Molecules **2004**, *9*, 1177-1193

molecules

ISSN 1420-3049
http://www.mdpi.org

Variable Connectivity Index as a Tool for Modeling Structure-Property Relationships

Milan Randić [1,*], Matevž Pompe [2], Denise Mills [3] and Subhash C. Basak [3]

[1] National Institute of Chemistry, Ljubljana, Slovenia and 3225 Kingman Rd. Ames, IA 50014, USA.
[2] Department of Chemistry and Chemical Technology, The University of Ljubljana, 1000 Ljubljana, Aškerčeva 5, Slovenia, E-mail: matevz.pompe@uni-Lj.si;
[3] Natural Resources Research Institute, 5013 Miller Trunk Hwy, Duluth, MN, USA. E-Mail: dmills@nrri.umn.edu and sbasak@nrri.umn.edu

* Author to whom correspondence should be addressed; E-mail: mrandic@msn.com

Received: 2 July 2004 / Accepted: 14 October 2004 / Published: 31 December 2004

Abstract: We report on the calculation of normal boiling points for a series of n = 58 aliphatic alcohols using the variable connectivity index in which variables x and y are used to modify the weights on carbon (x) and oxygen atoms (y) in molecular graphs, respectively. The optimal regressions are found for x = 0.80 and y = -0.90. Comparison is made with available regressions on the same data reported previously in the literature. A refinement of the model was considered by introducing different weights for primary, secondary, tertiary, and quaternary carbon atoms. The standard error in the case of the normal boiling points of alcohols was slightly reduced with optimal weights for different carbon atoms from s = 4.1°C (when all carbon atoms were treated as alike) to s = 3.9 °C.

Keywords: Flexible molecular descriptors, Variable connectivity index, Aliphatic alcohols, Normal boiling points.

Introduction

As can be seen from browsing through the literature despite the availability of various methodologies, such as the Principal Component Analysis (PCA), the Ridge Regression (RR), the Partial Least Square (PLS), and the Artificial Neural Networks (ANN), the Multivariate Regression Analysis (MRA) continues to be widely used in studies of structure-property relationships. The situation is likely to continue because different methodologies have different advantages and limitations. With respect to MRA, in recent years we witnessed several important developments associated with molecular descriptors that continue to keep MRA as an active route for structure-property studies. These developments include:

(1) Development of numerous novel molecular descriptors;
(2) Development of user-friendly computer software for MRA;
(3) Development of orthogonalization procedure for descriptors;
(4) Development procedures for interpretation of descriptors;
(5) Development of "flexible molecular descriptors."

Here, we will be concerned with the use of flexible molecular descriptors in MRA, a development that started a dozen years ago and appears to be catching up in recent time. However, before entering the domain of flexible descriptors, let us briefly elaborate on the remaining four points listed above because those interested in structure-property-activity studies should consider and, if possible, combine all the five aspects of MRA in order to arrive at better models for use in QSAR and QSPR (Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships, respectively).

For an exhaustive compilation of numerous molecular descriptors one should consult the Handbook of Molecular Descriptors by Todeschini and Consonni [1], while a recent book by Devillers and Balaban [2] lists numerous articles on novel molecular descriptors. The development of novel topological indices (TIs) goes in two directions:

(1) Generalization of functional dependencies between distance or adjacency matrices and various properties and
(2) development of molecular descriptors designed for heteroatoms.

In spite of the great number of successful applications of topological indices in QSAR and QSPR studies, their structural interpretation still remains the main bottle-neck in the understanding of their nature. Due to a large number and wide diversity of graph invariants different TIs, it is difficult to find general relations. Thus it is desirable to obtain some generalization of TIs that makes further interpretation more effective. Lucic *et al.* generalized Wiener index for the modeling of boiling points [3]. Estrada proposed a generalization of several well-known TIs by using a vector-matrix-vector multiplication procedure [4-7].

A number of molecular descriptors that have been designed for use with molecules containing heteroatoms can be related to several graph theoretical indices (referred often as topological indices) and can be viewed as specifically generalized topological indices. Examples include the well-known valence connectivity indices of Kier and Hall [8, 9], which are derived from the connectivity index [10] and the "higher order" connectivity indices [11], descriptors derived from distance matrix of molecules containing heteroatoms [12], Balaban's topological index J for heteroatom-containing molecules [13-15] derived from the topological index J [16], Zefirov and Palyulin's "solvation connectivity index" [17-19]. We will see later how the variable connectivity index may be viewed not merely as a generalization of the "valence" connectivity indices but as an "unlimited" collection of "valence" connectivity indices.

As a user-friendly software for MRA we should mention CODESSA, developed by Katritzky, Lobanov and Karelson [20]. In CODESSA, there are some 400 molecular descriptors that can be evaluated for a given set of structures and used in MRA.

Orthogonalization of molecular descriptors, a procedure that is only slowly gaining increased attention, has several important properties for MRA [21-24]. First, it leads to a stable step-wise regression equation in which the coefficients of the already used descriptors do not change when additional (orthogonal) descriptors are added. Thus, this allows one to view the coefficients of the regression equation as indicators of the relative importance of the descriptors used. It also allows one to condense the structural information contained in two or more descriptors into a single descriptor [25, 26]. Finally, it also allows one, by using retro-regression [27], to find the best subset of descriptors for regression using *n* descriptors.

Recently a relatively general procedure has been outlined which allows one to partition molecular descriptors in terms of bond contributions [28, 29]. Such partitioning provides some insight into the relative importance of the individual bonds for the particular molecular property. When this approach is applied to different topological indices, one sees that, for some indices, terminal CC bonds play a more important role and the interior CC bonds make lesser contributions, while in other indices the opposite is the case [30, 31]. This, of course, is important for a better understanding of molecular models.

