The Variable Connectivity Index $^1\chi^f$ versus the Traditional Molecular Descriptors: A Comparative Study of $^1\chi^f$ Against Descriptors of CODESSA

Milan Randić*,† and Matevž Pompe‡

National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia, and Department of Chemistry and
Chemical Technology, University of Ljubljana, Ljubljana, Aškerčeva 5, Slovenia

Received August 12, 2000

In this study we compared the prediction abilities of the variable connectivity index $^1\chi^f$ (not included in CODESSA) with topological indices available from CODESSA. We selected the boiling points of $n = 100$ alcohols as the property and examined the pool of 56 topological indices. Prediction capabilities of the developed models were evaluated by clasical training/test set approach. RMS errors calculated from the prediction set for the MLR models obtained from CODESSA software with 1, 2, 3, 4, and 5 parameters were 9.06, 5.69, 5.40, 4.9, and 3.37 °C, respectively. Using the variable connectivity index with weights $x = 0.10$ and $y = -0.92$ for carbon and oxygen atom respectively, we obtain regression BP $= 38.12\ ^1\chi^f - 37.56$ with the correlation coefficient $r = 0.9915$, RMS error 4.21 °C calculated from the test set, and Fisher ratio $F = 5691$. Prediction capability of the variable connectivity index was better than for MLR regression model with up to four parameters.

1. INTRODUCTION

Topological indices[1−4] have emerged as molecular descriptors of choice in studies of structure−property activity and rational drug design. They are in particular inescapable in the development of successful multilinear regression analysis (MRA) and other related statistical methodologies (e.g., the principal component analysis,[5] the pattern recognition,[6] the cluster analysis,[7] artificial neural networks[8]) as well as in screening combinatorial libraries. Recent quantitative structure−property relationship (QSPR) and quantitative structure−activity relationship (QSAR) studies have been facilitated greatly by development of powerful computer software, such as POLLY,[9] MOLCONN,[10] and particularly CODESSA,[11] which allow fast computation of large number of various molecular descriptors and subsequent statistical analysis of such by multivariate regression analysis (MRA) or other statistical methodologies. These programs evaluate hundreds of topological indices and in the case of CODESSA also other geometrical and quantum-chemical molecular descriptors.

CODESSA includes two advanced procedures for systematic development of the multilinear QSAR/QSPR equations (1) *The Heuristic Method* and (2) *The Best Multilinear Regression Method.* The Heuristic Method is usually used to obtain preliminary screening of the library of descriptors in order to select a subset of descriptors that may be of interest and importance for the study under consideration. It produces several alternative best regression models for a single-parameter correlation and points to descriptors that are highly interrelated. The program then eliminates highly correlated descriptors using pairwise correlation matrix of descriptors and thus further reduces the descriptor pool. This last step, however, lacks theoretical justification. To illustrate this point consider regression of molar refraction (MR) for octanes with $^1\chi$ and $^2\chi$, the first-order[12] and the second-order[13] connectivity indices

$$\mathrm{MR} = 4.6951\ ^1\chi + 1.3720\ ^2\chi + 17.5482$$

with the correlation coefficient $r = 0.971$.[14] A simple regression of MR for the same data with $^1\chi$ and $^2\chi$ shows no correlation, $r_1 = 0.087$ and $r_2 = 0.177$, for $^1\chi$ and $^2\chi$, respectively. If we would adopt one-parameter significance criterion $r > 0.200$, we would have missed the above quite good regression equation of MR with both $^1\chi$ and $^2\chi$. At the same time we would also have miss the two-parameter multiple linear regression (MLR) model because no one would normally use both parameters with very high intercorrelation ($r = 0.9630$) in the same equation.

The Heuristic Methods of CODESSA report all the best two-parameter regressions setting a stage for further stepwise introduction of additional descriptors. Once the type of descriptors has been selected one can use *The Best Multilinear Regression Method* option of CODESSA for a more systematic and thorough search of the selected pool of descriptors (which may include up to 300 descriptors and 150 structures). In this way one can obtain the best two-parameter model, the best three-parameter model, etc., based on the highest r^2 value (r is the correlation coefficient). In this study we will make a comparison of the results that are obtained using CODESSA with the result obtained by using a simple regression model based on the variable connectivity index $^1\chi^f$.

2. VARIABLE CONNECTIVITY INDEX $^1X^F$

The connectivity index χ, also referred to as the first-order connectivity index $^1\chi$ or the Randić index (order 1), is a bond

* Corresponding author fax: (515)292-8629; e-mail: milan.randic@drake.edu. Current address: 3225 Kingman Rd, Ames, IA 50015. On leave from Department of Mathematics & Computer Science, Drake University, Des Moines, IA 50311.

† National Institute of Chemistry.

‡ University of Ljubljana.

10.1021/ci000119r CCC: $20.00
Published on Web 04/12/2001

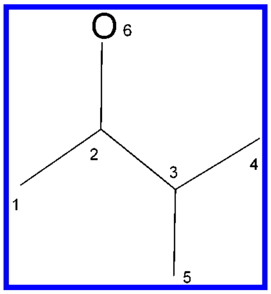

Figure 1. Hydrogen supressed molecular graph for 3-methyl-2-butanol.

