PREDICTION OF GAS CHROMATOGRAPHIC RETENTION INDICES USING VARIABLE CONNECTIVITY INDEX†

Milan Randić
Department of Mathematics and Computer Science, Drake University, Des Moines, IA 50311, USA

Subash C. Basak
Natural Resources Research Institute, University of Minnesota at Duluth, Duluth, MN 55811, USA

Matevž Pompe
Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1000 Ljubljana, Slovenia

Marjana Novič
National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia

†This paper is dedicated to the memory of Professor Drago Kolar

Received 24-10-2000

Abstract

We have re-examined gas chromatographic retention indices of alkanes (48 compounds), and alcohols (31 compounds), combining all molecules into a single set (n=79) using variable connectivity index $^{1}\chi^{f}$. By varying the weight for oxygen atom we obtained the regression characterized by the correlation coefficient r = 0.9933, the standard error s=14.24 retention time units, and Fisher ratio F = 5695. Use of the simple connectivity index $^{1}\chi$, which does not differentiate carbon and oxygen atoms, gives regression with the standard error four times larger.

Introduction

To derive a structure-property regression one has to select suitable molecular descriptors.[1,2] Even though there are several hundreds of descriptors available for use,[3] often they would show limited ability to correlate with a selected molecular property. This justifies continuing interest in construction of novel topological indices. However, rather then expanding the existing large pool of descriptors we would like to advocate use of *a novel kind* of topological indices, which can be modified during the search for best regressions. These indices can be contrasted to all hitherto designed topological indices, which are numerically fixed once structure is selected. Novel indices may have conceptual similarity to the traditional indices in the sense that for special values of their variables they may reduce to one of the known numerically fixed index. In this paper in particular we consider variable connectivity index $^{1}\chi^{f}$. Although the variable connectivity indices were introduced in quantitative structure-activity relationship, QSAR, already 10 years ago,[4,5] apparently their potential has been overlooked and until very recently they have not received due attention.[6-9]

We will illustrate use of the variable connectivity index $^1\chi^f$ and will demonstrate their ability to yield regressions of very high quality. We will re-examine data on chromatographic retention indices for a subset of alkanes and alcohols that have been recently studied by two of the present authors.[10]

Variable connectivity index

One can calculate the connectivity index $^1\chi$ [11] (known also as Randić connectivity index of order 1,[12] by combining the row sums of the adjacency matrix of molecular graph. If zeros on the diagonal of the adjacency matrix are replaced by weights x, y, ..., which characterize different kind of atoms, one obtains the augmented adjacency matrix (Table 1). From the row sums of augmented matrix one can obtain the flexible connectivity index $^1\chi^f$ similarly as one can calculate the connectivity index from the row sums of the adjacency matrix. One simply combines the row sums S_i, S_j of the matrix corresponding to bond (i, j) by using the algorithm $1/\sqrt{(S_i,S_j)}$. In Table 2 we show the construction of $^1\chi$ and $^1\chi^f$ for 2-methyl-3-pentanol.

Table 1. Augmented adjacency matrix for 2-methyl-3-pentanol.

Atom no.	1	2	3	4	5	6	7	Row sum
1	x	1	0	0	0	0	0	1+x
2	1	x	1	0	0	1	0	3+ x
3	0	1	x	1	0	0	1	3+x
4	0	0	1	x	1	0	0	2+x
5	0	0	0	1	x	0	0	1+x
6	0	1	0	0	0	x	0	1+x
7	0	0	1	0	0	0	y	1+y

The connectivity index $^1\chi^f$ is now a function of two variables. By varying these variables one can change the relative contributions to the connectivity index of atoms of different kind. In Table 3 we illustrate this for 2-methyl-3-pentanol by keeping x = 0, and varying y, the variable depicting the role of oxygen atoms in alcohols. Such change of variables can often reduce the standard error of a regression drastically, as will be illustrated on the retention indices of alkanes and alcohols.