Flexible Molecular Descriptors

Typical MRA analysis starts with a selection of molecular descriptors from a pool of available descriptors. For instance, as already mentioned, CODESSA offers several hundreds of descriptors to choose from, which include indicator variables, quantum chemically computed quantities, and graph theoretically derived or modified molecular descriptors. The procedure and criteria for the *selection* of descriptors and alternatives is a non-trivial problem in view of interrelation of many descriptors. This is an important topic that continues to receive attention and may well result in diverse approaches yet to be fully evaluated. In contrast to approaches based on choosing between "fixed" molecular descriptors, the notion of "flexible" molecular descriptors varies descriptors already used by modifying them slightly and thus creating novel descriptors to be tested for their performance. The process continues till, for a given type of descriptor, one has explored the domain of the variable parts of the descriptors and found optimal parameters that minimize the overall standard error of the regression.

The first variable descriptor was the variable connectivity index [32, 33]. The index is constructed by introducing one or more variables associated with atoms of different kinds or different types. In this way, the bond (m,n) that connects atoms having m and n nearest neighbors, respectively, (which contributed $1/\surd$ (m·n) to the connectivity index $^1\chi$) makes the contribution $1//\surd$ [(m+x)(n+y)], where x and y are variables to be selected during the regression analysis. For example, in a study of normal boiling points of smaller aliphatic alcohols [28], it was found that the standard error was reduced from 7.9°C when the connectivity index $^1\chi$ is used, that is when x = y = 0, to 3.3°C when x = 1.50 and y = -0.85. This is quite an impressive improvement in the simple regression for the normal boiling points of smaller alcohols. Similar results were obtained when the normal boiling points of smaller amines were considered [34], when it was found that the standard error was reduced from 3.5°C when the connectivity index $^1\chi$ was used, (that is when x = y = 0), to 1.9°C when x = 1.25 and y = -0.65. These two cases well illustrate the "power" of the variable connectivity index and suggest by extension that, in general, variable indices are likely to offer significant improvements in the regression analysis.

Recently, in addition to the use of variable connectivity indices [35-43], several different flexible molecular descriptors have been considered in the literature, including the use of variable paths weights [44, 45], variable distance related indices [46, 47], and other variable descriptors [48]. Thus, we may well be at the beginning of a novel direction in structure-property-activity studies that will be dominated by the use of variable molecular descriptors that not only lead to better regressions with fewer molecular descriptors in comparison with regressions using "fixed" descriptors, but, because variable descriptors offer novel interpretation to the descriptors [49], may also offer guidance in the refinement of molecular models, as will be explained later in this article. An important consequence of the introduction of the variable molecular connectivity index should not be overlooked, however: variable connectivity not only clearly points to limitations of the "valence"-based molecular descriptors but has actually demonstrated that there are no "universal" valence descriptors, and that, at best, they may be suitable for a narrow selection of molecular properties.

Normal boiling Points of Aliphatic Alcohols Revisited

We have selected a set of 58 aliphatic alcohols, previously considered in the literature, in order to compare the performance of different variable descriptors to that of the variable connectivity index. One of the earlier studies is based on use of path numbers in which variable weight has been introduced for paths that include the oxygen atom [45]. Using paths of length one and two, the optimal variable weight, x, is 2.6, and the associated standard error, s, is 4.0°C. When paths of length three are also included, the optimal variable weight increases slightly to 3.1, with an associated standard error of 3.9°C.

Recently Krenkel, Castro and Toropov [48] reported on their analysis of the same 58 aliphatic alcohols using "CW"(correctional weight) descriptors. In the upper part of Table 1, we have collected the statistical parameters for several multivariate regression analyses of the normal boiling points of alcohols, and in the lower part of Table 1 we have summarized the results of the present study on the same set of alcohols. Observe the distinction between "descriptors" and "variables," the former indicating the number of molecular descriptors that appear in the regression equation and the latter indicating the "flexibility" of the descriptors, which is the inherent power to adjust and modify so that they can account for different kinds of atoms or different environments for similar atoms.

Let we see an example. If we use MRA the number of descriptor represent the number of parameters of the model minus one if the constant is used (eq. 1).

$$(\text{Property}) = (\text{coef. } 1)\ (\text{desc. } 1) + \ldots + (\text{coef. } n)\ (\text{desc. } n) + (\text{const.}) \qquad (1)$$

We could see that the number of parameters for the above model is equal n+1. The flexibility of model is achieved by selection of certain number of descriptors from the large pool of available descriptors. In the case of variable topological indices usually simple linear regression is used (eq. 2) and the number of parameter of the model is equal 2.

$$(\text{Property}) = (\text{coef.})\ (\text{desc.}(\text{variable } 1, \text{variable } 2, \ldots)) + (\text{const.}) \qquad (2)$$

We can see that the number of variables has no influence on the number of parameters of the model, because variables represent just the flexibility of the single descriptor. It is true that by increasing the number of variables we are increasing the possibility to obtain a chance correlation but the same is true also in the case of MRA where chances to obtain a random correlation are increased by using larger pool of possible descriptors.

The correlations using weighted paths involve several molecular descriptors, while the regression of Krenkel, Castro and Toropov [48] as well as all of the models using the connectivity index are simple regressions using one molecular descriptor.

Table 1. Statistical data on regression of the normal boiling points of alcohols. We grouped data from different sources ending with results from this study

	Descriptors	**Variables**	**Standard error**	**Correlation coefficient**	**Outliers**	**Ref.**
Weighted paths p_1, p_2	2	1	4.1 °C	0.994	0	[45]
Weighted paths p_1, p_2	2	1	3.6 °C	0.995	3	[46]
Weighted paths p_1, p_2, p_3	3	1	3.6 °C	0.995	0	[45]
Weighted paths p_1, p_2, p_3*	3	2	2.9 °C	0.997	3	[46]
Correlation weights	1	7			0	[48]
Linear		Training	2.9 °C	0.995	0	
		Test	3.0 °C	0.995		
Quadratic		Training	3.0 °C	0.995	0	
		Test	2.8 °C	0.995		
Cubic		Training	2.9 °C	0.995	0	
		Test	2.9 °C	0.995		
Connectivity index	1	2	4.1 °C	0.994	0	This work
Connectivity index	1	2	3.5 °C	0.995	2	This work
Connectivity index	1	2	3.2 °C	0.996	3	This work
Connectivity index	1	2	2.9 °C	0.997	5	This work
Connectivity index	1	5	3.9 °C	0.994	0	This work
Connectivity index	1	5	3.2 °C	0.996	2	This work
Connectivity index	1	5	2.9 °C	0.997	3	This work
Connectivity index	1	5	2.6 °C	0.997	5	This work

