only has rich nutrition, but also gives people enjoyment [1]. Wine, as a good source of polyphenols has received attention, largely due to its in vitro inhibitory effect on low-density lipoprotein (LDL) oxidation. Under the generic term of ‘‘wine’’, there is a diversity of quality which is quite unique among the products and determined mainly by interaction between grapes, yeasts and technology. Wine is a natural product resulting from a number of biochemical reactions, which begin during ripening of the grapes and continue during harvesting, throughout the alcoholic fermentation, clarification and after bottling. Many of these reactions are left to nature and microorganisms present on the grapes [2].

Monitoring of contaminant residues in wine is essential because diversified kinds of pesticides are frequently applied in viticulture and their residues in grapes might withstand the fermentation process and find their presence in wine. In addition to pesticides, commodities such as grapes and wines may get contaminated by the persistent environmental contaminants. In case of wine, due to the presence of some lipophilic components, the corks might attract environmental pollutants which in turn might contaminate the wines. In grapes, which are the starting material for wine, the residues of pesticides may appear from both direct applications as well as indirect sources like contaminated agro-inputs, drift from adjoining fields, etc. On the other hand, the dioxin-like polychlorinated biphenyls (PCBs) might appear in wine through exposure to environment and undergo biomagnification because of their lipophilicity. These compounds can cause reproductive and developmental toxicity, immunotoxicity,

hepatotoxicity, and cancer in humans and wildlife [3]. Polycyclic aromatic hydrocarbons (PAHs), also known as poly-aromatic hydrocarbons or polynuclear aromatic hydrocarbons, are potent atmospheric pollutants that consist of fused aromatic rings and do not contain heteroatoms or carry substituents [4]. As a pollutant, they are of concern because some compounds have been identified as carcinogenic, mutagenic, and teratogenic. Since the pesticides, PCBs and PAHs are inherently toxic in nature; therefore, it is necessary to evaluate the safety aspects of wines with regards to their residues to ensure consumer safety. Furthermore, in addition to causing health hazards, the presence of such contaminant residues may also affect the sensory quality of wines, which in turn could affect the marketability of the products. Cork taint is considered as a major organoleptic defect in wine, which produces mouldy, musty aroma, and could result in significant financial loss to the wine industry.

In addition to pesticides, commodities such as grapes and wines may get contaminated by persistent environmental contaminants. In many developing countries, the pesticide use is not being properly regulated leading to residues in food and food commodities which poses health hazards to the consumers. The treated fruits and vegetables are picked/harvested without taking into consideration the withholding periods. Pesticide residues above the maximum residue limit (MRL) in the crops at harvest are a cause of great concern. These residues make food commodities hazardous for human consumption as well as pollute the environment. Unambiguous identification and quantification of contaminant residues in food at such trace levels demand use of selective sample preparation techniques coupled with highly sensitive and sophisticated instrumentation methods.

Comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC×GC–TOFMS) offers unprecedented separation power in multiresidue analysis. Combination of a long non-polar with a short and polar capillary column connected in series through a thermal modulator provides enormous peak capacity, which is utilized in separating mixture of large number of compounds in single chromatographic run. The TOF mass analyzer further enhances the separation process on the basis of relative flight times of ions as decided by their mass/charge (m/z) ratio. Although quite a number of papers described the separation efficiency of this technique [5,6], the literature evaluating the quantitative performance of this technique in pesticide residue analysis in agricultural products is rather limited.

The quality regulations and food safety standards are becoming more stringent in most countries. Thus, target-oriented residue analysis using quadrupole or ion trap GC–MS with selected ion monitoring (SIM) [7] or tandem mass spectrometry (MS/MS) [8] is not always sufficient to provide complete information about the contamination status of any food sample as these techniques involve targeted acquisition of only selected compounds included in the screening program and at non-target full scan mode, it may not be possible to achieve the desired level of sensitivity for all analytes. As in SIM or MS/MS modes, we loose valuable mass spectral information, the uncertainty level of analysis increases and library-based

screening is not possible. On the other hand, in full scan mode, the closely eluting compounds may affect the quality of each other's mass spectra and this in turn may result in false positives/negatives. Co-elution may also lead to over-estimation or under-estimation of residues. Sometimes, use of long column with slow multi-step temperature programming is useful in resolving co-elution problems but the industry and the regulatory bodies expect a rapid turn-around time, and thus, we cannot afford to have a long GC run to separate multiple compounds. Unambiguous identifications of residues are thus challenging with ^{1}D GC–MS especially when sample history or contamination sources are unknown. Under such situation, GC×GC–TOFMS provide novel solution in providing high peak capacity, adequate sensitivity in full scan (mass range: 5–1000 amu) due to high mass analyzer efficiency and acquisition rate as high as complete 200 spectra/s, generating large number of data points across a narrow peak.

In addition to that chromatographic retention prediction methodologies can be valuable starting points for GC×GC method development. A promising approach is the use of quantitative structure–retention relationship (QSRR) [9]. QSRR are statistically derived relationships between chromatographic parameters and descriptors related to the molecular structure of the analytes. In QSRR these descriptors are used to model the molecular interaction of the analytes with a given stationary phase and eluent. In chromatography, QSRR have been applied to: (*i*) gaina better understanding of the molecular mechanism of the chromatographic separation process; (*ii*) identify the most informative structure related properties of analytes; (*iii*) characterize stationary phases, and (*iv*) predict retention for new analytes [10].

There is a trend to develop QSRR from a variety of methods. In particular, genetic algorithm (GA) is frequently used as search algorithms for variable selection in chemometrics and QSRR. The GA provides a "population" of models, from which it could be difficult to identify the most significant or relevant models (which may be preferred in certain uses, e.g. regulatory toxicology prediction).