Table 2. Construction of Variable Connectivity Index

Bond	Connectivity Index	Variable Connectivity Index
1-2	$1/\surd(1 \cdot 3)$	$1/\surd(1 + x)\,(3 + x)$
2 – 3	$1/\surd(3 \cdot 3)$	$1/(3 + x)$
3 – 4	$1/\surd(2 \cdot 3)$	$1/\surd(2 + x)\,(3 + x)$
4-5	$1/\surd(1 \cdot 2)$	$1/\surd(1 + x)\,(2 + x)$
2-6	$1/\surd(1 \cdot 3)$	$1/\surd(1 + x)\,(3 + x)$
3-7	$1/\surd(1 \cdot 3)$	$1/\surd(3 + x)\,(1 + y)$
$^{1}\chi = 1/\surd 3 + 1/\surd 9 + 1/\surd 6 + 1/\surd 2 + 1/\surd 3 + 1/\surd 3 = 3.180739$		
$^{1}\chi^{f} = 1/\surd(1 + x)\,(3 + x) + 1/(3 + x) + 1/\surd(2 + x)\,(3 + x) + 1/\surd(1 + x)\,(2 + x) + {}$ $+ 1/\surd(1 + x)\,(3 + x) + 1/\surd(3 + x)\,(1 + y) = f(x,y)$		

Table 3. Variation of $^{1}\chi^{f}$ as a function of y

Y	$^{1}\chi^{f}$	y	$^{1}\chi^{f}$	y	$^{1}\chi^{f}$
- 0.99	8.376892	- 0.50	3.419886	2	2.936722
- 0.95	5.185378	- 0.25	3.270056	5	2.861599
- 0.90	4.429131	0	3.180739	10	2.777467
- 0.80	3.894383	0.25	3.119787	100	2.660837
- 0.70	3.657481	0.5	3.074793	1000	2.621646
- 0.60	3.516260	1	3.011637	∞	2.603389

Property – property regressions

In Fig. 1 we show regression between retention indices of alcohols versus the retention indices of alkanes having the same skeletal forms. This is a property-property regression. A parallelism between experimental properties for corresponding compounds suggests that the same descriptor may characterize the retention indices for both sets of compounds when considered separately.

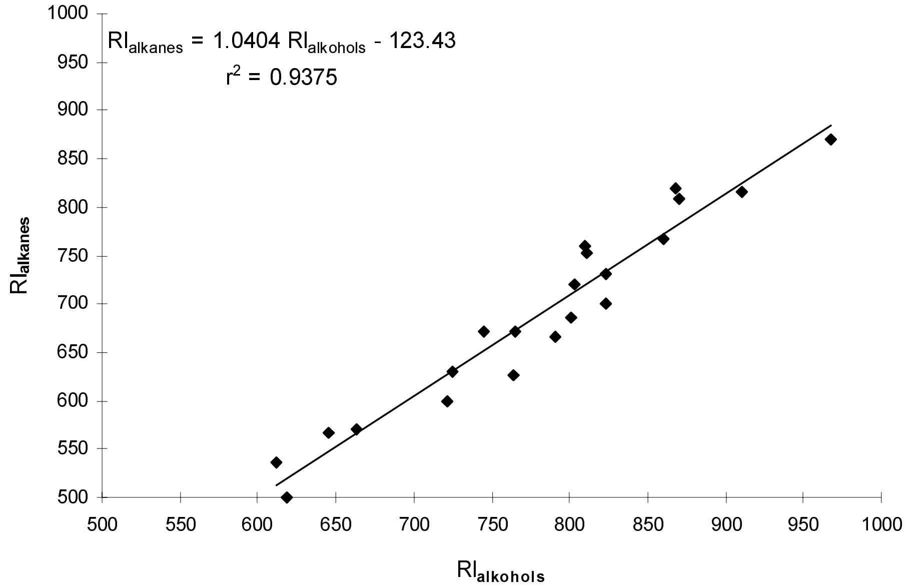

Figure 1. Regression between the retention indices of alcohols versus the retention indices of alkanes having the same skeletal forms