There is no doubt that the approach of Krenkel, Castro and Toropov [48] gave better results that those of [45], evidenced not only by the smaller standard error and fewer outliers but also by the fact that they used half of the data for the "training set" and half of the data as the "prediction set." However, their results are not totally surprising in view of the fact that Castro *et al* employed four different weights for primary, secondary, tertiary and quaternary carbon atoms, in addition to using three weights for different atomic types (hydrogen, carbon and oxygen). It may be of interest to consider how much the introduction of different weights for paths of different lengths may improve the results based on variable path numbers, although this remains to be seen. Interest in such a comparison arises from the fact that, on one hand, one uses several variable descriptors (paths of different lengths) in MRA and, on the other hand, one uses a single variable descriptor (CW descriptor) in a simple regression but expressed with twice as many variables. Future applications are likely to show how these two alternative approaches complement each other, how much they overlap, and how much they differ in various applications.

We see from Table 1 that on using a single variable connectivity index that discriminates between carbon atoms and oxygen atoms (the variable weights for carbon and oxygen are 0.80 and -0.90, respectively) one obtains a regression correlation of similar quality as those based on two and three variable path numbers. Thus, in this respect, the variable connectivity indices are more "powerful" descriptors than the variable path numbers. We should add, that although the novel "correlation weights" indices do offer a good regression, the indices themselves show high degeneracy, with about half of the molecules considered having the same numerical values for their indices, as indicated in Table 2 below, and hence the same applies to their computed normal boiling points:

Table 2. Degeneracy of the Correlation Weights descriptors of smaller alcohols

Alcohol	Exp.	Calc.	Alcohol	Exp.	Calc.
2-pentanol	119.0	117.13	2-M-2-hexanol	142.5	143.30
3-pentanol	115.3	117.13	3-M-3-hexanol	142.4	143.30
2-hexanol	139.9	136.57	3-E-3-pentanol	142.5	143.30
3-hexanol	135.4	136.57	2,3-MM-2-pentanol	139.7	136.84
2-M-1-pentanol	148.0	148.68	2,3-MM-3-pentanol	139.0	136.84
3-M-1-pentanol	152.4	148.68	3,3-MM-2-pentanol	133.0	137.43
4-M-1-pentanol	151.8	148.68	2,3-MM-3-pentanol	136.0	137.43
2-E-1-butanol	146.5	148.68	2-nonanol	198.5	194.89
2-M-2-pentanol	121.4	123.86	3-nonanol	194.7	194.89
3-M-3-pentanol	122.4	123.86	4-nonanol	193.0	194.89
2,2-MM-1-butanol	136.8	136.55	5-nonanol	195.1	194.89
3,3-MM-1-butanol	143.0	136.55	2,6-MM-4-heptanol	178.0	181.99
3-heptanol	156.8	156.01	3,5-MM-4-heptanol	187.0	181.99
4-heptanol	155.0	156.01			

Degeneracy among topological indices is quite common and not necessarily detrimental because different compounds, having identical molecular descriptors, may also have identical or very similar magnitudes for selected molecular properties. In the case of alcohols, for instance, as we can see from

Table 2, this is the case with 2-methyl-2-hexanol, 3-methyl-3-hexanol and 3-ethyl-3-pentanol, having normal boiling points of approximately 142.4-142.5. However, there are cases wherein molecules have the same index but are reported to have significantly different normal boiling points, such as 3-methyl-2-pentanol and 2-methyl-3-pentanol, or 2,6-dimethyl-4-heptanol and 3,5-dimethyl-4-heptanol, where the two isomers differ by 7.7 ^{0}C and 9.0 ^{0}C. The variable connectivity index also shows degeneracy with respect to the set of 58 alcohols, but only for the isomers listed below:

Alcohol	**Exp.**	**Calc.**	**Alcohol**	**Exp.**	**Calc.**
2-M-1-pentanol	148.0	149.44	3-nonanol	194.7	195.85
3-M-1-pentanol	152.4	149.44	4-nonanol	193.0	195.85
3-heptanol	156.8	156.19	5-nonanol	195.1	195.85
4-heptanol	155.0	156.19			

In view of the very good regression reported by Krenkel, Castro and Toropov for their variable index, it appears that their index may have even greater potential than hitherto displayed if it could be further modified by introducing additional variables that could differentiate among the many isomers showing the degenerate numerical values for the index as currently calculated.

Another matter that needs to be addressed, when one uses variable descriptors or other descriptors selected from a large pool of descriptors, is the risk of “chance” correlation. The risk of “chance” correlation can be assessed by randomizing the input data and determining whether the descriptors provide a satisfactory regression for randomly assigned input data. If they do, then the descriptors used for the particular regression should be rejected; if they don’t, then the regression is significant. In view of the fact that we have a sizable sample (n = 58), one such random test may suffice to convince readers that the variable connectivity descriptors are encoding specific molecular structural features rather than adjusting to any meaningless set of input data. Using random number tables we have randomized normal boiling points. The order of the 58 normal boiling point entries is the following:

39	14	30	42	5	41	48	24	58	38	46	11	13	8	20
40	35	37	2	55	54	44	57	31	7	53	47	19	27	22
23	50	15	51	3	4	25	34	33	49	26	29	1	43	36
12	9	6	10	32	45	17	16	28	21	52	50	18		