Partial least square (PLS) is the most commonly used multivariate calibration method. Moreover, non-linear statistical treatment of QSRR data is expected to provide models with better predictive quality as compared with related PLS models. In this perspective, artificial neural network (ANN) modelling has become quite common in the QSRR field [11]. Extensive use of ANN, which does not require the "a priori" knowledge of the mathematical form of the relationship between the variables, largely rests on its flexibility (functions of any complexity can be approximated. In recent years, nonlinear kernel-based algorithm as kernel partial least squares (KPLS) has been proposed [12]. KPLS can efficiently compute latent variables in the feature space by means of nonlinear kernel functions. Compare to other nonlinear PLS methods, the main advantage of the kernel-based algorithm is that it does not involve nonlinear optimization; thus it essentially requires only linear algebra, which makes it as simple as the conventional linear PLS. In addition, because of its ability to use different kernel functions, KPLS can handle a wide range of nonlinearities. In the present work, a QSRR study has been carried out on the GC×GC–TOFMS system retention times (tR) for 185 compounds in wine and

grape by using structural molecular descriptors. The present study is a first research on QSRR of the compounds in wine and grape against the tR, using GA-KPLS and L-M ANN.

2. Computational

2.1. Data set

Retention times (tR) of 185 compounds including 160 pesticides and 25 environmental organic pollutants in wine and grape were taken from the literature [13]. The selected pesticides (160) included all the GC-amenable chemicals registered in Indian agriculture, which are currently available in market. Twenty-five environmental contaminants selected for the study included 12 dioxin-like polychlorinated biphenyls (PCBs), bisphenol A and 12 polyaromatic hydrocarbons (PAHs). The names of these compounds are presented in Table 1. Sample components are identified and measured by the GC×GC–TOFMS.

Pegasus IV GC×GC–TOFMS system (Leco, St. Joseph, MI, USA) including an Agilent 6890N GC system (Agilent Technologies, USA) and equipped with a CTC Combipal (CTC Analytics, Switzerland) autosampler was used for analysis. Dry nitrogen gas (INOX Air Product, Mumbai, India), liquid nitrogen and compressed air were provided for modulation. Ultra-pure grade helium (Brin's Oxygen Company, Kolkata, India) was used as the carrier gas. Other equipment used in this project included high-speed homogenizer (DIAX-900, Heidolph, Germany), low-volume concentrator (TurboVap LV; Caliper Life Sciences, Russelsheim, Germany), non-refrigerated centrifuge (Eltek, Mumbai, India), refrigerated centrifuge (Kubota, Japan) and a microcentrifuge (Microfuge Pico, Kendro D-37520, Osterode, Germany).

The GC×GC separation was performed by injecting 2 μ L (splitless) on a DB-5MS capillary column (5% phenyl polysilphenylenesiloxane; 30m×0.25mm, 0.25 μ m) connected in series to a Varian V-17 capillary column (50% phenyl, 50% dimethylpolysiloxane; 1m×0.10mm, 0.10 μ m) as the secondary column. Helium was used as the carrier at the corrected constantflowrate of 1.5 mL/min. The injector port was set at 250°C. A gooseneck splitless liner (78.5mm×6.5mm, 4mm) from Restek Corporation (PA, USA) was used. Transfer line temperature was maintained at 305°C. Electron impact ionization was achieved at 70 eV and the ion source temperature was set at 240°C. The mass spectrum of perfluorotributylamine was used to tune the mass spectrometer. The detector voltage was set at −1750V and the data acquisition was carried out within the mass range of 50–550 m/z at acquisition rate of 250 spectra/s at 2-D mode.

2.2. Molecular modeling and theoretical molecular descriptors

The derivation of theoretical molecular descriptors proceeds from the chemical structure of the compounds. In order to calculate the theoretical descriptors, molecular structures were constructed with the aid of HyperChem version 7.0. The final geometries were obtained with the semi-empirical AM1 method in HyperChem program. The molecular structures were optimized using Fletcher-Reeves algorithm until the root mean square gradient was 0.01 kcal mol^{-1}. Some quantum descriptor such as dipole moment and orbital energies of HOMO and LUMO was calculated by using the HyperChem

software. The resulted geometry was transferred into Dragon program, to calculate 1497 descriptors, which was developed by Todeschini *et al* [14]. To reduce the original pool of descriptors to an appropriate size, the objective descriptor reduction was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute either no information or whose information content is redundant with other descriptors present in the pool. As a result, a total of 1061 theoretical descriptors were calculated for each compound in the data sets.

2.3. Genetic algorithm for descriptor selection

To select the most relevant descriptors with GA, the evolution of the population was simulated [15]. Each individual of the population, defined by a chromosome of binary values, represented a subset of descriptors. The number of the genes at each chromosome was equal to the number of the descriptors. The population of the first generation was selected randomly. A gene was given the value of one, if its corresponding descriptor was included in the subset; otherwise, it was given the value of zero. The number of the genes with the value of one was kept relatively low to have a small subset of descriptors that is the probability of generating zero for a gene was set greater. The operators used here were crossover and mutation. The application probability of these operators was varied linearly with a generation renewal. For a typical run, the evolution of the generation was stopped, when 90% of the generations had taken the same fitness. The molecules of the training and the test sets, on which the GA technique was performed, are shown in Table 1. In this paper, size of the population is 30 chromosomes, the probability of initial variable selection is 5:V (V is the number of independent variables), crossover is multi Point, the probability of crossover is 0.5, mutation is multi Point, the probability of mutation is 0.01 and the number of evolution generations is 1000. For each set of data, 3000 runs were performed.

Table 1. The data set, the corresponding observed, predicted tR values, relative error and RMSE by L-M ANN model for the calibration, prediction and test sets.

Entry	Name	tR_{Exp}	tR_{Cal}	RE	RMSE
	Calibration Set				
1	Naphthalene	450	481	6.96	31.30
2	Diflubenzuron	455	429	5.82	26.50
3	Dichlorvos	465	461	0.90	4.20
4	*4*-Bromo-*2*-chlorophenol	495	454	8.38	41.50
5	Diuron	515	551	7.03	36.20
6	Metoxuron	575	582	1.23	7.10
7	*trans*-Mevinphos	580	559	3.62	21.00
8	Acephate	600	585	2.50	15.00
9	Acynephthylene	635	682	7.35	46.70