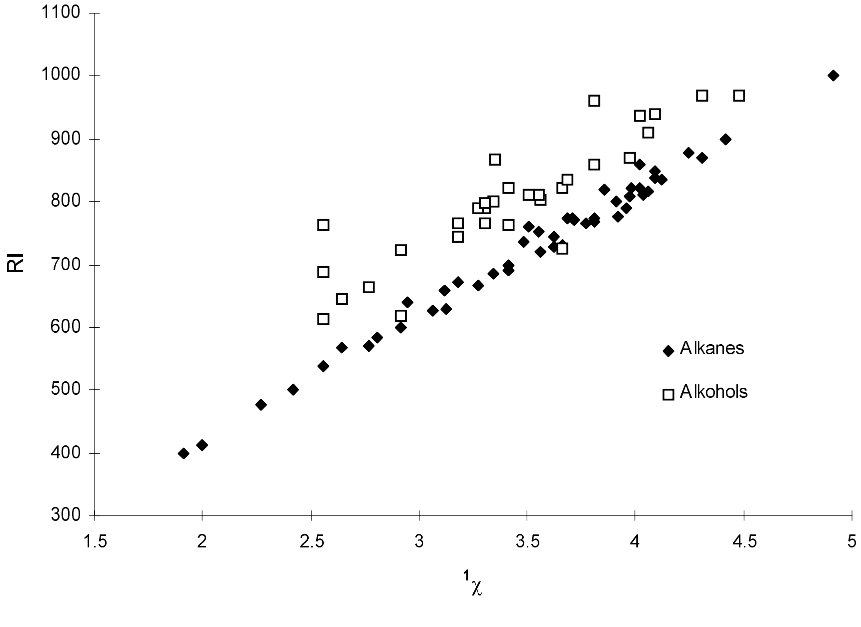

Figure 2. The regression of retention indices RI for alkanes (n=48) and alcohols (n=31) against the connectivity index $^1\chi$

In Fig. 2 we show that indeed it is possible to describe RI of alkanes and alcohols by the same molecular descriptor, the connectivity index. The plot of Retention Indices (RI) for alkanes gives regression of very high quality: the correlation coefficient r=0.9930, the standard error of estimate s =15.25, and the Fisher ratio F = 3234. In the case of alcohols the correlation with the connectivity index $^{1}\chi$ is somewhat less satisfactory, as could be expected in view that the connectivity index does not discriminates carbon and oxygen atoms. The corresponding statistical parameters are: r=0.8888, s = 45.98, and F =109. It is not difficult to identify the points corresponding to alkanes, which make a very good linear regression and points corresponding to alcohols that lie above and showing a greater scatter.

Search for optimal variable weights

Because the regression of RI against $^{1}\chi$ for alkanes is very good we will consider only variation of y, the weight describing oxygen atom, and will keep x = 0, which for alkanes reduces the variable index $^{1}\chi^{f}$ to the simple connectivity index $^{1}\chi$. In Table 4 we show variations of the correlation coefficient r, standard error s, and Fisher ratio F as a function of y. As y approaches the value - 0.70 the standard error s approaches the minimum value 14.24. For comparison we included in the last line in the table the corresponding statistical parameters when the connectivity index $^{1}\chi$ is used as descriptor that is when carbon atoms and oxygen atoms are not differentiated. However, when the distinction between carbon and oxygen is made the standard error has been reduced by factor of four.

Table 4. Variations of the correlation coefficient r, the standard error s, and the Fisher ratio F as a function of y.

weight	r	s	F
-0.90	0.8780	59.01	259
-0.80	0.9748	27.49	1471
-0.75	0.9887	18.47	3355
-0.70	0.9933	14.24	5695
-0.65	0.9928	14.73	5314
-0.60	0.9895	17.80	3616
0.00	0.8891	56.44	290

When y = - 0.70 the scattered points belonging to alcohols have shifted towards those of alkanes leading to very good regression line shown in Fig. 3.