By using the ordinary connectivity index (that is not differentiating carbon and oxygen atoms), which gives the standard error of s = 8.6°C and the regression coefficient r = 0.971 if true normal boiling points are modeled, we obtain after randomization of normal boiling points a simple regression with the standard error 90.7°C and the regression coefficient r = 0.044. This clearly points to the non-random nature of the connectivity index χ. In fact that the difference between the largest and the smallest normal boiling points (corresponding to 1-decanol and methanol with 230.3°C and 64.7°C, respectively) is less than twice the standard error of the randomized input entries shows inability of the connectivity index to fit random numbers. However, we have to show that the same is true also for the variable connectivity index if the results summarized in Table 1 are to hold unchallenged. Hence, we introduced two variables, variable x to modify the weights of the carbon atoms and variable y to modify the role of the oxygen atom. After a search for the optimal values for the two variables, we obtain a standard error and coefficient of regression of s = 89.9°C and r = 0.133, respectively, with x = 2 and y = 10^{15}. Thus, clearly the variable connectivity index, while having the flexibility to adjust to specific structural variations of molecules (alcohols), cannot adjust to fit random data. This ought to suffice to dispel concerns that flexible molecular descriptors can be made to suit random data.

Model Refinement

The variable molecular indices, including of course the variable connectivity index, have shown flexibility in adjusting to specific requirements that individual molecular properties may require and consequently produced high quality regressions. In addition, these indices have another important property that should be considered – they offer increased latitude in molecular modeling that was hitherto not so readily accessible. Variable descriptors make it possible to test modifications of existing models in order to find better models, as will be illustrated here using normal boiling points of alcohols modeled with the variable connectivity index.

As we have seen from Table 1, the variable connectivity index with two variables, one for carbon atoms (x) and one for the oxygen atom (y), produced the regression equation that is accompanied with the standard error of s = 4.1°C. The optimal values found for the variables are x = 0.80 and y = -0.90, which can be interpreted to mean that the carbon-oxygen bond has a greater role than the carbon-carbon bonds. One may now raise the question: does the enhanced role of the oxygen atom also influence the adjacent carbon atom? In other words, do the carbon atoms that are bonded to oxygen play a somewhat more important role in comparison with carbon atoms that are more distant from the oxygen atom?

With the notion of the variable connectivity index, the above question can be readily answered. All we need is to introduce an additional variable weight to be associated with the carbon atom that is adjacent to the oxygen. When we did this and varied three variables (two for carbon atoms and one for oxygen), we found no significant improvement in the regression. The conclusion is that there is no essential difference between carbon atoms, regardless of whether they are bonded to oxygen or not in the case of saturated compounds. There is, of course, a difference between carbon atoms having sp^3 and sp^2 hybridization in compounds having C=O bonds.

The next step in trying to improve the model based on discrimination between carbon atoms is to consider different weights for the primary, secondary, tertiary and quaternary carbon atoms. Indeed, when considering aqueous solubility of alcohols, Cammarata [50] observed that the primary, secondary, tertiary, and quaternary carbon atoms make somewhat different contributions to the total surface area, to be used as a structural descriptor for aqueous solubility of aliphatic alcohols. The variable connectivity index allows one to test if this is also the case with the normal boiling points of alcohols. Using four different weights for carbon atoms and a single variable for the oxygen atom, we have examined the normal boiling points of alcohols and obtained the results listed in the lower part of Table 1. As we can observe by differentiating the primary, secondary, tertiary, and quaternary carbon atoms, we were able to reduce the standard error from 4.1°C to 3.9°C, not dramatic but certainly a significant improvement. It turns out that the initial carbon "weights" of 0.80 have changed into: 0.80, 0.80, 0.96, 1.00, for the primary, secondary, tertiary, and quaternary carbon atoms respectively, thus showing no change for the primary or secondary carbon atoms but showing some decrease in the contributions associated with the tertiary and quaternary carbon atoms. Recall that an increase in the "weight", because of the inverse square root contributions, means a decrease to the bond additivity that contributes to the connectivity index.

Table 3. The connectivity index and variable connectivity indices for alcohols considered

	$^1\chi$	$^1\chi^f(x, y)$	$^1\chi^f$ (x_1,x_2,x_3,x_4,y)
Carbon atoms (x):	**0**	**0.80**	**0.80, 0.80, 0.96, 1.00,**
Oxygen atom (y):	**0**	**- 0.90**	**- 0.90**
Methanol	1.0000	2.3570	2.3570
Ethanol	1.4142	2.3353	2.3353
1-propanol	1.91421	2.6924	2.6924
2-propanol	1.73205	2.3869	2.3382
1-butanol	2.41421	3.0495	3.0495
2-butanol	2.27006	2.7566	2.7094
2-M-1-propanol	2.27006	2.9611	2.9393
2-M-2-propanol	2.00000	2.4640	2.4142
1-pentanol	2.91421	3.4067	3.4067
2-pentanol	2.8998	3.1137	3.0666
3-pentanol	2.8081	3.1262	3.0806
2-M-1-butanol	2.8081	3.3308	3.3104
3-M-1-butanol	2.7701	2.8420	2.7936
2-M-2-butanol	2.5607	3.3183	3.2964
3-M-2-butanol	2.7176	3.0759	3.0131
2,2-MM-1-propanol	2.5607	3.1832	3.1571
1-hexanol	3.4142	3.7638	3.7638
2-hexanol	3.2701	3.4709	3.4237
3-hexanol	3.3081	3.4834	3.4377
2-M-1-pentanol	3.3081	3.6879	3.6676
3-M-1-pentanol	3.3081	3.6879	3.6676
4-M-1-pentanol	3.2701	3.6754	3.6535
2-M-2-pentanol	3.0607	3.1991	3.1507
3-M-2-pentanol	3.1807	3.4021	3.3365
4-M-2-pentanol	3.1259	3.3824	3.3134
2-M-3-pentanol	3.1807	3.4021	3.3428
3-M-2-pentanol	3.1213	3.2200	3.1729
2-ethyl-1-butanol	3.3461	3.7004	3.6816
2,2-MM-1-butanol	3.1213	3.5612	3.5365
2,3-MM-1-butanol	3.1807	3.6066	3.5663
3,3-MM-1-butanol	3.0607	3.5404	3.5142
2,3-MM-2-butanol	2.9880	3.1517	3.0825
3,3-MM-2-butanol	2.9434	3.2593	3.1884
1-heptanol	3.9142	4.1210	4.1210
3-heptanol	3.8081	3.8405	3.7949

Table 3. Cont.