10	Acynephthene	660	643	2.56	16.90
11	Omethoate	750	686	8.52	63.90
12	Fluorene	765	719	6.00	45.90
13	Monocrotophos	840	764	9.01	75.70
14	Methabenzthiazuron	860	872	1.34	11.50
15	α-Hexachlorocyclohexane	885	818	7.56	66.90
16	Dimethoate	905	906	0.11	1.00
17	Atrazine	915	845	7.66	70.10
18	Fluchloralin	935	986	5.40	50.50
19	β-Hexachlorocyclohexane	945	932	1.38	13.00
20	γ-Hexachlorocyclohexane	950	1011	6.46	61.40
21	Pyremethanil	965	878	9.07	87.50
22	Paraoxon methyl	980	999	1.96	19.20
23	Flufenoxuron	985	1074	8.98	88.50
24	Kitazin	990	990	0.00	0.00
25	Anthracene	995	1072	7.78	77.40
26	δ-Hexachlorocyclohexane	1010	930	7.97	80.50
27	Spiroxamine:1	1025	936	8.69	89.10
28	Chloropyriphos-methyl	1030	975	5.31	54.70
29	Vinclozoline	1035	1125	8.67	89.70
30	Alachlor	1040	1001	3.75	39.00
31	Methyl parathion	1045	1034	1.05	11.00
32	Metalaxyl	1050	1155	9.99	104.90
33	Phenchlorphos	1060	999	5.76	61.10
34	Carbaril	1065	1065	0.03	0.30
35	Spiroxamine:2	1070	1025	4.19	44.80
36	Demeton-*S*-methyl sulfone	1080	1117	3.45	37.30
37	Phenitrothion	1085	1184	9.16	99.40
38	Linuron	1100	1159	5.33	58.60
39	Chlorpyriphos-ethyl	1105	1101	0.36	4.00
40	Tetraconazole	1120	1037	7.38	82.70
41	Triadimefon	1125	1034	8.09	91.00
42	Aldrin	1130	1020	9.70	109.60
43	*cis*-Chlorfenvinphos	1160	1142	1.55	18.00
44	Bioallethrin	1175	1249	6.33	74.40
45	Penconazole	1185	1300	9.71	115.10
46	Trifloxystrobin acid metabolite	1195	1193	0.17	2.00
47	Phenthoate	1200	1203	0.25	3.00
48	Prallethrin	1205	1234	2.42	29.20
49	Triadimenol:1	1210	1225	1.21	14.60

50	Epoxyheptachlor	1215	1153	5.12	62.20
51	Triadimenol:2	1230	1309	6.44	79.20
52	Captan	1235	1177	4.74	58.50
53	Butachlor	1240	1303	5.07	62.90
54	Fluoranthene	1245	1285	3.20	39.80
55	*o,p*-DDE	1250	1179	5.72	71.50
56	Thiabendazole	1255	1281	2.04	25.60
57	α-Endosulfan	1290	1307	1.32	17.00
58	Oxadiazon	1305	1347	3.19	41.60
59	Pyrene	1310	1227	6.34	83.00
60	Profenofos	1315	1351	2.73	35.90
61	Oxyfluorfen	1320	1233	6.59	87.00
62	*p,p*-DDE	1330	1341	0.79	10.50
63	Buprofezin	1335	1368	2.46	32.90
64	Flusilazole	1335	1390	4.12	55.00
65	PCB IUPAC No.077	1340	1388	3.54	47.50
66	*o,p*-DDD	1350	1452	7.53	101.70
67	Dieldrin	1360	1363	0.22	3.00
68	Ethion	1410	1357	3.79	53.50
69	PCB IUPAC No.118	1415	1494	5.58	79.00
70	β-Endosulfan	1435	1380	3.85	55.30
71	Trifloxystrobin	1455	1579	8.54	124.20
72	PCB IUPAC No. 126	1470	1523	3.57	52.50
73	*cis*-Propiconazole	1485	1359	8.51	126.40
74	Flupicolide	1495	1427	4.55	68.00
75	*p,p*-DDT	1505	1358	9.76	146.90
76	Endosulfan sulfate	1510	1428	5.46	82.40
77	Propargite	1520	1525	0.32	4.90
78	Triphenyl phosphate	1535	1522	0.83	12.80
79	PCB IUPAC No. 157	1565	1629	4.06	63.60
80	Iprodione	1570	1596	1.65	25.90
81	Tetramethrin	1585	1588	0.18	2.90
82	Phosmet	1600	1464	8.51	136.20
83	PCB IUPAC No.167	1605	1507	6.14	98.50
84	PCB IUPAC No. 169	1615	1576	2.41	38.90
85	Dicofol	1620	1601	1.17	19.00
86	Fenazaquin	1625	1713	5.42	88.00
87	Chrysene	1635	1574	3.72	60.80
88	Phosalone	1655	1701	2.76	45.70
89	λ-Cyhalothrin	1660	1822	9.73	161.50

90	PCB IUPAC No. 156	1670	1671	0.06	1.00
91	Diafenthiuron	1705	1858	8.96	152.70
92	Fenarimol	1715	1828	6.56	112.50
93	PCB IUPAC No. 189	1720	1864	8.35	143.60
94	*cis*-Permethrin	1755	1619	7.74	135.80
95	Bitertanol	1755	1865	6.25	109.60
96	Cyfluthrin:1	1810	1958	8.17	147.80
97	Cyfluthrin:2	1820	1816	0.22	4.00
98	Cyfluthrin:3	1830	1692	7.56	138.30
99	Cypermethrin:1	1850	1947	5.25	97.20
100	Flucythrinate:1	1865	1948	4.43	82.70
101	Cypermethrin:3	1875	1902	1.44	27.00
102	Cypermethrin:4	1885	1981	5.08	95.80
103	Flucythrinate:2	1890	1952	3.26	61.70
104	Etofenprox	1895	2040	7.67	145.30
105	Fluvalinate:1	1980	1890	4.56	90.20
106	Fluvalinate:2	1990	1981	0.45	9.00
107	Benzo(*a*)pyrene	1995	1982	0.65	13.00
108	Difenoconazole:1	2085	1891	9.29	193.70
109	Difenoconazole:2	2095	1920	8.38	175.50
110	Deltamethrin	2125	2274	7.01	149.00
111	Dimethomorph-2	2265	2193	3.20	72.40
	Prediction Set				
112	Isoproturon	460	465	1.11	5.10
113	*cis*-Mevinphos	575	617	7.25	41.70
114	Fenbucarb	745	771	3.54	26.40
115	Demeton-S-methyl	775	689	11.10	86.00
116	Carbofuran	905	822	9.18	83.10
117	Diazinon	935	929	0.62	5.80
118	Etrimphos	970	961	0.93	9.00
119	Phenanthrene	985	1000	1.52	15.00
120	Fenchlorphos-oxon	1005	1111	10.50	105.50
121	Malaoxon	1025	951	7.22	74.00
122	Metribuzine	1040	947	8.94	93.00
123	Heptachlor	1065	1061	0.38	4.00
124	Malathion	1085	1175	8.26	89.60
125	Phenthion	1120	951	15.09	169.00
126	Fipronil	1165	1054	9.51	110.80
127	Cyprodinil	1180	1129	4.30	50.70
128	Quinalphos	1205	1200	0.41	5.00