Table 1. The Adjacency Matrix, the Augmented Adjacency Matrix, the Row Sums, the Connectivity Index, and the Variable Connectivity Index for 3-Methyl-2-butanol

	1	2	3	4	5	6	row sum
$^1\chi = 4/\sqrt{3} + 1/3 = 2.642734$							
1	0	1	0	0	0	0	1
2	1	0	1	0	0	1	3
3	0	1	0	1	1	0	3
4	0	0	1	0	0	0	1
5	0	0	1	0	0	0	1
6	0	1	0	0	0	0	1
$^1\chi^v = 3/\sqrt{(1+x)(3+x)} + 1/(3+x) + 1/\sqrt{(3+x)(1+y)} = \mathrm{f}(x, y)$							
1	x	1	0	0	0	0	$1 + x$
2	1	x	1	0	0	1	$3 + x$
3	0	1	x	1	1	0	$3 + x$
4	0	0	1	x	0	0	$1 + x$
5	0	0	1	0	x	0	$1 + x$
6	0	1	0	0	0	y	$1 + y$

additive molecular structural invariant. It can be computed by first classifying the bond into the (m, n) bond type, where m, n are the (graph theoretical) valences of atoms forming the bond. The connectivity index is given by summing the contributions of the form $1/\sqrt{(m\cdot n)}$ over all the bonds of hydrogen suppressed molecular graph. The example of a hydrogen suppressed molecular graph for 3-methyl-2-butanol is shown in Figure 1. Alternatively, $^1\chi$ can be obtained from the adjacency matrix of a molecular graph by using the row sums and the algorithm $1/\sqrt{(RS_i\cdot RS_j)}$ and summing over all bonds (i, j). RS_i and RS_j are row sums for ith and jth rows, respectively. This is illustrated in Table 1 (top part) on 3-methyl-2-butanol.

The variable connectivity index $^1\chi^f$ is a generalized connectivity index constructed from augmented adjacency matrix.[15,16] To obtain $^1\chi^f$ one replaces the diagonal zero entries of the adjacency matrix by variables *x*, *y*, *z*, ... Each variable characterizes different kinds of atoms in a molecule or even the same atom in structurally distinct environment. Clearly $^1\chi^f$ remains a bond additive quantity; however, the contributions of individual bonds are now modified by the presence of the variables *x*, *y*, *z*, ... It can be computed in an analogous manner as $^1\chi$ from the row sums of the modified adjacency matrix. In Table 1 (the lower part) this is illustrated again on 3-methyl-2-butanol where the variables *x* and *y* indicate carbon and oxygen atoms, respectively. Instead of 1 and 3 as the row sums of the adjacency matrix now we have as the row sums $(1 + x)$, $(3 + x)$, and $(1 + y)$. The generalized connectivity index becomes a function $\mathrm{f}(x, y)$ of two variables. It can be calculated only after one selects numerical values for *x* and *y*. The case $x = 0$, $y = 0$ reduces the variable connectivity index $^1\chi^f$ to $^1\chi$.

Table 2. Comparison of the Standard Error When the Connectivity Index Is Used and When the Variable Connectivity Index Is Used as Found in Recent High Quality Regression Studies

compounds	property	$^1\chi$ s	$^1\chi^v$ s	ref
alcohols	boiling points	7.86	3.30	17
amines	boiling points	3.488	1.907	20
smaller alkanes	boiling points	2.928	2.481	21
alkanes + cycloalkanes	boiling points	4.15	3.18	22
sulfides	boiling points	2.71	1.326	23
alcohols	toxicity in mice	0.1297	0.0957	24
alkane + alcohols	retention indices	56.44	14.24	25
amino acids	partial volume	11.118	5.861	26
amino acids	crystal density	0.184	0.040	27
amino acids	partition coefficient	0.565	0.269	27
amino acids	relaxation rate	0.125	0.044	27

As has been demonstrated by varying the variables *x*, *y*, *z*, ... so that the standard error of a regression is minimized one can substantially improve the correlation coefficient of the regression *r*, the standard error *s*, and the Fisher ratio *F*, of numerous regressions. We should emphasize that the variable connectivity index is a representative of a novel class of topological indices. In contrast to the traditional topological indices, such as Wiener index W, Hosoya index Z, Balaban's index J, and others, which are all numerically "fixed" once a structure has been selected, the variable descriptors, such as the variable connectivity index, can take different values during the search for optimal regression. Its value will depend on the selection of the variable weights *x*, *y*, *z*, ... Therefore the connectivity index can adjust to individual requirements that different molecules and different molecular properties may require. That different molecular properties may require distinct descriptors for the same set of compounds has been confirmed with the use of the variable path numbers on the boiling points, the octanol/water partition coefficient log P, the cavity surface area, and the molecular solubilities of alcohols.[17,18]

In Table 2 we report recent results on the high quality regressions obtained using $^1\chi^f$.[16,19−26] For comparison we also show the standard errors for the same data when the simple connectivity index $^1\chi$ is used instead of $^1\chi^f$. Observe the visible improvement of the regression parameters when the variable index is used instead of the traditional topological indices. Typically the standard error has been reduced by a factor from two to four. In view of such an exceptional performance of the variable connectivity index it would be of interest to see how it compares with the performance of the powerful CODESSA software. To find out we selected the boiling points of 100 alcohols for which we sought the best CODESSA results.