	$^{1}\chi$	$^{1}\chi^{f}(x, y)$	$^{1}\chi^{f}\ (x_1,x_2,x_3,x_4,y)$
4-heptanol	3.8081	3.8405	3.7949
2-M-2-hexanol	3.5607	3.5563	3.5079
3-M-3-hexanol	3.6213	3.5771	3.5301
3-E-3-hexanol	3.6820	3.5980	3.5523
2,3-MM-2-pentanol	3.4814	3.4923	3.4259
3,3-MM-2-pentanol	3.5040	3.6373	3.5678
2,2-MM-3-pentanol	3.4814	3.6290	3.5596
2,3-MM-3-pentanol	3.5040	3.5007	3.4341
2,4-MM-3-pentanol	3.2201	3.4148	3.3711
1-octanol	4.4142	4.4781	4.4781
2-octanol	4.2701	4.1852	4.1380
2-E-1-hexanol	4.3461	4.4147	4.3959
2,3,3-MMM-3-pentanol	3.8107	3.7307	3.6686
1-nonanol	4.9142	4.8353	4.8353
2-nonanol	4.7701	4.5423	4.5360
3-nonanol	4.8081	4.5548	4.5423
4-nonanol	4.8081	4.5548	4.5423
5-nonanol	4.8081	4.5548	4.5423
7-M-1-octanol	4.7701	4.7468	4.7250
2,6-MM-4-heptanol	4.5197	4.3779	4.3217
3,5-MM-4-heptanol	4.6294	4.4173	4.4048
3,5,5-MMM-1-hexanol	4.4545	4.5359	4.5097
1-decanol	5.4142	5.1924	5.1924

In Table 3 we have listed the connectivity index and two variable connectivity indices for each of the 58 alcohols in order to illustrate the inherent flexibility of variable molecular indices. While for all alcohols shown in Table 3, both variable connectivity indices have increased values as compared to χ, the changes are different for different molecules. For instance, 2-butanol, 2-M-1-propanol, 2-pentanol, and 2-M-2-butanol have all the same $\chi = 2.27006$, but the corresponding values for the variable indices are visibly different: 2.75658, 2.96111, 3.11372, and 3.31825, respectively. These values again slightly change when we introduce different weights for the primary, secondary, tertiary, and quaternary carbon atoms, becoming respectively: 2.70941, 2.93925, 3.06655, and 3.29639. A close look at Table 3 shows that when introducing weights for primary, secondary, tertiary, and quaternary carbon atoms, some indices did not change at all, some have decreased in magnitude slightly, while others show somewhat greater decreases in magnitude. Clearly those alcohols that have no tertiary or quaternary carbon atoms have not changed (in this particular application).

From the above results, we conclude that modeling based on differentiation of CH_3, CH_2, CH and C carbon atoms has some merit and makes a significant, though not dramatic improvement on the accompanying regression equation. In Table 4 below we have listed the optimal regression equations.

Table 4. The regression equation based on differentiation between primary, secondary, tertiary and quaternary carbon atoms. In the lower part of the table regression equation was listed when five outliers have been removed from the set of n = 58 alcohols

Variables	Regression Equation	Outliers
0.80 0.80 0.96 1.00 -0.90	$53.964\ \chi - 49.003$ s = 3.9°C $r^2 = 0.988$ r = 0.994	0
0.80 0.80 0.96 1.00 -0.90	$53.255\ \chi - 45.599$ s = 2.6°C $r^2 = 0.994$ r = 0.997	5

In addition to the regression equation based on all 58 compounds, we also included the regression equation in which we removed five compounds that show deviations larger than two standard deviations. With removal of the outliers, the standard deviation has dropped down to a very remarkable 2.6°C. In Table 5 we have listed the experimental normal boiling points as well as computed normal boiling points and the residuals. The asterisk in the last two columns indicates that the compounds were viewed as outliers, including methanol, with a computed normal boiling point 13.5 °C from the experimental value. The other outliers show differences of between 10.5°C and 6.2°C. Observe that all five of the excessive residuals are negative, meaning that all five computed normal boiling points are larger than the corresponding experimental values. It would be premature to speculate on the origin of this observation, and in particular to speculate on the possibility of impure in experimental samples, because it is difficult to imagine that the normal boiling point of methanol would not be reasonably accurate. In the case of methanol it is likely that the simplified representation of this alcohol by hydrogen suppressed graphs, which reduces molecular graph to simple single edge K_2 graph, may represent an oversimplified scheme. However the other four outliers: 2,2-dimethyl-1-propanol, 3,3-dimethyl-2-pentanol, 2,2-dimethyl-3-petnanol, and 2,6-dimethyl-4-heptanol have no special structural features not present in other alcohols that could suggest their different behavior. Thus it remains an open question whether their departure represents limitations of the model, being the "tail" compounds in a Gaussian distributions of the residuals, or whether there may be some unspecified causes for their departure from the computed regression line, not excluding a possibility of impurities and other experimental errors. In any case it would be of interest to explore if other molecular models of similar accuracy would also point to the same subset of compounds as outliers or not. If that will be the case and if experimental errors are eliminated as cause of disagreement (by repeating the normal boiling point measurements) these compounds would present an interesting challenge for theoretical studies that would account for the anomaly in their normal boiling points.

Table 5. The experimental and calculated normal boiling points for alcohols. The last two columns show results when five outliers have been removed from the set of n = 58 alcohols.