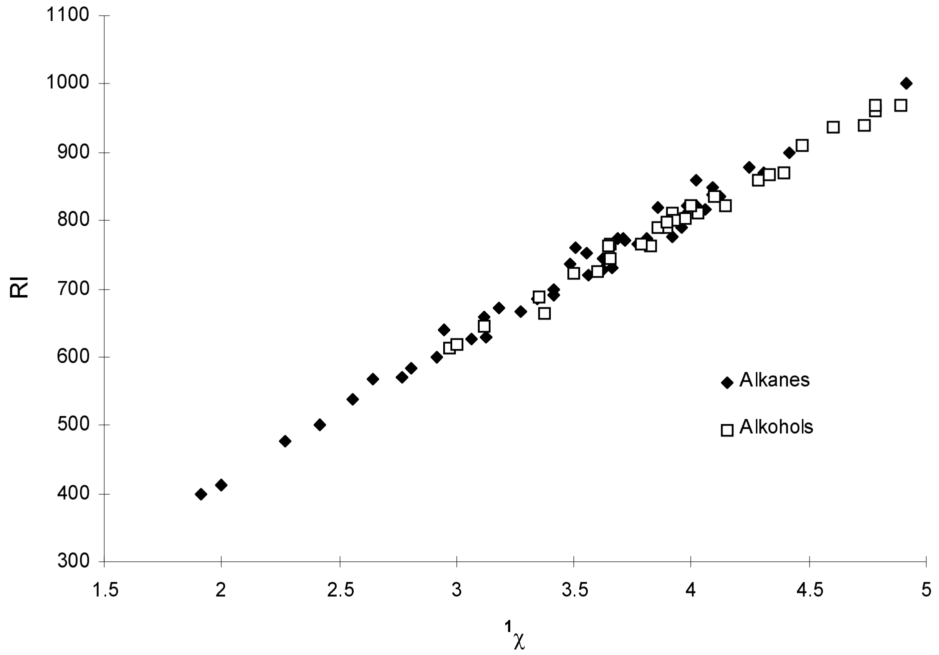

Figure 3. Correlation of RI for alkanes and alcohols if optimal weight (y =-0.70) for oxygen is used.

The following regression coefficient r, the standard error s, and the Fisher ratio F were obtained:

n = 78 r = 0.9933 s =14.24 F = 5695

The following linear regression equation is obtained:

RI =193.3894 (± 2.5625) $^1\chi^f$ + 41.1493 (± 9.5772).

The quadratic correlation:

RI = 227.7620 (± 19.7045) $^1\chi^f$ - 4.8957 (± 2.7833) $(^1\chi^f)^2$ -17.1215 (± 34.4494).

makes but a minor improvement:

n = 78 r = 0.9936 s =14.05 F = 2927

From the linear and the quadratic equation one can construct the regression equation:

RI =193.38941 $^1\Omega$ - 4:8957 $^2\Omega$ + 41.1493 associated with orthogonalized descriptors $^1\Omega$ and $^2\Omega$, where $^1\Omega$ is $^1\chi^f$ and $^2\Omega$ is the residual of the regression of $(^1\chi^f)^2$ against $^1\chi^f$.[13-16]

Table 5 The retention indices RI, connectivity indices for x = 0, y = 0, and x = 0, y = -0.70, the calculated retention indices RIcalc, and the residuals

ID	molecule	RI	0, 0	0, -0.70	RIcalc	Residual
1	22MM3	412.32	2.00000	2.00000	427.93	-15.61
2	2M4	475.28	2.27006	2.27006	480.15	-4.87
3	22MM4	536.80	2.56066	2.56066	536.35	0.45
4	23MM4	567.28	2.64273	2.64273	552.23	15.05
5	2M5	569.68	2.77006	2.7700(	576.85	-7.17
6	3M5	584.24	2.80806	2.80806	584.20	0.04
7	22MM5	625.60	3.06066	3.06066	633.05	-7.45
8	24MM5	629.84	3.12590	3.12590	645.67	-15.83
9	223MMM5	639.68	2.94338	2.94338	610.37	*29.31
10	33MM5	658.88	3.12132	3.12132	644.78	14.10
11	2M6	666.56	3.27006	3.27006	673.54	-6.98
12	23MM5	671.68	3.18074	3.18074	656.27	15.41
13	3E5	686.00	3.34607	3.34607	688.24	-2.24
14	224MMM5	689.92	3.41650	3.41650	701.86	-11.94
15	22MM6	719.36	3.56066	3.56066	729.74	-10.38
16	25MM6	728.40	3.62590	3.62590	742.36	-13.96
17	24MM6	731.92	3.66390	3.66390	749.71	-17.79
18	223MMM5	737.12	3.48138	3.48138	714.41	22.71
19	33MM6	743.52	3.62132	3.62132	741.47	2.05
20	234MMM5	752.40	3.55342	3.55342	728.34	24.06
21	233MMM5	759.36	3.50404	3.50404	718.79	**40.57
22	2M7	764.88	3.77006	3.77006	770.24	-5.36
23	4M7	767.20	3.80806	3.80806	777.59	-10.39
24	34MM6	770.56	3.71874	3.71874	760.31	10.25
25	3M7	772.32	3.80806	3.80806	777.59	-5.27
26	2244MMMM5	772.72	3.70711	3.70711	758.07	14.65
27	33ME5	774.00	3.68198	3.68198	753.21	20.79
28	225MMM6	776.32	3.91650	3.91650	798.56	-22.24
29	224MMM6	789.12	3.95451	3.95451	805.91	-16.79
30	244MMM6	808.72	3.97716	3.97716	810.29	-1.57
31	235MMM6	812.00	4.03658	4.03658	821.78	-9.78
32	22MM7	815.36	4.06066	4.06066	826.44	-11.08
33	2234MMMM5	819.60	3.85406	3.8540fi	786.48	*33.12
34	223MMM6	821.60	3.98138	3.98138	811.11	10.49
35	223MME5	822.24	4.01939	4.01939	818.46	3.78
36	33MM7	835.76	4.12132	4.12132	838.17	-2.41
37	234MEM5	836.48	4.09142	4.09142	832.39	4.09
38	234MMM6	849.12	4.09142	4.09142	832.39	16.73
39	2334MMMM5	858.00	4.01651	4.01651	817.90	**40.10
40	3M8	870.24	4.30806	4.30806	874.28	-4.04
41	33EE5	877.20	4.24264	4.24264	861.63	15.57