3. THE BEST MULTILINEAR REGRESSION RESULTS BY CODESSA

In Table 3 we list 56 molecular descriptors that we selected from CODESSA to be tested against the variable connectivity index $^1\chi^f$. The indices can be grouped into the indicator variables, the topological indices, and the information theoretic indices. The indicator variables count various kinds of bonds and atoms, and the topological indices include the connectivity indices of different order, the valence connectivity indices of different order, and Kier's kappa shape indices. The kappa indices are based on a comparison of the path numbers in a molecule and in the extreme structures

Table 3. The Indicator Variables, the Topological Indices, and the Information Theoretical Indices from CODESSA Tested in This Work

indicator variables	topological indices	information theoretic indices
number of atoms	Wiener index	av information content (order 0)
number of C atoms	Randić index (order 0)	information content (order 0)
relative number of C atoms	Randić index (order 1)	av structural information content (order 0)
number of H atoms	Randić index (order 2)	structural information content (order 0)
relative number of H atoms	Randić index (order 3)	av complementary information content (order 0)
number of O atoms	Kier & Hall index (order 0)	complementary information content (order 0)
relative number of O atoms	Kier & Hall index (order 1)	av bonding information content (order 0)
number of bonds	Kier & Hall index (order 2)	bonding information content (order 0)
number of single bonds	Kier & Hall index (order 3)	av information content (order 1)
relative number of single bonds	Kier shape index (order 1)	information content (order 1)
number of double bonds	Kier shape index (order 2)	av structural information content (order 1)
relative number of double bonds	Kier shape index (order 3)	structural information content (order 1)
molecular weight	Kier flexibility index	av complementary information content (order 1)
relative molecular weight	Balaban index	complementary information content (order 1)
gravitation index (all bonds)		av bonding information content (order 1)
gravitation index (all pairs)		bonding information content (order 1)
		av information content (order 2)
		information content (order 2)
		av structural information content (order 2)
		structural information content (order 2)
		av complementary information content (order 2)
		complementary information content (order 2)
		av bonding information content (order 2)
		bonding information content (order 2)

(the linear chains and the star graphs having the same number of vertices). The information theoretic indices use the Shannon formula of Information Theory[27] applied to various partitioning of molecules into the constituent elements. CODESSA searches for the best single, the best two-descriptor combination, the best three-descriptor combination, etc. The data set of 100 alcohols (Table 4) was randomly divided into two subsets. The first one contained 70 alchohols and was used for the creation of the models. The remaining 30 compounds formed a test set and were used for the evaluation of the prediction capabilities of the created models. In Table 4 the compounds representing the test set are marked with the italic letter b. In Table 5 we show the regression results obtained for the $n = 70$ boiling points of alcohols. The Fisher ratio (F) and the regression coefficient (r) were calculated using a training set. On the other hand the root-mean squared (RMS) error was calculated from the test set. The best single descriptor selected by CODESSA is the simple connectivity index ${}^1\chi$ (labeled in CODESSA as Randić index of order 1). The accompanied RMS error is 9.48 °C, however, still too large. Already using two descriptors the RMS error shows considerable improvement to 5.69 °C. With use of three descriptors the RMS error is further reduced to 5.40 °C and finally reduced again with the use of four descriptors: RMS = 4.91 °C. Finally, the use of five descriptors in multilinear regression gives a respectable MRS error of 3.37 °C.

Observe that the reduction in the standard error with addition of new descriptors in the stepwise regressions is not paralleled by a stepwise increase of the Fisher ratio F. One would like to see in a stepwise regression at each step a decrease in the standard error and an increase in the Fisher ratio. As we see this is not the case with data in Table 5 as in the third and the fourth step the value of F has decreased. Even more important is to observe that the exhaustive search by CODESSA for the best descriptors select *different* descriptors at different stages of the stepwise regression. Thus the Randić index (order 1), which is the best single descriptor, does not appear in the best two-parameter regression. The best two descriptors are the information content (order 1) and Kier flexibility index. In the next step Kier and Hall index (order 3) is added, but in the following step all the descriptors that make the best three-descriptor combination are displaced by the connectivity indices and the valence connectivity indices of Randić and Kier and Hall. This unpredictive behavior of descriptors in multilinear regressions during the search for best descriptors has been reported in several studies by Lučić, Trinajstić, and co-workers.[28−32]

The described unpredicted behavior of descriptors in stepwise regression makes orthogonalization (and attempts to interpret regression equations) impossible. Recently an approach that makes it possible to arrive at orthogonalized descriptors in such a situation, referred to as the "retro-regression", was outlined.[33,34] Briefly, one selects the best result obtained in stepwise regression constructed by an exhaustive search of combinatorial possibilities. For example, this could be the last regression equation of Table 5 (if found to be statistically significant). Once the stepwise solution has been selected one considers the structure-space defined by all the descriptors that occur in the so selected regression equation. Next one seeks a representation for the selected structure-space. This can be accomplished by backtracking and eliminating one of the descriptors at each step in a stepwise fashion. In this way one can order the selected set of descriptors that define the structure-space and use them to construct orthogonalized descriptors.

The unpredicted behavior of stepwise regression deserves more attention. While the proposed retroregression will cure this problem, it does not tell why this has happened. Why the information content (order 1) and the Kier flexibility index have replaced the already found best single descriptor? The answer may be not because the information content (order 1) and Kier flexibility index have merits on their own, but because they correlate to large extent with, the already found best descriptor, the Randić index (order 1). A way to reduce, if not to eliminate completely, the unpredicted behavior of the stepwise regressions using exhaustive searching for the best combination of descriptors could be to use

Table 4. List of the 100 Alcohols Considered with the Experimental and Calculated Boiling Points[a]