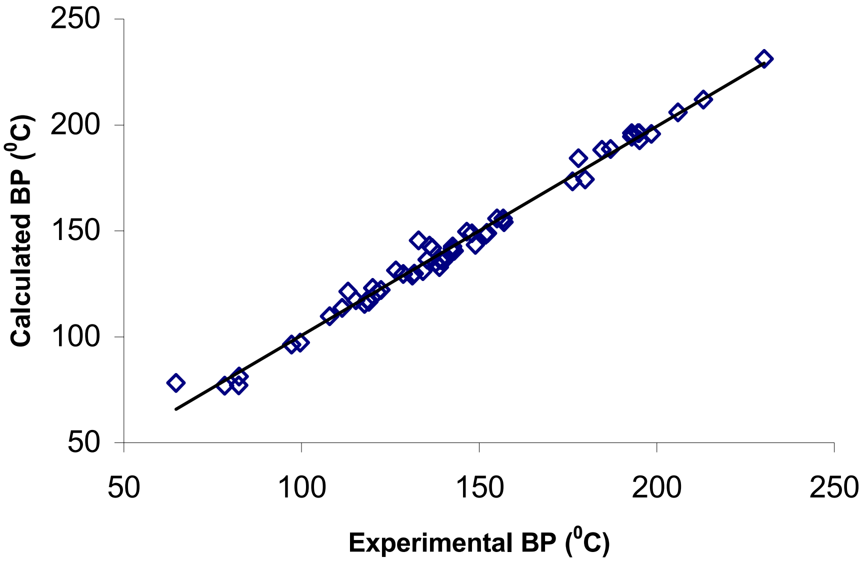
Alcohol	Exp	Calc	Res	Calc*	Res*
methanol	64.7	78.2	-13.5	*	*
ethanol	78.3	77.0	1.3	78.8	-0.5
1-propanol	97.2	96.3	0.9	97.8	-0.6
2-propanol	82.3	77.2	5.1	78.9	3.4
1-butanol	117.0	115.6	2.1	116.8	0.9
2-butanol	99.6	97.2	2.4	98.7	0.9
2-M-1-propanol	107.9	109.6	-1.7	110.9	-3.0
2-M-2-propanol	82.4	81.3	1.1	83.0	-0.6
1-pentanol	137.8	134.9	3.0	135.8	2.0
2-pentanol	119.0	116.5	2.5	117.7	1.3
3-pentanol	115.3	117.2	-1.9	118.5	-3.2
2-M-1-butanol	128.7	129.6	-0.9	130.7	-2.0
3-M-1-butanol	131.2	128.9	2.3	129.9	1.3
2-M-2-butanol	102.0	101.8	0.2	103.2	-1.1
3-M-2-butanol	111.5	113.6	-2.1	114.9	-3.4
2,2-MM-1-propanol	113.1	121.4	-8.3	*	*
1-hexanol	157.0	154.1	2.9	154.8	2.2
2-hexanol	139.9	135.8	4.2	136.7	3.2
3-hexanol	135.4	136.5	-1.1	137.5	-2.1
2-M-1-pentanol	148.0	148.9	-0.9	149.7	-1.7
3-M-1-pentanol	152.4	148.9	3.5	149.7	2.7
4-M-1-pentanol	151.8	148.2	3.6	149.0	2.8
2-M-2-pentanol	121.4	121.0	0.4	122.2	-0.8
3-M-2-pentanol	134.2	131.1	3.2	132.1	2.1
4-M-2-pentanol	131.7	129.8	1.9	130.9	0.9
2-M-3-pentanol	126.6	131.4	-4.8	132.4	-5.8
3-M-2-pentanol	122.4	122.2	0.2	123.4	-1.0
2-ethyl-1-butanol	146.5	149.7	-3.2	150.5	-4.0
2,2-MM-1-butanol	136.8	141.8	-5.0	142.7	-5.9
2,3-MM-1-butanol	149.0	143.5	5.6	144.3	4.7
3,3-MM-1-butanol	143.0	140.6	2.4	141.6	1.5
2,3-MM-2-butanol	118.6	117.3	1.3	118.6	0.0
3,3-MM-2-butanol	120.0	123.1	-3.1	124.2	-4.2
1-heptanol	176.3	173.4	2.9	173.9	2.4
3-heptanol	156.8	155.8	1.0	156.5	0.3
4-heptanol	155.0	155.8	-0.8	156.5	-1.5
2-M-2-hexanol	142.5	140.3	2.2	141.2	1.3
3-M-3-hexanol	142.4	141.5	0.9	142.4	0.0

Table 5. Cont.

Alcohol	Exp	Calc	Res	Calc*	Res*
3-E-3-hexanol	142.5	142.7	-0.2	143.6	-1.1
2,3-MM-2-pentanol	139.7	135.9	3.8	136.9	2.9
3,3-MM-2-pentanol	133.0	143.5	-10.5	*	*
2,2-MM-3-pentanol	136.0	143.1	-7.1	*	*
2,3-MM-3-pentanol	139.0	136.3	2.7	137.3	1.7
2,4-MM-3-pentanol	138.8	132.9	5.9	133.9	4.9
1-octanol	195.2	192.7	2.5	192.9	2.3
2-octanol	179.8	174.3	5.5	174.8	5.0
2-E-1-hexanol	184.6	188.2	-3.6	188.5	-3.9
2,3,3-MMM-3-pentanol	152.2	149.0	3.2	149.8	2.4
1-nonanol	213.1	211.9	1.2	211.9	1.2
2-nonanol	198.5	195.8	2.7	196.0	2.5
3-nonanol	194.7	196.1	-1.4	196.3	-1.6
4-nonanol	193.0	196.1	-3.1	196.3	-3.3
5-nonanol	195.1	196.1	-1.0	196.3	-1.2
7-M-1-octanol	206.0	206.0	0.0	206.0	-0.0
2,6-MM-4-heptanol	178.0	184.2	-6.2	*	*
3,5-MM-4-heptanol	187.0	188.7	-1.7	189.0	-2.0
3,5,5-MMM-1-hexanol	193.0	194.4	-1.4	194.6	-1.6
1-decanol	230.2	231.2	-1.0	230.9	-0.7

In Fig. 1 we have plotted the calculated normal boiling points against the experimental normal boiling points.