Table 5. Continued

ID	molecule	RI	0, 0	0, -0.70	RIcalc	Residual
42	4	400.00	1.91421	1.91421	411.34	-11.34
43	5	500.00	2.41421	2.41421	508.03	-8.03
44	6	600.00	2.91421	2.91421	604.73	-4.73
45	7	700.00	3.41421	3.41421	701.42	-1.42
46	8	800.00	3.91421	3.91421	798.12	1.88
47	9	900.00	4.41421	4.41421	894.81	5.19
48	10	1000.00	4.91421	4.91421	991.51	8.49
49	2M2OH4	612.00	2.56066	2.97353	616.20	-4.20
50	1OH4	619.04	2.91421	2.99810	620.95	-1.91
51	3M2OH4	645.04	2.64273	3.11948	644.42	0.62
52	2OH5	663.04	2.77006	3.37655	694.14	*-31.10
53	3M2OH5	765.04	3.18074	3.65748	748.47	16.57
54	3M1OH4	689.04	2.56066	3.35394	689.77	-0.73
55	4M2OH5	724.96	3.66390	3.60264	737.86	-12.90
56	1OH5	722.00	2.91421	3.49810	717.64	4.36
57	2M3OH5	744.96	3.18074	3.65748	748.47	-3.51
58	24MM2OH5	762.00	3.41650	3.82937	781.71	-19.71
59	33MM1OH4	764.00	2.56066	3.64455	745.97	18.03
60	3OH6	766.00	3.30806	3.78480	773.09	-7.09
61	2M2OH6	803.04	3.56066	3.97353	809.59	-6.55
62	2MlOH5	790.00	3.30806	3.89195	793.81	-3.81
63	24MM3OH5	811.04	3.55342	4.03016	820.54	-9.50
64	4M1OH5	790.96	3.27006	3.85394	786.46	4.50
65	23MM3OH5	810.00	3.50404	3.91691	798.64	11.36
66	2E1OH4	801.04	3.34607	3.92995	801.16	-0.12
67	3M1OH5	798.00	3.30806	3.89195	793.81	4.19
68	5M3OH6	823.04	3.66390	4.14064	841.91	-18.87
69	3E3OH5	834.00	3.68198	4.09485	833.05	0.95
70	1OH6	823.04	3.41421	3.99810	814.34	8.70
71	4OH7	860.00	3.80806	4.28480	869.78	-9.78
72	224MMM3OH5	868.00	3.35406	4.33080	878.68	-10.68
73	35MM3OH6	870.00	3.97716	4.39003	890.13	-20.13
74	2M2OH7	911.04	4.06066	4.47353	906.28	4.76
75	6M2OH7	936.96	4.01651	4.60264	931.25	5.71
76	4E3OH6	938.96	4.09266	4.73349	956.56	-17.60
77	4OH8	960.00	3.80806	4.78480	966.48	-6.48
78	3OH8	968.00	4.30806	4.78480	966.48	1.52
79	36MM3OH7	970.00	4.47716	4.89003	986.83	-16.83

M = methyl; E = ethyl; OH = alcohol; 3 = propane; 4 = butane; etc.