		BP (°C)				BP (°C)	
structure	name	exptl	predicted	structure	name	exptl	predicted
1	ethanol	78	80.5	51	2,3,3-trimethyl-2-butanol	131	128.0
2	propanol	97.1	98.7	52	octanol	195.1	189.4
3	2-propanol	82.4	80.3	53	6-methyl-1-heptanol[b]	188.6	184.3
4	butanol	117.6	116.8	54	4-methyl-1-heptanol[b]	188	185.5
5	2-methyl-1-propanol	108.1	111.7	55	2-octanol	180	172.2
6	2-butanol	99.5	99.6	56	2,5-dimethyl-1-hexanol	179.5	180.3
7	2-methyl-2-propanol	82.4	82.8	57	4-octanol[b]	176.3	173.5
8	pentanol	138	135.0	58	6-methyl-3-heptanol[b]	174	168.3
9	3-methyl-1-butanol[b]	131	129.8	59	5-methyl-3-heptanol	172	169.5
10	2-methyl-1-butanol[b]	128	131.0	60	3-octanol[b]	171	173.5
11	2-pentanol	119.3	117.8	61	5-methyl-2-heptanol[b]	170	168.3
12	3-pentanol	116.2	119.0	62	4-methyl-3-heptanol	170	170.1
13	3-methyl-2-butanol	112.9	113.2	63	2,4,4-trimethyl-1-pentanol	168.5	172.8
14	2-methyl-2-butanol	102.3	103.0	64	2-methyl-3-heptanol	167.5	168.9
15	hexanol	157.6	153.1	65	3-methyl-2-heptanol	166.1	168.9
16	3-methyl-1-pentanol[b]	153	149.2	66	3,4-dimethyl-2-hexanol	165.5	165.5
17	4-methyl-1-pentanol[b]	151.9	148.0	67	2-methyl-4-heptanol	164	168.3
18	2-methyl-1-pentanol	149	149.2	68	3-methyl-3-heptanol	163	159.4
19	2-ethyl-1-butanol	147	150.4	69	3-methyl-4-heptanol	162	170.1
20	2,3-dimethyl-1-butanol[b]	144.5	144.6	70	4-methyl-4-heptanol[b]	161	159.4
21	3,3-dimethyl-1-butanol	143	140.4	71	2-methyl-3-ethyl-3-pentanol[b]	160	157.1
22	2-hexanol	140	135.9	72	2,3-dimethyl-2-hexanol	160	154.4
23	2,2-dimethyl-1-butanol	136.5	142.4	73	2,3,4-trimethyl-3-pentanol	156.5	150.9
24	3-hexanol	135	137.2	74	2-methyl-3-ethyl-2-pentanol	156	155.6
25	3-methyl-2-pentanol	134.3	132.6	75	2-methyl-2-heptanol	156	157.4
26	4-methyl-2-pentanol[b]	131.6	130.8	76	2,5-dimethyl-2-hexanol	154.5	152.3
27	2-methyl-3-pentanol	126.5	132.6	77	2,2,4-trimethyl-3-pentanol	150.5	157.1
28	3-methyl-3-pentanol[b]	122.4	123.1	78	2,4,4-trimethyl-2-pentanol	147.5	144.7
29	2-methyl-2-pentanol	121.1	121.1	79	nonanol[b]	213.3	207.6
30	3,3-dimethyl-2-butanol	120.4	124.2	80	7-methyloctanol	206	202.4
31	2,3-dimethyl-2-butanol	118.4	116.9	81	3-nonanol	195	191.6
32	heptanol	176.4	171.3	82	2-nonanol	193.5	190.4
33	4-methyl-1-hexanol	173	167.3	83	5-nonanol[b]	193	191.6
34	5-methyl-1-hexanol	170	166.1	84	4-nonanol	192.5	191.6
35	3-methyl-1-hexanol	169	167.3	85	4-ethyl-4-heptanol	182	179.5
36	2-methyl-1-hexanol	164	167.3	86	2-methyl-2-octanol	178	175.6
37	2-heptanol[b]	159	154.1	87	2,6-dimethyl-3-heptanol	175	181.9
38	2,4-dimethyl-1-pentanol	159	150.2	88	2,6-dimethyl-4-heptanol[b]	174.5	181.3
39	3-heptanol[b]	157	155.3	89	2,6-dimethyl-2-heptanol	173	170.4
40	4-heptanol	156	155.3	90	3,6-dimethyl-3-heptanol[b]	173	172.4
41	5-methyl-2-hexanol	151	148.9	91	2,2,3-trimethyl-3-hexanol	156	166.3
42	5-methyl-3-hexanol[b]	148	150.2	92	decanol[b]	231.1	225.7
43	2-methyl-2-hexanol	143	139.3	93	3,7-dimethyl-1-octanol	212.5	216.6
44	2-methyl-3-hexanol[b]	143	150.7	94	2-decanol[b]	211	208.5
45	3-methyl-3-hexanol	143	141.2	95	4-decanol[b]	210.5	209.8
46	3-ethyl-3-pentanol	142	143.2	96	3,6-dimethyl-3-octanol[b]	202.2	191.8
47	2,3-dimethyl-3-pentanol[b]	139.7	137.0	97	3-ethyl-3-octanol[b]	199	197.7
48	2,4-dimethyl-3-pentanol	138.7	146.1	98	2,6-dimethyl-4-octanol	195	200.7
49	2,2-dimethyl-3-pentanol[b]	135	143.5	99	2,7-dimethyl-3-octanol	193.5	200.0
50	2,4-dimethyl-2-pentanol	133.1	134.1	100	3-ethyl-2-methyl-3-heptanol	193	193.4

[a] M = methyl, E = ethyl, C_n-OH is the alcohol having n carbon Atoms in a chain, i.e., C_2-OH = ethanol, C_3-OH = propanol, C_4-OH = butanol, etc. [b] Test set.

orthogonalized molecular descriptors[35−38] instead of nonorthogonal. Amić and co-workers[39] considered orthogonalized descriptors and found that the order in which descriptors are orthogonalized can influence the statistical parameters of a regression. They, however, have not applied orthogonalization in a stepwise fashion as advocated here. We think that after the first best descriptor has been found one should make all other descriptors orthogonal to it. One then continues an exhaustive search for the best combination of two descriptors that are orthogonal to the already found best single descriptor. It may happen that the already found best single descriptor remains the best in the combinations of two descriptors, but this also need not be the case. If this is not the case, the so found best two descriptors (orthogonal on the best single descriptor) certainly deserve the attribute of the best combination of two descriptors *on their own merits.* Otherwise they may be found as the best two descriptors because they to a greater or a lesser degree parallel properties of the already found best single descriptor. Work is underway to see to what extent the so called "orthogonalized" procedure will reduce the unpredicted behavior of molecular descriptors previously discussed.