129	Methidathion	1245	1142	8.27	103.00
130	Folpet	1250	1263	1.02	12.80
131	Vamidothion	1260	1424	13.02	164.00
132	Imazalil	1300	1363	4.85	63.00
133	Isoprothiolane	1310	1451	10.78	141.20
134	Bisphenol A	1335	1235	7.48	99.80
135	Endrin	1410	1461	3.63	51.20
136	*o,p*-DDT	1435	1482	3.25	46.70
137	Benalaxyl	1470	1479	0.60	8.80
138	Ediphenphos	1495	1414	5.39	80.60
139	PCB IUPAC No.081	1365	1534	12.38	169.00
140	PCB IUPAC No. 123	1540	1709	10.97	169.00
141	Biphenthrin	1565	1750	11.83	185.10
142	Fenpropathrin	1595	1664	4.35	69.40
143	Benz(*a*)anthracene	1625	1733	6.65	108.00
144	Oryzalin	1740	1738	0.11	2.00
145	Cyfluthrin:4	1840	1905	3.52	64.80
146	Benzo(*b*)fluoroathene	1885	2134	13.21	249.00
147	Indoxacarb	2075	2423	16.79	348.30
148	Azoxystrobin	2170	2131	1.81	39.20
	Test Set				
149	Methamidophos	465	521	11.96	55.60
150	Sulfosulfuron	520	577	10.94	56.90
151	Propoxur	750	779	3.85	28.90
152	Phorate	850	786	7.52	63.90
153	Thiometon	905	1062	17.38	157.30
154	*cis*-Phosphamidon	945	931	1.48	14.00
155	Chlordene	980	1013	3.36	32.90
156	Chlorothalonil	985	850	13.68	134.70
157	*trans*-Phosphamidon	1010	1003	0.74	7.50
158	Propanil	1035	1213	17.23	178.30
159	Carbofuran-*3*-OH	1045	1202	15.06	157.40
160	Pirimiphos-methyl	1065	1065	0.01	0.10
161	Dichlofluanid	1100	1281	16.42	180.60
162	Parathion	1125	1174	4.39	49.40
163	*trans*-Chlorfenvinphos	1185	1266	6.83	80.90
164	Pendimethalin	1170	1269	8.48	99.20
165	Procymidone	1215	1335	9.86	119.80
166	*cis*-Chlordane	1250	1356	8.50	106.20
167	Paclobutrazole	1265	1399	10.55	133.50

168	*trans*-Chlordane	1285	1353	5.31	68.20
169	Hexaconazole	1305	1454	11.42	149.00
170	Kresoxim-methyl	1330	1403	5.46	72.60
171	Myclobutanil	1340	1450	8.18	109.60
172	PCB IUPAC No. 105	1435	1698	18.31	262.70
173	Triazophos	1455	1628	11.92	173.40
174	*trans*-Propiconazole	1495	1488	0.47	7.00
175	Tebuconazole	1535	1652	7.63	117.10
176	PCB IUPAC No.114	1405	1401	0.28	4.00
177	Oxycarboxin	1560	1523	2.37	37.00
178	Captafol	1565	1806	15.37	240.50
179	Phenothrin	1620	1783	10.04	162.60
180	Azinphos-methyl	1686	1514	10.18	171.70
181	*trans*-Permethrin	1765	1886	6.85	120.90
182	Cypermethrin:2	1865	1933	3.64	67.80
183	Fenvalerate	1985	2345	18.16	360.40
184	Esfenvalerate	2020	1989	1.53	31.00
185	Dimethomorph-1	2200	2023	8.05	177.10

2.4. Nonlinear model

2.4.1. Artificial neural network

An artificial neural network (ANN) with a layered structure is a mathematical system that stimulates biological neural network, consisting of computing units named neurons and connections between neurons named synapses [16]. All feed-forward ANN used in this paper are three-layer networks. Each neuron in any layer is fully connected with the neurons of a succeeding layer. The Levenberg–Marquardt back propagation algorithm was used for ANN training and the linear functions were used as the transformation functions in hidden and output layers.

3. Results and discussion

3.1. Linear model

3.1.1. Results of the GA-PLS model

To investigate whether there is a linear relationship existing between the descriptors and compounds, the widely used PLS approach is also applied in the present work. PLS is a linear modeling technique where information in the descriptor matrix X is projected onto a small number of underlying (latent) variables called PLS components, referred to as latent variables. The matrix Y is simultaneously used in estimating the latent variables in X that will be most relevant for predicting the Y variables. For regression analysis, data set was separated into training and test sets. The prediction error sum of squares (PRESS) obtained in the cross-validation was calculated each time that a new principal component (PC) was added to the model. The optimum number of PLS factors is the one that minimizes PRESS. Appropriate models with low PRESS and high correlation coefficients were obtained.

Consequently, among different models, the best model was chosen. It is obvious that as the number of descriptors increase the R^2 will increase. Figure 1a shows the plot of PRESS versus the number of factors for the PLS model. The best GA-PLS model contains 8 selected descriptors in 3 latent variables space. These descriptors were obtained constitutional descriptors (number of atoms (nAT) and number of Hydrogen atoms (nH)), WHIM descriptors (1st component symmetry directional WHIM index / weighted by atomic masses (G1m) and 3st component symmetry directional WHIM index / weighted by atomic polarizabilities (G3p)), charge descriptors (relative negative charge (RNCG) and total squared charge (Q2)) and quantum chemical descriptors (dipole moment (μ) and high occupied molecular orbital (HOMO)). For this in general, the number of components (latent variables) is less than number of independent variables in PLS analysis. The predicted values of tR are plotted against the experimental values for training and test sets in Figure 1b. The R^2, mean relative error (RE) and RMSE for training and test sets were (0.913, 0.871), (9.03, 12.26) and (245.08, 301.63), respectively. The PLS model uses higher number of descriptors that allow the model to extract better structural information from descriptors to result in a lower prediction error.