In Table 5 we have listed the experimental retention indices (RI), the connectivity index ${}^{1}\chi$ and the optimal connectivity index ${}^{1}\chi^{f}$ used to calculate RI. The last column of Table 5 gives the residuals for the linear regression. A closer look at the residuals indicates that only two show somewhat larger departure from the regression, # 21, 2,3,3-

trimethylpentane and # 39, 2,3,3,4-tetramethylpentane (indicated by the double asterisks in Table 5), for another three compounds the residuals are marginally close to 2s, twice the standard error. It is significant that all the "outliers" are alkanes, except # 52, which is 2-pentanol. This suggests that variation of x may further improve the regression and reduce the residuals for alkanes.

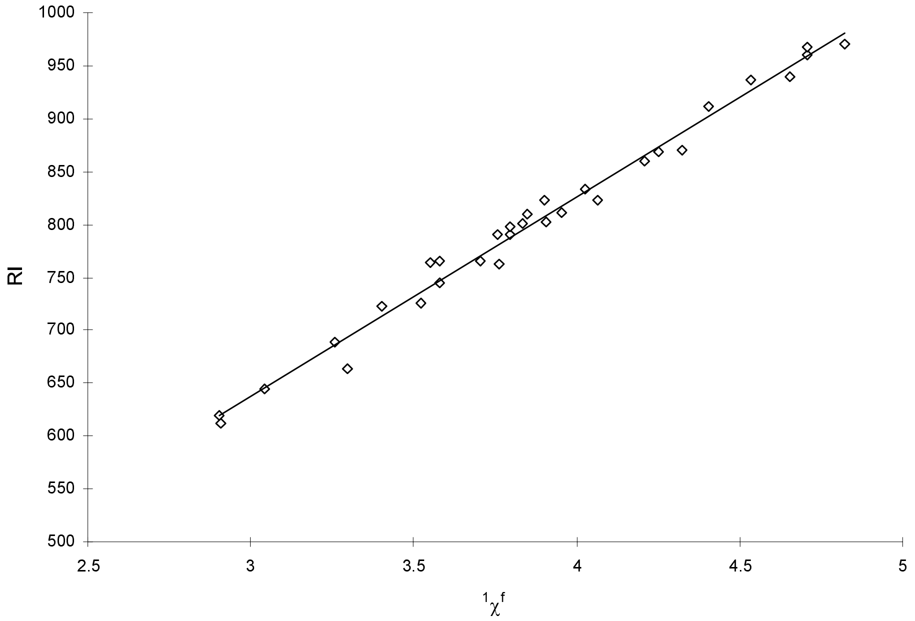

Figure 4. Correlation of RI of alcohols using optimal $^1\chi^f$ (x = 0, y = -0.65)

In Fig. 4 we have plotted the correlation of RI for alcohols alone using the optimal connectivity index $^1\chi^f$ (with x = 0, y = -0.65) in order to illustrate the effect of discrimination of carbon and oxygen atoms. Fig. 4 should be compared with Fig. 2 in order to see how $^1\chi^f$ was able to reduce the scatter of points of Fig. 2 and result in very high quality regression. The regression of Fig. 4 has the following statistical parameters: n=31, r =0.9936, s = 11.32; F = 2249. If we compare the above with the statistical parameters characterizing the regression of alkanes alone we see that now correlation of alcohols alone yielded a smaller standard error, which further support that variation of x, the weight for carbons, may improve the results somewhat.

As we can see from Fig. 4 the variable weight reduced dramatically the scatter of points for alcohols. Moreover, it also has brought the points of alcohols in line with the correlation of points corresponding to alkanes, as we can see by comparing Fig. 2 and

Fig. 3. Hence, we succeeded to describe the gas chromatographic retention of alkanes and alcohols by a single descriptor.