4. COMPARISON OF CODESSA RESULTS WITH THE RESULTS BASED ON THE VARIABLE CONNECTIVITY INDEX

In Table 4 we have listed the experimental and the calculated boiling points for the 100 alcohols considered.

Table 5. The Best Stepwise Regression Using from One to Five Descriptors Based on the 56 Molecular Descriptors Listed in Tables 3 and 4

	X	DX	*t*-test	
0	23.484	4.476	5.246	intercept
1	34.92	1.189	29.35	Randić index (order 1)
	$r = 9627$	RMS = 9.06	$F = 861$	
0	24.527	2.937	8.35	intercept
1	2.085	0.080	25.95	information content (order 1)
2	9.497	0.462	20.57	Kier flexibility index
	$r = 0.9864$	RMS = 5.69	$F = 1212$	
0	29.265	3.407	8.59	intercept
1	1.800	0.138	13.05	information content (order 1)
2	9.592	0.446	21.50	Kier flexibility index
3	5.136	2.058	2.50	Kier & Hall index (order 3)
	$r = 0.9876$	RMS = 5.40	$F = 872$	
0	46.320	3.136	14.77	intercept
1	68.228	4.752	14.36	Randić index (order 1)
2	−15.002	3.712	−4.04	Randić index (order 2)
3	30.130	3.696	8.15	Kier & Hall index (order 2)
4	−29.668	4.494	−6.60	Kier & Hall index (order 0)
	$r = 0.9911$	RMS = 4.91	$F = 899$	
0	147.46	12.448	11.85	intercept
1	281.48	25.39	11.09	Randić index (order 1)
2	30.494	6.621	4.61	Randić index (order 2)
3	40.211	3.277	12.27	Kier & Hall index (order 2)
4	−10.853	1.144	−9.48	molecular weight
5	21.625	3.260	6.63	Kier & Hall index (order 3)
	$r = 0.9946$	RMS = 3.37	$F = 1167$	

[a] X indicates the coefficient of the descriptors listed in the last column, DX is the standard error for the coefficient next column gives the *t*-test, *r* is the regression coefficient, *s* = standard error for regression equation, and *F* is Fisher ratio.

The same training/test set procedure as used for the validation of CODESSA models was used for the evaluation of prediction capabilities of the linear regression model using variable connectivity index. The training set was used for the creation of the model, and the test set was used for the calculation of the RMS error in prediction. Using the weights $x = 0.10$ and $y = -0.92$ for the carbon atoms and the oxygen atom, respectively, we obtain the simple regression equation based on $^1\chi^f$

$$BP = 38.12\ {}^1\chi^f - 37.56$$

with R = 0.991 and RMS = 4.21.

When this is compared with the results obtained by CODESSA it clearly stands out. The variable connectivity index $^1\chi^f$ gave visibly better statistics than the best four descriptors derived from CODESSA. The CODESSA results were obtained after examination, in an exhaustive search, of all combinations of four descriptors from the 56 descriptors of Table 3. That means that simple regression based on $^1\chi^f$ is better than any of the 316 251 multivariate regressions using the combinations of four descriptors at a time. It was only after use of five descriptors that CODESSA produced a regression with a smaller standard error ($s = 3.29$ °C) than the variable connectivity index ($s = 4.02$ °C).

There is yet another important advantage of the variable connectivity index over the best combination of four descriptors. Use of a single descriptors $^1\chi^f$ allows relatively simple interpretation of the results of the regression analysis in comparison with interpretation of linear combination of nonorthogonal descriptors. Consider an interpretation of the linear combination BP = 46.609 + 68.013 Randić index (order 1) − 15.654 Randić index (order 2) + 30.698 Kier and Hall index (order 2) − 29.466 Kier & Hall index (order 0).

Despite that all the indices included in the above regression equation are individually well defined and numerically can be easily obtained, it is unclear what is the meaning of their linear combination. The same is true for any linear combination of nonorthogonal descriptors, as is well illustrated by the principal component analysis (PCA). The principal components (PC), being eigenvectors of a correlation matrix, are among themselves orthogonal. However, each individual PC is a linear combination of nonorthogonal descriptors. This makes interpretation of PCs difficult, if not impossible, as has been illustrated on many occasions. For example, Cramer[40] applied the PCA to common physicochemical properties of a selection of molecules and found interrelation suggesting three dominant components. However, when interpreting the results (which are mathematically rigorous and precise) at best he could use such vague and undefined concepts as the "compactness" and the "bulk" as their meaning. Similarly Basak and co-workers[41−44] have in several papers examined the principal components using topological indices for the characterization of molecules. They found that the first principal component strongly correlates with variables that quantify shape and size of molecules. The next important factor is molecular complexity that is encoded in the second principal component, while the higher order principal components are strongly correlated with invariants, which quantify such subtle structural factors as branching, cyclicity, etc. Again the *quantitative* results of PCA can at best be described by *qualitative*, intuitively understood but undefined quantities, such as the size, the shape, the complexity, the branching, etc.