Figure 1. Calculated normal boiling points plotted against the experimental normal boiling points for 58 alcohols studies



Concluding Remarks

One should not be surprised that different models may give regressions of similar statistical quality. The selection of descriptors depends also on the models considered and interpretability of the descriptors. In the case of the variable connectivity index we can conclude not only that the carbon-oxygen bond makes a visibly greater contribution to the bond additivity of the normal boiling points of alcohols than do carbon-carbon bonds, but in view of the high quality correlation, we could identify several outliers. By excluding these outliers, the standard error for the regression of well over 50 alcohols is about 3.6°C, and even 2.6°C when five outliers have been removed from the set. It may be premature to speculate on why outliers have occurred, except perhaps for methane, the hydrogen suppressed molecular graph of which may have introduced oversimplification, because it is difficult to think that experimental error of over 10°C would be possible. However, we observe that for all five outliers the experimental normal boiling points are lower than those calculated, which may hint to a systematic rather than random displacement of computed normal boiling points, and consequently this might imply some structural factor not taken into account by variable descriptors that may be behind the difference between the computed and experimental normal boiling points. A closer look at the outliers shows that in three cases the OH group is next to a neopentyl fragment, thus being somewhat shielded. As a consequence in these compounds one expects weaker intermolecular hydrogen bonds that would be associated with lowering of the normal boiling points. In the fourth outliner we have two methyl groups on each side of OH that could produce a similar effect of shielding. Thus observed larger negative residuals may suggest presence of weaker hydrogen bonds and possible influence of steric hindrance due to some crowding of hydrogens from neighboring methyl or methylene groups.

If this explanation proves to hold we could even speculate on possible use of residuals in similar situations as a measure of the "weakening" of hydrogen bonds. We should mention that we exclude methanol from this discussion, as in the case of methanol no shielding could be present, but methanol clearly makes a class on its own and presents challenge on its own.

Acknowledgments

We thank Professor A. T. Balaban (Texas A & M University at Galveston) for useful comments on possible role of hydrogen bond on the normal boiling points of alcohols and numerous other comments on the manuscript that contributed to an improved presentation of the material. We would also like to acknowledge the financial support by the Ministry of Education, Science and Sport of the Republic of Slovenia (Grants P1-0017 and P1-0153). This is contribution number 371 from the Center for Water and the Environment of the Natural Resources Research Institute. Research was supported by Grant F49620-02-1-0138 from the United States Air Force, Cooperative Agreement Number 572112 from the Agency for Toxic Substances and Disease Registry, and Cooperative Agreement Number 58-1275-4-313 from the United States Department of Agriculture Research Services.

References

1. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors* (Methods and Principles in Medicinal Chemistry, vol. 11; Mannhold, R.; Kubinyi, H.; Timmerman, H., Eds.; Wiley-VCH: New York, **2000**
2. Devillers, J; Balaban, A. T., Eds. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: Amsterdam, **1999**
3. Lučić, B.; Miličević, A.; Nikolić, S.; Trinajstić, N. On variable Wiener index. *Ind. J. Chem. A* **2003**, *42*, 1279-1282
4. Estrada, E. Three-dimensional generalized graph matrix, Harary descriptors and a generalized interatomic Lennard-Jones potential. *J. Phys. Chem. A* **2004**, *108*, 5468-5473
5. Estrada, E. Generalized Graph Matrix, Graph Geometry, Quantum Chemistry and the Optimal Description of Physicochemical Properties. *J. Phys. Chem. A* **2003**, *107*, 7482-7489
6. Estrada, E.; Gutierrez, Y. The Balaban *J* index in the multidimensional space of generalized topological indices. Generalizations and QSPR *Match* **2001**, *44*, 155-167
7. Estrada, E. Generalization of topological indices. *Chem. Phys. Lett.* **2001***,* 336, 248-252
8. Kier, L. B.; Hall, H. L. Molecular Connectivity. VII. Specific treatment of heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806-1809
9. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, **1976**
10. Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc*. **1975**, 97, 6609-6615
11. Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular connectivity V: Connectivity series concept applied to density. *J. Pharm. Sci*. **1976**, 65, 1226-1230
12. Barysz, M.; Jashari, G.; Lall, R. S.; Srivastave, V. K.; Trinajstić, N. On the distance matrix of molecules containing heteroatoms. In *Applications of Chemical Topology and Graph Theory*. King., R. B. Ed.; Elsevier: Amsterdam, **1983**; pp. 222-227
13. Balaban, A. T. Chemical graphs. 48. Topological index J for heteroatom-containing molecules taking into account periodicities of elements properties. *Math. Chem. (MATCH)* 1986, 21, 115-122
14. Ivanciuc, O; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, T. Comparison of weighted schemes for molecular graph descriptors: Application in quantitative structure – retention relationship for alkylphenols in gas-liquid chromatography. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 732-743
15. Ivanciuc, O; Ivanciuc, T.; Balaban, T. Design of topological indices. Part 10. Parameters based on electronegativity and covalent radius for the computation of molecular graph descriptors for heteroatom-containing molecules. *J. Chem. Inf. Comput. Sci*. **1998**, 38, 395-401
16. Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett*. **1982**, 80, 399-404
17. Antipin, I. S., Arslanov, N. A., Palyulin, V. A., Konovalov, A. I., Zefirov, N. S. Solvation topological index. Topological description of dispersion interaction (in Russian). *Dokl. Akad. Nauk. SSSR*, **1991**, 316, 925-927
18. Antipin, I. S., Arslanov, N. A., Palyulin, V. A., Konovalov, A. I., Zefirov, N. S. Prognosis of enthalpy[y of nonspecific salvation of organic nonelectrolytes (in Russian). *Dokl. Akad. Nauk. SSSR*, **1993**, 316, 173-176; [*Chem. Abstr*. **1993**, *120*, 133743]