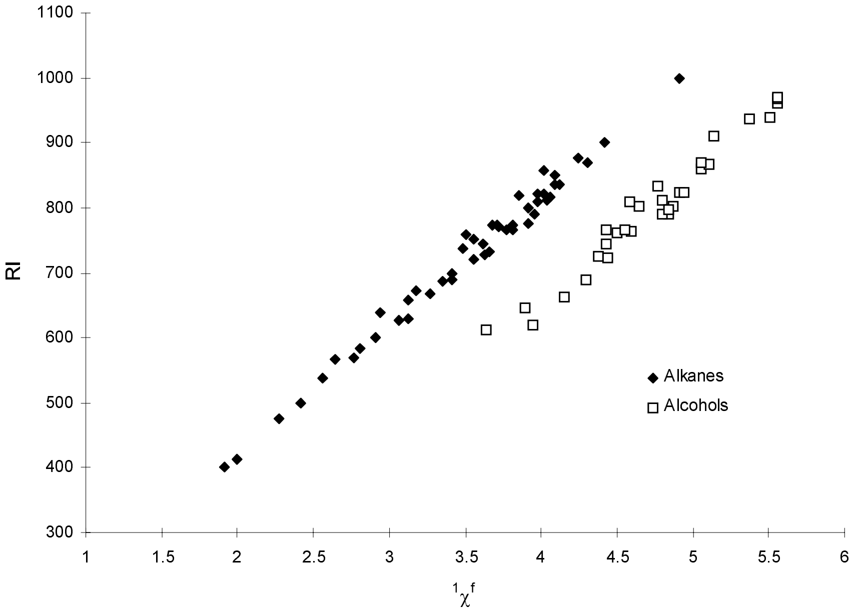

Figure 5. Correlation of RI for alkanes and alcohols showing a separately regressions because not optimal weight for oxygen was used.

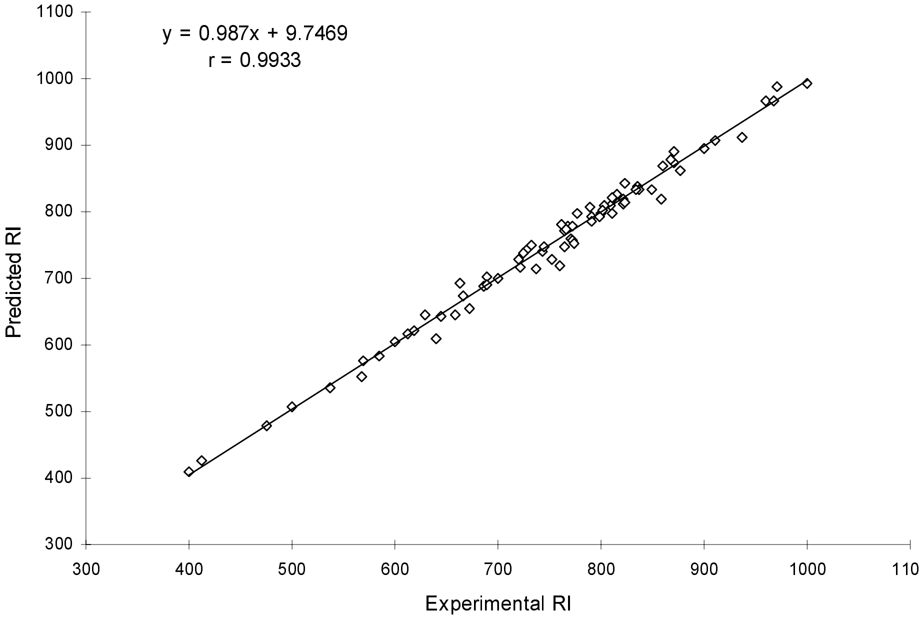

Figure 6. Calculated Retention Indices against the experimental Retention Indices

In order to better see how was this possible we illustrate in Fig. 2 the "working" of the flexible descriptor $^{1}\chi^{f}$. Clearly some differentiation between carbon atoms and oxygen atoms was essential for deriving high quality regression. However, as we can see from Fig. 5, by increasing the weight y for oxygen beyond the optimal value the correlation lines for alkanes and alcohols do not coincide. Fig. 5 illustrates the combined regression of RI of alkanes and alcohols by assuming y = - 0.90. As we see from Fig. 5 the points corresponding to alcohols have moved too far to the right. A more negative value of the variable y then needed results in even larger increase of the contribution of oxygen to the connectivity descriptor then required. The optimal value of y is one that leads to an overlap of the separate linear regressions for alkanes and alcohols.

In Fig. 6 we show the plot of calculated Retention Indices (RI) against the experimental RI. The corresponding statistical parameters are: r = 0.9933, s =14.14, and Fisher ratio F = 5695.