In addition to the difficulty of interpretation of individual linear combination of topological (and other) descriptors the problem is also compounded by lack of interpretation of most of topological indices. The physicochemical interpretation of many topological indices, including the well-known Wiener index,[45] is unclear. The use of qualitative and ambiguous concepts only further complicates the situation. For example, what constitutes a shape index? A shape index should, in our view, be independent of the size of a molecule, just as a shape of an ellipses is independent of its size. However, the Kier's shape indices[46−57] depend on the size of the molecule considered to some extent, as has been recently noticed.[52,53] The problem of interpretation of topological indices has recently been reconsidered.[54]

5. ON INTERPRETATION OF THE REGRESSION BASED ON THE VARIABLE CONNECTIVITY INDEX

In contrast to the difficulties encountered with an interpretation of a linear combination of descriptors the variable connectivity index $^1\chi^f$ bypasses these perplexities. The simple connectivity index introduced different weights to different bond types in a molecule so that more "exposed" terminal CC bonds have a greater weight than "buried" internal CC bonds. The weights *x* and *y* in generalized connectivity index formally augment the valence for carbon and oxygen differently and independently. The negative weights increase the role of oxygen atoms in alcohols, which now play a dominant role in determining the relative boiling points of alcohols. The value for $y = -0.92$ gives to the contribution of the C−O bond much greater weight than the contributions of the remaining C−C bonds. For example, in the case of 3-methyl-2-butanol (for which the calculation of the variable connectivity was illustrated in Table 1) we have, when the values $x = +0.10$ and $y = -0.92$ are substituted in the expression for ${}^1\chi^f$, the following numerical results:

bond	contribution	numerical value
CH−CH$_3$	$1/\sqrt{(1+x)(3+x)}$	0.541530
CH−CH	$1/(3+x)$	0.322581
CH−OH	$1/\sqrt{(3+x)(1+y)}$	2.008048
total C−C bonds contribution		1.947172
molecule		3.9552

As we see from the above all four C−C bonds make a smaller contribution to the molecular connectivity index than the single C−O bond. Moreover, the variable connectivity index offers insights into the relative magnitudes of the boiling points among some isomers.

We can better understand isomeric variations in the BP among isomers for primary, secondary, and tertiary alcohols when we compare the C−O bond contributions to ${}^1\chi^f$ in different alcohols. The contributions of C−O bonds in primary, secondary, and tertiary alcohols will be given by $1/\sqrt{(1+x)(1+y)}$, $1/\sqrt{(2+x)(1+y)}$, and $1/\sqrt{(3+x)(1+y)}$, respectively. For the optimal values of x and y these contributions are 3.370999, 2.439750, and 2.008048, respectively. This points to the relative magnitudes of BP: the primary alcohols will have the largest boiling point, the secondary will be the next, and the tertiary alcohols will have the smallest boiling point. Nevertheless, in few cases this regularity is violated. For example, the boiling points for 2-methyl-3-hexanol and 2-methyl-2-hexanol are both reported as 143, despite that the former is *secondary* and the latter *tertiary* alcohol. Using the optimal variable connectivity index the predicted values for the boiling points of the two alcohols are 150.7 and 141.2. This should been expected in view of the relative contribution to the variable connectivity index by secondary and tertiary C−OH group. Similar predictions follow also from the best regression equation obtained by CODESSA. We find that the predicted values for the boiling points of the two alcohols using CODESSA are 143.65 °C and 135.20 °C, respectively, pointing again to an "excessive" difference for the two boiling points reported experimentally to be the same. If the experimental data are correct (and at this point there are no indications that this may not be the case), it remains to be investigated why such discrepancies occur. Could this be due to some experimental considerations (impurities, etc.), or are there some structural features in which the two alcohols differ that the indices used do not adequately characterize? For example, atoms in tertiary structures are more "crowded" than atoms in corresponding secondary structure. Could this cause a "shift" in the boiling points between the two types of alcohols beyond the values calculated by topological indices that do not take into account crowding of atoms? There are molecular descriptors that can characterize "crowding" of atoms. Besides the count of paths of length three P$_3$, the polarity index of Wiener, recently additional such descriptors were considered.[20,56,57] Investigation of the possible role of "crowding" of atoms is, however, outside the scope of the present study (and if making a significant contribution it is likely to equally affect the "fixed" descriptors of CODESSA as well as the variable connectivity index).

6. ON LIMITATION OF THE "FIXED" MOLECULAR DESCRIPTORS

Let us illustrate limitations of the traditional topological indices designed for heteroatoms by returning to the best single descriptor regression for the boiling points of alcohols as found by CODESSA. The best descriptor found by CODESSA was Randić index (order 1), and not Kier and Hall index (order 1), that discriminates carbon and oxygen atoms. One may ask why the simple connectivity index is a better descriptor for systems having heteroatoms than valence connectivity index. One would expect that an index that differentiates between carbon atoms and oxygen atoms is bound to be better suited for characterization of heteroatomic systems than a simple connectivity index that does not discriminate between carbon and oxygen. But apparently it is not in this particular application.

There are additional illustrations of the same phenomenon already cited in the book of Kier and Hall. For example, in considering partition coefficients of hydrocarbons Kier and Hall found that ${}^1\chi$ is the best correlation variable for ethers

$$\log P = -1.411 + 0.988\ {}^1\chi$$

$$r = 0.9680, \quad s = 0.091, \quad N = 12$$

and that the addition of ${}^1\chi^v$ does not give significant improvement.[58]

The situation, although somewhat unexpected is not surprising, because in numerous applications of both connectivity indices ${}^1\chi$ and ${}^1\chi^v$ it was observed that sometimes ${}^1\chi$ has been a better descriptor for some properties of heteroatomic systems. Kier and Hall[50] found that ${}^1\chi$ gives a very good correlation for the boiling points of aliphatic ethers ($r = 0.9851$; $s = 5.69$ °C). This regression is only slightly improved by inclusion of ${}^1\chi^v$ ($r = 0.9882$; $s = 5.39$ °C). But why ${}^1\chi^v$ used alone did not give better results? The question could have been considered some 20 years ago, although it is not clear whether it could have been resolved.