19. Zefirov, N. S.; Palyulin, V. A. QSAR for boiling points of “small” sufides. Are the “High-Quality Structure-Property-Activity Regressions” the real high quality QSAR models?” *J. Chem. Inf. Comput. Sci*. **2001**, 41, 1022-1027
20. Katrizky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA (COmprehensive Descriptors for Structural and Statistical Analysis), University of Florida, Gainesville. FL. **1995**
21. Randić, M. Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci*. **1991**, 31, 311-320
22. Randić, M. Orthogonal molecular descriptors. *New J. Chem*. **1991**, 15, 517-525
23. Randić, M. Correlation of enthalpy of octanes with orthogonal connectivity indices. *J. Mol. Struct. (Theochem)* **1991**, 233, 45-59
24. Randić, M. Fitting of nonlinear regressions by orthogonalized power series. *J. Comput. Chem*. **1992**, 14, 363-370
25. Wu, L.; Zhang, W-J. Comparison of different methods for variable selection. *Anal. Chim. Acta* **2001**, 446, 477-483
26. Lučić, B.; Amić, D.; Trinajstić, N. Nonlinear multivariate regressions outperforms several concisely designed neural networks on three QSRP data set. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 403-413
27. Randić, M. Retro-regression – another important multivariate regression improvement. *J. Chem. Inf. Comput. Sci*. **2001**, 41, 602-606
28. Randić, M.; Zupan, J. M. On the structural interpretation of topological indices. In: "*Topology in Chemistry. Discrete Mathematics of Molecules*"; Rouvray, D. H.; King, R. B., Eds.; Horwood Publishing Series in Chemical Science, Horwood Publ. Ltd.: Chichester, U.K. **2002**; pp. 249-291.
29. Randić, M.; Zupan, J. On intereptation of well-known topological indices. *J. Chem. Inf. Comput. Sci*. **2001**, 41, 550-560
30. Randić, M.; Balaban, A. T.; Basak, S. C. On structural interpretation of several distance related topological indices. *J. Chem. Inf. Comput. Sci*. **2001**, 41, 593-601
31. Randić, M.; Basak, N. Novel graphical matrix and novel distance based molecular descriptors. *Croat. Chem. Acta* (in press)
32. Randić, M. Novel graph theoretical approach to heteroatoms in quantitative structure-activity relationships. *Chemometrics Intel. Lab. Systems* **1991**, 10, 213-227
33. Randić, M. On computation of optimal parameters for multivariate analysis of structure-property relationship. *J. Comput. Chem*. **1991**, 12, 970-980
34. Randić, M.; Dobrowolski, J. Cz. Optimal molecular connectivity descriptors for nitrogen-containing molecules. *Int. J. Quantum Chem*. **1998**, 70, 1209-1215
35. Randić, M. High quality structure-property regressions: boiling points of smaller alkanes. *New J. Chem*. **2000**, 24, 165-171
36. Randić, M.; Basak. S. C. Construction of high quality structure-property-activity regressions: the boiling points of sulfides. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 899-905
37. Randić, M.; Pompe, M. The variable connectivity index ${}^{1}\chi^{f}$ versus the traditional descriptors: A comparative study of 1chif against descriptors of CODESSA. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 631-638.
38. Randić, M.; Mills, D.; Basak, S. C. On use of variable connectivity index ${}^{1}\chi^{f}$ in QSAR: Toxicity of aliphatic ethers. *J. Chem. Inf. Comput. Sci*. **2001**, 41, 614-618

39. Randić, M.; Pompe, M. The variable connectivity index $^1\chi^f$ versus the traditional molecular descriptors: A comparative study of $^1\chi^f$ against descriptors of CODESSA. *J. Chem. Inf. Comput. Sci*. **2001**, 41, 631-638
40. Randić, M.; Plavšić, D.; Lerš, N. Variable connectivity index for cycle-containing structures. . *J. Chem. Inf. Comput. Sci*. **2001**, 41, 657-662
41. Randić, M.; Mills, D.; Basak, S. C. On use of variable connectivity index for characterization of amino acids. *Int. J. Quantum Chem*. *Int. J. Quantum Chem.* **2000**, *80,* 1199-1209
42. Liu, D.; Zhong, C. Modeling of the heat capacity of polymers with the variable connectivity index. *Polymer J*. **2002**, 34, 954-961
43. Zhong, C.; He, J.; Xia, Z; Li, Y. Estimation of activity for Efavirenz analogous with the K 103N mutant of HIV reverse transcriptase using variable connectivity indices. *Bioorg. Med. Chem. Lett.* (submitted)
44. Randić, M.; Pompe, M. On characterization of the CC double bond in alkenes. *SAR & QSAR in Environ. Res*. **1999**, 10, 451-471
45. Randić, M.; Basak. S. C. Optimal molecular descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci*. **1999**, 39, 261-266
46. Randić, M.; Basak. S. C. A new descriptor for structure-property and structure-activity correlations. *J. Chem. Inf. Comput. Sci*. **2001**, 41, 650-656
47. Randić, M.; Pompe, M. The variable molecular descriptors based on distance related matrices, *J. Chem. Inf. Comput. Sci*. **2001**, 41, 575-581
48. Krenkel, G.; Castro, E. A.; Toropov, A. A. Improved molecular descriptors based on the optimization of correlation weights of local graph invariants. *J. Mol. Struct. - Theochem*. **2001**, 542, 107-113
49. Randić, M.; Pompe, M. (to be published)
50. Cammarata, A. Molecular topology and aqueous solubility of aliphatic alcohols, *J. Pharm. Sci.* **1979**, 68, 839-842

© 2004 by MDPI (http://www.mdpi.org). Reproduction is permitted for noncommercial purposes.