Conclusions

We have seen how modification of traditional topological indices can enormously increase their power in quantitative structure-property relationship (QSPR) and QSAR studies. In particular use of a single descriptors, the variable connectvtiy index, gave a very high quality regression for chromatographic RI. From the present study we have seen that the role of oxygen atoms is significantly more important for chromatographic retention times than the role of carbon atom. From Table 3, and the part of Table 5 corresponding to alcohols, we see that contribution of oxygen atoms has been by about four times greater than the contributing carbon atoms.

In comparison with the traditional MRA studies, for which typically variable connectivity index can replace three to four topological indices based on fixed numerical values, we see advantages of the variable indices not only for improving the statistical quality of the regression but also making the interpretation of the results possible and more meaningful.

Acknowledgment

This is contribution number 279 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by Grant F49620-98-1-0015 from the U. S. Air Force and by Ministry of Science and Technology of the Republic of Slovenia.

References and notes

1. M. Randić, Topological Indices, "*The Encyclopedia of Computational Chemistry*," Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer III, H. F.; Schreiner, P. R (Eds.); John Wiley & Sons: Chichester, 1999, pp. 3018-3032.
2. M. Karelson, V. S. Lobanov, A. R. Katritzky, Qunatum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027-1043.
3. R. Todeschini, V. Consonni, *The Handbook of Molecular Descriptors,* Series of Methods and Principles in Medicinal Chemistry - Vol. 11, Eds. R.Mannhold, H. Kubinyi, G. Timmerman, Wiley-VCH, New York, 2000.
4. M. Randić, Novel Graph Theoretical Approach to Heteroatom in Quantitative Structure-Activity Relationship, *Chemometrics Intel. Lab. Syst.* **1991**,*10*, 213-227.
5. M. Randić, On Computation of Optimal Parameters for Multivariate Analysis of Structure-Property Relationship, *J. Comput. Chem*. **1991**, *12*, 970-980.
6. M. Randić and J. Cz. Dobrowolski, Optimal Molecular Connectivity Descriptors for Nitrogen-Containing Molecules, *Int. J. Quant. Chem.* **1998**, *70*, 1209-1215.
7. M. Randić, High quality structure-property regressions. Boiling points of smaller alkanes, *New J. Chem*. **2000**, *24*, 165-171.
8. M. Randić and S. C. Basak, On construction of high quality structure- property-activity regressions. Boiling points of sulfide, *J. Chem. Inf. Comput. Sci.* (submitted)
9. M. Randić, D. Mills, S. C. Basak, On characterization of physical properties of amino acids, *Int. J. Quantum. Chem.* **2000**, *80*, 1199-1209.
10. M. Pompe and M. Novič, Prediction of gas-chromatographic retention indices using topological descriptors, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 59.
11. M. Randić, On the characterization of molecular branching, *J. Am. Chem. Soc*. **1975**, *97*, 6609-6615.
12. A. R. Katritzky, V. Lobanov, and M. Karelson, CODESSA (COmprehensive DEscriptors for Structural and Statistical Analysis), University of Florida, Gainesville.
13. M. Randić, Orthogonal molecular descriptors, *New J. Chem* . **1991**,*15*, 517-525.
14. M. Randić, Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J. Chem. Inf Compuf. Sci*. **1991**, *31*, 311-370.
15. M. Randić, Fitting of non linear regressions by orthogonalized power series. *J. Comput., Chem*. **1993**, *14*, 363-370.
16. M. Randić, Curve fitting paradox. *Int. J. Quant. Chem: Quant. Biol. Symp* . **1994**, *21*, 215-225.

Povzetek

Pri našem delu smo analizirali kodirne sposobnosti variabilnega indeksa povezanosti (${}^1\chi^f$) za napovedovanje retencijskih indeksov dobljenih s plinsko kromatografijo. Z uporabo linearne regresije smo retencijske indekse napovedovali za 79 spojin, ki so vsebovale 48 alkanov in 31 alkoholov. S spreminjanjem uteži za kisikov atom v variabilnem indeksu povezanosti smo dobili linearni regresijski model z naslednjimi karakteristikami: korelacijski koeficient r=0.9933, standardna napaka s=14,24 retencijske enote in Fisherjevim koeficientom F=5695. Z uporabo klasičnega indeksa povezanosti, ki ne razlikuje med kisikovimi in ogljikovimi atomi, smo dobili štirikrat večjo standardno napako.