Similarly Basak and co-workers[59] in a comparative study of lipophilicity of barbiturates using topological indices and a quadratic regression found that ${}^1\chi$ better correlates than ${}^1\chi^v$ with molar concentration of applied drug producing anesthesia in mice and with experimental log(1/C) values causing inhibition of arbacia egg cell division, thus showing that in some applications of connectivity indices in QSAR again discrimination of carbon and oxygen as implied by the valence connectivity index apparently is ineffective, if not counterproductive. In another comparative study Basak et al.[60] considered aquatic toxicity of alcohols using several information-theoretic indices based on atomic neighborhoods and interatomic distances as well as Wiener index W, the connectivity index ${}^1\chi$ and valence connectivity index ${}^1\chi^v$, and log P. The best quadratic regression for data on 10 alcohols was obtained with ${}^1\chi$ ($r = 0.99$, $s = 0.13$, and $F = 1090$). The valence connectivity index for comparison gave the following statistical parameters: $r = 0.99$, $s = 0.19$, and $F = 512$.

With introduction of the variable connectivity index, however, insight and understanding into this limited "superiority" of $^1\chi^v$ over $^1\chi$ and vice versa has come to light. Briefly, the answer lay in the fact that *different* molecular properties require *distinct* optimal parameters. In other words, there is no universal valence connectivity index that would apply to all properties of heteroatomic structures. Hence, any set of preselected "rules" that fix the relative weights for heteroatoms in topological indices may better suit some molecular properties but will equally fail several others. The answer is the variable molecular descriptors.

7. CONCLUDING REMARKS

As has been demonstrated on several recent structure−property regressions variable molecular descriptors, and the variable connectivity index in particular, allow one to use simple, where other molecular descriptors require multivariate, regression. By using fewer descriptors in MRA interpretation of its results will become easier. Interpretation of regression analysis continues to plague current use of MRA. Variable descriptors will not displace the traditional "fixed" molecular descriptors of today. This is not only because their use implies more computational time but also because there are instances where a variable connectivity index did not give better results that the best alternative topological indices. In addition use of "fixed" topological indices may be a matter of choice in application to combinatorial libraries. At least that is likely to continue for a while in preliminary screenings of huge combinatorial libraries containing over 100 000−1 000 000 virtual compounds.

However, in applications to MRA using a smaller number of compounds (few hundreds at most), we would not be surprised to see that many currently used topological indices, particularly those adapted for heteroatomic structures, will be found obsolete. We can justify such expectations on already accumulated experience with the variable connectivity index and variable path numbers that clearly show that selection of descriptors depends on the property considered. Yet today's hetero-descriptors are fixed and "fixed" descriptors will at best approach optimal descriptors for few properties at most.

REFERENCES AND NOTES

(1) Trinajstić, N. *Chemical Graph Theory;* CRC Press: Boca Raton, FL, 1992; Chapter 10, pp 225−273.

(2) Randić, M. *Topological Indices.* In *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, 1998; pp 3018−3032.

(3) Balaban, A. T. *Historical developments of topological indices,* in: *Topological Indices and Related Descriptors in QSAR and QSPR;* Devillers, J., Balaban, A. T., Eds; in press.

(4) Randić, M.; Basak, S. C. Variable Molecular Descriptors, a chapter in a book published by Visva Bharaty University, Santiniketan, Bengal, India (in print).

(5) Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24,* 417−441 and 489−520.

(6) Okuyama, T.; Miyashita, Y.; Kanaya, S.; Katsumi, H.; Sasaki, S. I.; Randić, M. Computer assisted structure-taste studies on sulfates by pattern recognition method using graph theoretical invariants. *J. Comput. Chem.* **1988**, *9,* 636−646.

(7) Everitt, B. *Cluster Analysis;* John Wiley & Sons: New York (about 1973).

(8) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists;* VCH: Weinhein, 1993.

(9) Basak, S. C. POLLY; Natural Resources Research Institute, Duluth, University of Minnesota: Duluth, MN.

(10) Hall, L. H. MOLCONN-Z; Hall Associates Consulting: Quincy, MA, 1991.

(11) Katritzky, A. R.; Lobanov, V.; Karelson, M. CODESSA (COmprehensive DEscriptors for Structural and Statistical Analysis); University of Florida, Gainesville, FL.

(12) Randić, M. On the characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(13) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity V: Connectivity series applied to density. *J. Pharm. Sci.* **1975**, *65,* 1226−1230.

(14) Randić, M. On characterization of chemical structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 7, 672−687.

(15) Randić, M. Novel graph theoretical approach to heteroatoms in QSAR. *Chemometrics Intel. Lab. Syst.* **1991**, *10*, 213−227.

(16) Randić, M. On computation of optimal parameters for multivariate analysis of structure−property relationship. *J. Comput. Chem.* **1991**, *12,* 970−980.

(17) Randić, M.; Basak, S. C. Optimal molecular descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261−266.

(18) Randić, M.; Basak, S. C. Multiple regression analysis with optimal molecular descriptors. *SAR QSAR Environ. Res.* In press.

(19) Randić, M.; Dobrowolski, J. Cz. Optimal molecular connectivity descriptors for nitrogen-containing molecules. *Int. J. Quantum Chem.* **1998**, *70,* 1209−1215.

(20) Randić, M. High quality structure − property regressions. Boiling points of smaller alkanes. *New J. Chem.* **2000**, *24,* 165−171.

(21) Randić, M.; Plavšić, D.; Lerš, N. Variable connectivity index for cycle-containing structures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 657−662.