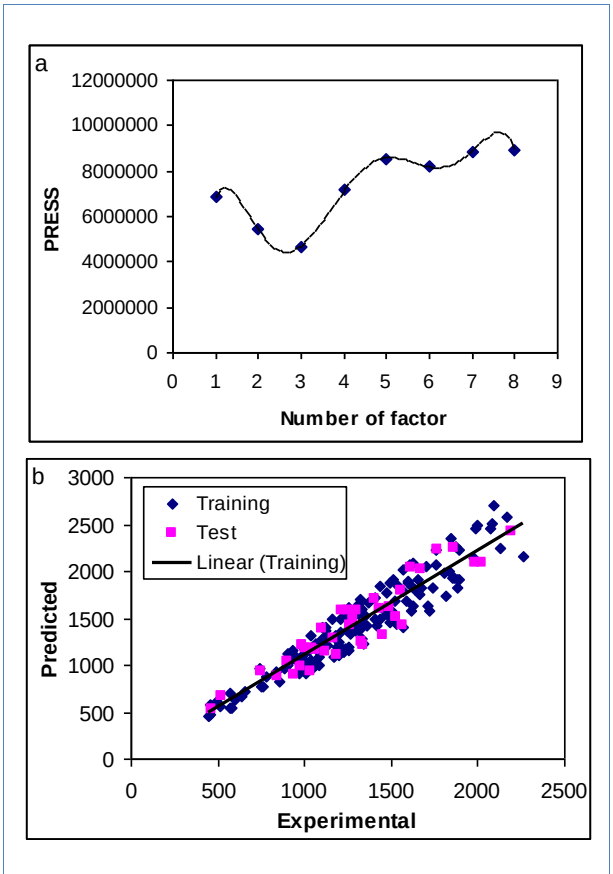


Figure 1. (a) PRESS versus the number of factors for the GA-PLS model, (b) Plots of predicted retention time against the experimental values by GA-PLS model

3.2. Nonlinear models

3.2.1 Results of the GA-KPLS model

In this paper a radial basis kernel function, $k(x,y)= \exp(||x-y||^2/c)$, was selected as the kernel function with $c = rm\sigma^2$ where r is a constant that can be determined by considering the process to be predicted (here r was set to be 1), m is the dimension of the input space and σ^2 is the variance of the data. It means that the value of c depends on the system under the study. Figure 2a shows the plot of PRESS versus the number of factors for the KPLS model. The 5 descriptors in 2 latent variables space chosen by GA-KPLS feature

selection methods were contained. These descriptors were obtained constitutional descriptors (nH), geometrical descriptors (gravitational index G2 (bond-restricted) (G2) and radius of gyration (mass weighted) (Rgyr)), molecular properties (topological polar surface area using N,O,S,P polar contributions (TPSA(Tot)) and quantum descriptors (lowest unoccupied molecular orbital (LUMO)). Figure 2b shows the plot of the GA-KPLS predicted versus experimental values for tR of all of the molecules in the data set. For the constructed model, three general statistical parameters were selected to evaluate the prediction ability of the model for the tR. The statistical parameters R^2, RE and RMSE were obtained for proposed models. Each of the statistical parameters mentioned above were used for assessing the statistical significance of the QSRR model. The R^2, RE and RMSE for training and test sets were (0.939, 0.901), (8.02, 9.91) and (183.21, 201.35), respectively. It can be seen from these results that statistical results for GA-KPLS model are superior to GA-PLS method. Inspection of the results reveals a higher R^2 and lower RMSE and RE for the GA-KPLS method compared with their counterparts for linear model. Also, a lower number of variables have appeared in the former model. This clearly shows the strength of GA-KPLS as a nonlinear feature selection method. This suggests that GA-KPLS hold promise for applications in choosing of variable for L-M ANN systems. This result indicates that the tR of compounds in wine and grape possesses some nonlinear characteristics.

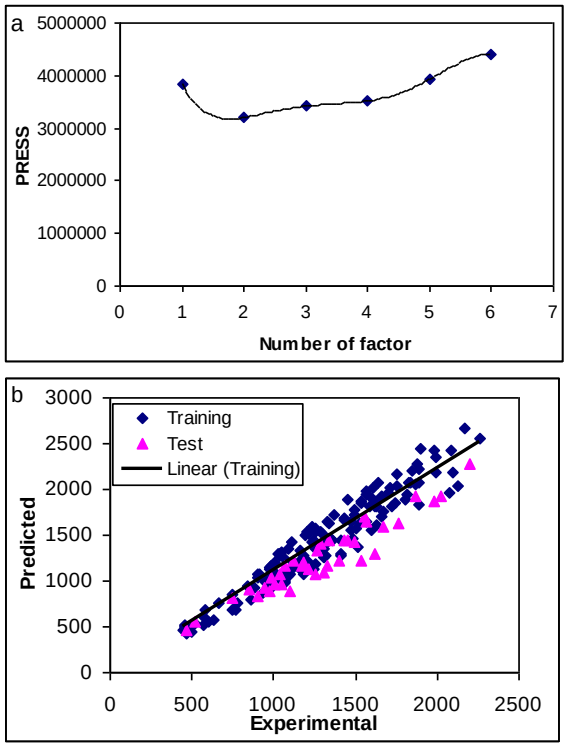


Fig. 2. (a) PRESS versus number of factors for the GA-KPLS model, (b) Plots of predicted tR against the experimental values by GA-KPLS model.

3.2.2. Results of the L-M ANN model

With the aim of improving the predictive performance of nonlinear QSRR model, L-M ANN modeling was performed. The networks were generated using the five descriptors appearing in the GA-PLS models as their inputs and tR as their output. For ANN generation, data set was separated into three groups: calibration and prediction (training) and test sets. All molecules were randomly placed in these sets. A three-layer network with a sigmoid transfer function was designed for each ANN. Before training the networks the input and output values were normalized between -1 and 1. The network was then trained using the training set by the back propagation strategy for optimization of the

weights and bias values. The procedure for optimization of the required parameters is given elsewhere. The proper number of nodes in the hidden layer was determined by training the network with different number of nodes in the hidden layer. The root-mean-square error (RMSE) value measures how good the outputs are in comparison with the target values. It should be noted that for evaluating the overfitting, the training of the network for the prediction of tR must stop when the RMSE of the prediction set begins to increase while RMSE of calibration set continues to decrease. Therefore, training of the network was stopped when overtraining began. All of the above mentioned steps were carried out using basic back propagation, conjugate gradient and Levenberge Marquardt weight update functions. It was realized that the RMSE for the training and test sets are minimum when three neurons were selected in the hidden layer and the learning rate and the momentum values were 0.8 and 0.4, respectively. Finally, the number of iterations was optimized with the optimum values for the variables. It was realized that after 18 iterations, the RMSE for prediction set were minimum.

The values of experimental, calculated, percent relative error and RMSE are shown in Table 1. The R^2, RE and RMSE for calibration, prediction and test sets were (0.967, 0.942, 0.922), (4.8, 6.48, 8.46) and (77.12, 113.14, 133.61), respectively. Inspection of the results reveals a higher R^2 and lowers other values parameter for the test set compared with their counterparts for other models. Plots of predicted tR versus experimental tR values by L-M ANN for calibration, prediction and test sets are shown in Figure 3a,3b, respectively.

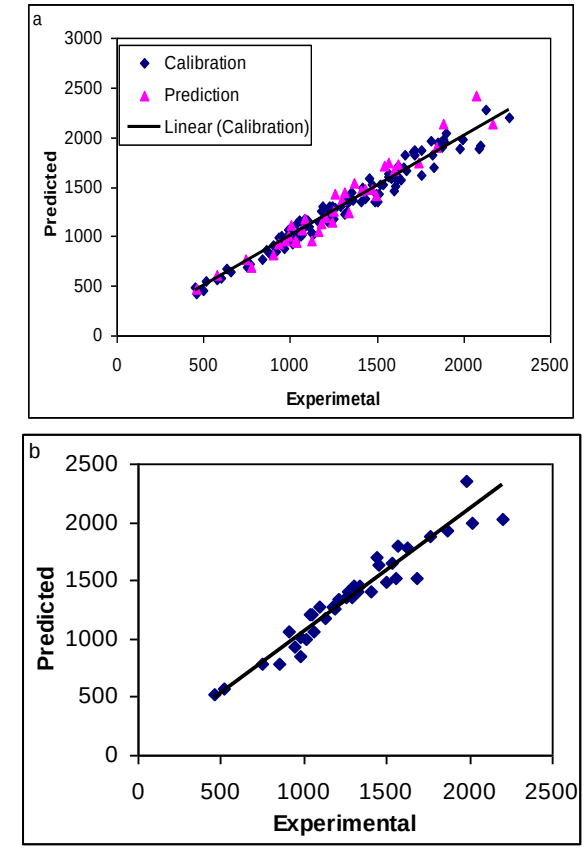


Fig 3. Plot of predicted tR obtained by L-M ANN against the experimental values (a) calibration and prediction sets of molecules and (b) for test set

The relative error and R^2 of test set for the GA-PLS model are 12.26 and 0.871 respectively and for the GA-KPLS model are 9.91 and 0.901 respectively which would be compared with the values of 8.46 and 0.922, respectively, for the L-M ANN model. Comparison between these values and other statistical parameters reveals the superiority of the L-M ANN model over other models. The key strength of neural networks, unlike regression analysis, is their ability to flexible mapping of the selected features by manipulating their functional dependence

implicitly. The statistical parameters reveal the high predictive ability of L-M ANN model. The whole of these data clearly displays a significant improvement of the QSRR model consequent to non-linear statistical treatment. Obviously, there is a close agreement between the experimental and predicted tR and the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero. As can be seen in this section, the L-M ANN is more reproducible than GA-KPLS for modeling the GC×GC–TOFMS retention time of compounds in wine and grape.

3.3. Model validation and statistical parameters

The applied internal (leave-group-out cross validation (LGO-CV)) and external (test set) validation methods were used for the predictive power of models. In the leave-group-out procedure one compound was removed from the data set, the model was trained with the remaining compounds and used to predict the discarded compound. The process was repeated for each compound in the data set. The predictive power of the models developed on the selected training set is estimated on the predicted values of test set chemicals. The data set should be divided into three new sub-data sets, one for calibration and prediction (training), and the other one for testing. The calibration set was used for model generation. The prediction set was applied deal with overfitting of the network, whereas test set which its molecules have no role in model building was used for the evaluation of the predictive ability of the models for external set.

In the other hand by means of training set, the best model is found and then, the prediction power of it is checked by test set, as an external data set. In this work, 60% of the database was used for calibration set, 20% for prediction set and 20% for test set, randomly (in each running program, from all 185 components, 111 components are in calibration set, 37 components are in prediction set and 37 components are in test set).

The result clearly displays a significant improvement of the QSRR model consequent to non-linear statistical treatment and a substantial independence of model prediction from the structure of the test molecule. In the above analysis, the descriptive power of a given model has been measured by its ability to predict partition of unknown compounds in wine and grape.

For the constructed models, some general statistical parameters were selected to evaluate the predictive ability of the models for tR values. In this case, the predicted tR of each sample in prediction step was compared with the experimental acidity constant. The PRESS (predicted residual sum of squares) statistic appears to be the most important parameter accounting for a good estimate of the real predictive error of the models. Its small value indicates that the model predicts better than chance and can be considered statistically significant.

$$PREES = \sum_{i=1}^{n} (y_i^{\wedge} - y_i)^2 \qquad \text{(Eq.1)}$$

Root mean square error (RMSE) is a measurement of the average difference between predicted and experimental values, at the prediction step. RMSE can be interpreted as the average prediction error, expressed in the same units as the original response values. The RMSE was obtained by the following formula:

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2]^{\frac{1}{2}} \quad \text{(Eq.2)}$$

The third statistical parameter was relative error (RE) that shows the predictive ability of each component, and is calculated as:

$$RE(\%) = 100 \times \left[\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{y_i} - y_i)}{y_i}\right] \quad \text{(Eq.3)}$$

The predictive ability was evaluated by the square of the correlation coefficient (R^2) which is based on the prediction error sum of squares (PRESS) and was calculated by following equation:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y_i} - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})} \quad \text{(Eq.4)}$$

Where y_i is the experimental tR in the sample *i*, $\hat{y}_i$ represented the predicted tR in the sample *i*, $\bar{y}$ is the mean of experimental tR in the prediction set and n is the total number of samples used in the test set.