(22) Randić, M.; Basak, S. C. On construction of high quality structure−property-activity regressions: The boiling points of sulfides. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 899−905.

(23) Randić, M.; Basak, S. C. On use of the variable connectivity index $^1\chi^f$ in QSAR: Toxicity of Aliphatic Ethers; Poster at Second Indo-U.S. Workshop on Mathematical Chemistry, Duluth, MN, May 30−June 3, 2000. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 614−618.

(24) Randić, M.; Basak, S. C.; Pompe, M.; Novic, M. Prediction of gaschromatographic retention indices using variable connectivity index. *Acta Chim. Slovenica* Submitted.

(25) Randić, M.; Mills, D.; Basak, S. C.; On use of variable connectivity index or characterization of amino acids. *Int. J. Quantum Chem.* **2000**, *80*, 1199−1209.

(26) Randić, M.; Mills, D.; Basak, S. C.; Pogliani, L. On characterization of several physicochemical properties of amino acids. *J. Chem. Inf. Comput. Sci.*, in press.

(27) Shannon, C. E. A mathematical theory of communication. *Bell. Syst. Technol.* J. **1948**, *27,* 379−423.

(28) Lučić, B.; Trinajstić, N. New developments in QSPR/QSAR modeling based on topological indices. *SAR QSAR Environ. Res.* **1997**, *7*, 45−62.

(29) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A new efficient approach for variable selection on multiregression: Prediction of gas chromatographic retention times and response factor. *J. Chem. Inf. Comput. Sci.* **1999**, *39,* 610−621.

(30) Amić, D.; Davidović-Amić, D.; Jurić, A.; Lučić,; Trinajstić, N. Structure−activity correlation of flavone derivatives for inhibition of cAMP phosphodiesterase. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1034−1038.

(31) Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D. The structure−property models can be improved using the orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35,* 532−538.

(32) Soskić, M.; Plavšić, D.; Trinajstić, N. Link between orthogonal and standard multiple linear regression models. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829−832.

(33) Randić M. Retro-Regression - - Another Important Multivariate Regression Improvement. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 602−606.

(34) Randić, M.; Pompe, M. Retro-regression - - A way to resolve multivariate regression ambiguities. *New. J. Chem.* Submitted for publication.

(35) Randić, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15,* 517−525.

(36) Randić, M. Resolution of ambiguities in structure−property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311−370.

(37) Randić, M. Fitting of non linear regressions by orthogonalized power series. *J. Comput. Chem.* **1993**, *14,* 363−370.

(38) Randić, M. Curve fitting paradox. *Int. J. Quantum Chem: Quantum Biol. Symp.* **1994**, *21,* 215−225.

(39) Amić, D.; Davidović-Amić, D.; Beslo, D.; Lučić, B.; Trinajstić, N. The use of the ordered orthgonalized multivariate linear regression in a structure−activity study of coumarin and flavonoid derivatives as inhibitors of aldose reductase. *J. Chem. Inf. Comput. Sci.* **1997**, *71,* 5581−586.

(40) Cramer, III, R. D. BC(DEF) parameters. 1. The intrinsic dmensiobality of intermolecular interactions in liquid state, *J. Am. Chem. Soc.* **1980**, *102,* 1837−1849.

(41) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Characterization of the molecular similarity of chemicals using topological invariants. *Advances Molecular Similarity* **1998**, *2,* 171−185.

(42) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting properties of molecules using graph invariants. *J. Math. Chem*. **1991**, *7,* 243−272.

(43) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* **1988**, *19,* 17−44.

(44) Basak, S. C.; Gute, B. D. Characterization of molecular structures using topological indices. *SAR QSAR Environ. Res.* **1997**, *7,* 1−21.

(45) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69,* 17−20.

(46) Kier, L. B. A shape index from molecular graphs. *Quant. Struct. − Act. Relat.* **1985**, *4,* 109−116.

(47) Kier, L. B. Shape indexes or orders one and three from molecular graphs. *Quant. Struct. − Act. Relat.* **1986**, *5,* 1−7.

(48) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research;* Academic Press: New York, 1976; p 165.

(49) Kier, L. B. Distinguishing atom differences in a molecular graph shape index. *Quant. Struct. − Act. Relat.* **1986**, *5,* 7−12.

(50) Kier, L. B. Indexes of molecular shape from chemical graphs. *Acta Pharm. Jugosl.* **1986**, *36,* 171−188.

(51) Kier, L. B. Indexes of molecular shape from chemical graphs. *Med. Res. Rev.* **1987**, *7,* 417−440.

(52) Randić, M. On characterization of the shape of molecular graphs. *SAR QSAR Environ. Res.* Submitted for publication.

(53) Randić, M. Novel shape descriptors for molecular graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41,* 607−613.

(54) Randić, M. On characterization of molecular attributed. *Acta Chim. Slovenica* **1998**, *45,* 239−252.

(55) Randić, M.; Zupan, J. On the interpretation of well-known topological indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550−560.

(56) Lukovits, I.; Linert, W. Polarity-numbers of cycle-containing structures. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 715−719.

(57) Randić, M. On characterization of spatial interference of close methyl groups by topological indices. *J. Math. Chem.* Submitted for publication.

(58) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research;* Academic Press: New York, 1976; pp 136−137.

(59) Basak, S. C.; Monsrud, L. J.; Frane, C. M.; Magnuson, V. R. A comparative study of lipophilicity and topological indices in biological correlation. *Acta Pharm. Jugosl.* **1986**, *36,* 81−95.

(60) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Comparative study of lipophilicity and topological indices in biological correlation. *J. Pharm. Sci.* **1984**, *73,* 429−437.

CI000119R