The main aim of the present work was to assess the performances of GA-KPLS and L-M ANN for modeling the retention time of compounds in wine and grape. The procedures of modeling including descriptor generation, splitting of the data, variable selection and validation were the same as those performed for modeling of the retention time of compounds in wine and grape.

3.4. Interpretation of Descriptors

It is well known that gas chromatographic retention time, unlike other molecular properties; strongly depend on interactions between eluents and stationary phases because interactions between eluents can be widely neglected. In the chromatographic retention of compounds in the intermediate polarity stationary phase's two important types of interactions contribute to the chromatographic retention of the compounds: the induction and dispersion forces. The dispersion forces are related to steric factors, molecular size and branching, while the induced forces are related to the dipolar moment, which should stimulate dipole-induced dipole interactions.

Constitutional descriptors are the most simple and commonly used descriptors, reflecting the molecular composition of a compound without any information about its molecular geometry. The most common constitutional descriptors are number of atoms, number of bound, absolute and relative numbers of specific atom type, absolute and relative numbers of single, double, triple, and aromatic bound, number of ring, number of ring divided by the number of atoms or bonds, number of benzene ring, number of benzene ring divided by the number of atom, molecular weight and average molecular weight [17].

The hydrogen bonding is a measure of the tendency of a molecule to form hydrogen bonds. This is related to the number of Hydrogen atoms (nH). Hydrogen-bonding may be divided into an

electrostatic term and a polarization/charge transfer term. Understandably, hydrogen bonding plays a significant role in retention behavior. Hydrogen bonding is not a true bond, but a very strong form of dipole–dipole attraction. Solvents with higher strength of hydrogen-bonding ability, the solutes should have higher retention times, i.e. they interact with mobile phase more strongly and eluted with lower speed. This implies that solutes with higher hydrogen bond ability should interact more with the mobile phase.

The geometrical descriptors are suitable for complex-behaved properties, because they take into account the 3D-arrangement the atoms without ambiguities (as those appearing when using chemical graphs), as well as they do not depend on the molecular size and thus they are applicable to a large number of molecules with great structural variance, which have a characteristic common to all of them. Gravitational index (G2) (bond-restricted) is a geometrical descriptor that reflecting the mass distribution in a molecule and defined as Eq. (5):

$$G_2 = \sum_{a=1}^{A} \left(\frac{m_i . m_j}{r_{ij}^2} \right)_a \qquad \text{(Eq.5)}$$

Where m_i and m_j are the atomic masses of the considered atoms; r_{ij} the corresponding interatomic distances; and A the number of all pairs of bonded atoms of the molecule. This index is related to the bulk cohesiveness of the molecules, accounting, simultaneously, for both atomic masses (volumes) and their distribution within the molecular space. This index can be extended to any other atomic property different from atomic mass, such as atomic polarizability, atomic, van der Waals volumetric [18].

Radius of gyration or gyradius (Rgyr), also referred to as gyradius, is the radial distance from a given axis at which the mass of a body could be concentrate d without altering the rotational inertia of the body about that axis. It is the name of several related measures of the size of an object, a surface, or an ensemble of points. It is calculated as the root mean square distance of the objects' parts from either its center of gravity or an axis. The gyradius (k) about a given axis can be computed in terms of the moment of inertia, and the total mass m;

$$K = \sqrt{\frac{1}{m}} \qquad \text{(Eq.6)}$$

The WHIM descriptors are built in such a way as to capture the relevant molecular 3-D information regarding the molecular size, shape, symmetry, and atom distribution with respect to some invariant reference frame. These descriptors are quickly computed from the atomic positions of the molecule atoms (hydrogens included). WHIM descriptors are based on principal component analysis of the weighted covariance matrix obtained from the atomic Cartesian coordinates. In relation to the kind of weights selected for the atoms different sets of WHIM descriptors can be obtained. Unitary weights (u), atomic mass (m), atomic van der Waals volume (v), atomic electronegativity (e), atomic polarizability (p) and atomic electrotopological state (s) are the available weighting schemes globally providing 66 directional and 33 global WHIM descriptors.

G1m variable is among WHIM descriptors and combines the atomic properties with 3D-structure information of molecules. This parameter is related to branching or cyclysity of molecules, and the value of this index increases from linear to more branched compounds.

TPSA (Tot) of a molecule is defined as the surface sum over of polar atoms. This molecular descriptor explains the electrostatic and polarization interactions between the solute and the solvent. All the interactions are obviously weak interactions such as higher multipole, dipole and induced-dipole interactions. So, TPSA (Tot) can be considered an important electrostatic descriptor compounds a QSRR study to understand the charge distribution of the molecules and use this information to project new compounds with desired properties. The values of TPSA were calculated by summarizing the respective fragmental constants of the two-dimensional structure of the considered compounds according to a procedure proposed by Ertl *et al.* [18].

Charge descriptors were defined in terms of atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such atoms, bonds, and molecular fragments. Electrical charges in the molecule are the driving force of electrostatic interactions, and it is well known that local electron densities or charge play a fundamental role in many physical-chemical properties and receptors-ligand binding affinity. Thus, charge based descriptors have been widely employed as chemical reactivity indices or as measures of weak intermolecular interactions.

Electric charge comes in two types, called positive and negative. Two positively charged substances, or objects, experience a mutual repulsive force, as do two negatively charged objects. Positively charged objects and negatively charged objects experience an attractive force. Relative negative charge (RNCG) is partial charge of the most negative atom divided by the total negative [17]:

$$RNCG = \frac{Q^{-}_{\max}}{Q^{-}} \qquad \text{(Eq.7)}$$

Although constitutional, WHIM, geometrical and charge descriptors are often successful in retention of these compounds, they cannot account for conformational changes and they do not provide information about electronic influence through bonds or across space. For that reason, quantum chemical descriptors are used in developing QSRR.

Quantum chemical descriptors can give great insight into structure and reactivity and can be used to establish and compare the conformational stability, chemical reactivity and inter-molecular interactions. They include thermodynamic properties (system energies) and electronic properties (LUMO or HOMO energy). Electronic properties may play a role in the magnitude in a biological activity, along with structural features encoded in indexes. The eigenvalues of LUMO and HOMO and their energy gap reflect the chemical activity of the molecule. LUMO as an electron acceptor represents the ability to obtain an electron, while HOMO as an electron donor represents the ability to donate an electron. The HOMO energy plays a very important role in the nucleophylic behavior and it represents molecular

reactivity as a nucleophyle. The energy of the LUMO is directly related to the electron affinity and characterizes the susceptibility of the molecule toward attack by nucleophiles. Electron affinity was also shown to greatly influence the chemical behavior of compounds, as demonstrated by its inclusion in the QSRR [19-21].

Polar functional groups account for many of the dipole–dipole, dipole-induced dipole and hydrogen bond interactions. Dipole moment is the measure of polarity of the molecule. Dipole moment describes the intramolecular electronic effect, which may be related to molecular reactivity. The activity of a molecule increases as the dipole moment is increases. Solutes should have higher retention times in a mobile phase with higher polarity/polarizability index.

From the above discussion, it can be seen that the particle size, hydrogen bonding and electrostatic interactions are the likely three factors controlling the tR of these compounds. All descriptors involved in the model, which have explicit physical meaning, may account for the structure responsible for the tR of these compounds.

4. Conclusion

The GA-PLS, GA-KPLS and L-M ANN modeling was applied for the prediction of the retention time values of 185 compounds in wine and grape. Three methods seemed to be useful, although a comparison between these methods revealed the slight superiority of the L-M ANN over the models. High correlation coefficients and low prediction errors confirmed the good predictability of three models. Application of the developed model to a testing set of 37 compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. Since the QSRR was developed on the basis of theoretical molecular descriptors calculated exclusively from molecular structure, the proposed model could potentially provide useful information about the tR of compounds. This procedure allowed us to achieve a precise and relatively fast method for determination of tR of different series of compounds in wine and grape to predict with sufficient accuracy the tR of new compound derivatives. The advantages of this study were that the number of the used descriptors was smaller and that all the used descriptors were related to retention of these compounds in GC×GC–TOFMS. To the best of our knowledge this is the first study for the prediction of retention time of compounds in wine and grape using GA-KPLS and L-M ANN.

Acknowledgment The authors are grateful for the partial financial support from Islamic Azad University of Ilam.

ORCID
Kaushik Pal: 0000-0002-9313-6497
Aydi Abdelkarim: 0000-0002-2928-7055

References

[1] B. Rankine, Making Good Wine: A Manual of Winemaking Practice for Australia and New Zealand, Sun Pan Macmillan, Australia, Sydney, 1995.
[2] M.J. Torija, N. Rozes, M. Poblet, J.M. Guillamon, A. Mas, *Antonie van Leeuwenhoek*., **2001**, *79*, 345-352.

[3] K. Hilscherova, M. Machala, K. Kannan, A.L. Blankenship, J.P. Giesy, *Environ Sci Pollut Res*., **2000**, *7*, 159-171.
[4] J.C. Fetzer, *Polycycl. Aromat. Compd.*, **2000**, *27*, 143.
[5] J. Zrostlıkova, J. Hajslova, T. Cajka, *J. Chromatogr. A.*, **2003**, *1019*, 173-182.
[6] M. Adahchour, J. Beens, R.J.J. Vreuls, A. Max Batenburg, U.A.Th. Brinkman, *J. Chromatogr. A*., **2004**, *1054*, 47-54.
[7] J.W. Wong, M.K. Hennessy, D.G. Hayward, A.J. Krynitsky, I. Cassias, F.J. Schenck, *J. Agric. Food Chem.*, **2007**, *55*, 1117-1124.
[8] J.L.M. Vidal, F.J.A. Liebanas, M.J.G. Rodrıguez, A.G. Frenich, J.L.F. Moreno, *Rapid Commun. Mass Spectrom*., **2006**, *20*, 365-361.
[9] R. Put, Y. Vander Heyden, *Anal. Chim. Acta*., **2007**, *602*, 164-172.
[10] R. Kaliszan, Structure, Retention in Chromatography. A Chemometric Approach, Harwood Academic Publishers, Amsterdam, 1997.
[11] H. Noorizadeh, A. Farmany, Chromatographia., **2010**, *72*, 563-569.
[12] H. Noorizadeh, A. Farmany, *Drug Test Anal*, **2012**., *4*, 151-157.
[13] S. Dasgupta, K. Banerjee, S.H. Patil, M. Ghaste, K.N. Dhumal, P.G. Adsule*, J. Chromatogr. A*., **2010**, *1217*, 3881-3889.
[14] R. Todeschini, V. Consonni, A. Mauri, M. Pavan., DRAGON-Software for the calculation of molecular descriptors; Version 3.0 for Windows, 2003.
[15] S. Ahmad, M.M. Gromiha, *J. Comput. Chem*., **2003**, *24*, 1313-1320.
[16] S. Kara, A.S. Güven, M. Okandan, F. Dirgenali., *Comput. Biol. Med.*, **2008**, *36*, 473-483
[17] P. Ghosh, M. Vracko, A.K. Chattopadhyay, M.C. Bagchi, *Chem Biol Drug Des*., **2008**, *72*, 155-162.
[18] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley/VCH, Weinheim, 2000.
[19] P. Thanikaivelan, V. Subramanian, J.R. Rao, B.U. Nair, *Chem. Phys. Lett.*, **2000**, *323*, 59-64.
[20] M.M. Heravi, H. Abdi Oskooie, Z. Latifi, H. Hamidi, *Adv. J. Chem. A*, **2018**, *1*, 7-11.
[21] A. Moghimi, M. Yari, *J. Chem. Rev.*, **2019**, *1*, 1-18.

How to cite this manuscript: Issa Amini, Kaushik Pal, Sharmin Esmaeilpoor, Aydi Abdelkarim, Prediction of two-dimensional gas chromatography time-of-flight mass spectrometry retention times of 160 pesticides and 25 environmental organic pollutants in grape by multivariate chemometrics methods, *Adv. J. Chem. A*, **2018**, *1(1)*, 12-31.