NEW IDEA FOR THE TOPOLOGICAL INDEX EVALUATION AND TREATISE MULTIPLE REGRESSION WITH THREE INDEPENDENT VARIABLES. SATURATED HYDROCARBONS USED LIKE A MODEL

E. CORNWELL

Departamento de Química Inorgánica y Analítica, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Casilla 233, Santiago, Chile

E-mail: ecornwel@ciq.uchile.cl

ABSTRACT

In QSRR discipline an easy novel to used parameter was designed (Vc) for evaluated classical topological index (W, $^{1}\chi$, Z, MTI) and two new generation ones (Xu, $^{1}\chi^{h}$). Regression between Vc and $^{1}\chi^{h}$ presented a correlation index (r) of 0,9992, a surprising high value in comparison with that founds commonly in QSPR/QSAR discipline. Through Vc parameter, an idea to treatise multiple three independent variable regression is present. Model of 35 saturated hydrocarbons were used.

INTRODUCTION

A mayor part of the current research in mathematical chemistry, chemical graph theory and quantitative structure-activity-property relationship studies involves topological indices. Topological indices (TIs) are numerical graph invariants that quantitatively characterize molecular structure.

A graph G = (V, E) is and ordered pair of two set V and E, the former representing a

nonempty set and the latter representing unordered pairs of elements of the set V. When V represent the atoms of the molecule and element of E symbolize covalent bonds between pairs of atoms, then G becomes a molecular graph. Such graph depicts the topological of the chemical species. A graph is characterized using graph invariants, an invariant may be a polynomial, a sequence of number, or a single number as the case used in the present article. A single number numerical graph invariant that characterize the molecular structure is called a topological index.

Application of graph theory to chemical and to structure-property-activity (QSPR/QSAR) relationships has led to the emergence of several critical graph-theoretical indices.

First application of graph-theoretical invariants in studies of structure-properties relationship (QSPR) was proposed by Weiner[1)] [Weiner index, (W)]. However, it was after Randic[2)] proposed a topological index for characterization of molecular branching [(Randic index, ($^1\chi$)] that dramatic expansion of studies in the area started. The two former topological indices indicated, plus Hosoya[3)] (Z), Schultz[4)] (MTI), Ren[5)] (X_u), and C.Yang and C Zhong[6)] ($^1\chi^h$) indices are evaluated using a new idea based on the three physicochemical properties of the molecules, the molar refraction index (MR) the critical pressure (Pcr) and the critical volume (Vcr). In previous report[7)] y proved that these physicochemical properties correlated so well with logarithmic relative retention time relative to n-hexane (log (t_{rr})) in GLC analysis by means of a linear relation (y = m*x + n). The novel relation proposed by the author (Vc) is a general idea and was proved in to 35 saturated hydrocarbons[8)] taken like a model, each one of these hydrocarbons are characteristize by a ordered set x_i, y_i, z_i , x_i is the molar refraction index[7)] (MR), y_i the critical pressure[7)] (Pcr) and z_i is a critical volume[7)] (Vcr) of an hydrocarbon i . Vc is the Euclidian distance of a particular set x_i, y_i and z_i to one hydrocarbon i belonging to the 34 hydrocarbon set respect to ethane with ordered set x_o, y_o and z_o The election of other referent hydrocarbon, produce results not satisfactory as well as methane, perhaps, the cause of that, is a molecular structure differences of any one of the 34 hydrocarbons relative to ethane.

All correlations treatise in this issue (y = mx + n) was referred to the linear regression between Vc and the topological indices cited or log t_{rr} versus all other variables in study. Other physical-chemical properties cited in my last published issue[7)] were used in three elements set to defined Vc and not good results were obtained as the three proposed one (MR, Pcr, Vcr)

Through a linear regression of Vc with all proposed topological index, permit us to order its in accordance with the magnitude correlation index (r), order that is the same when we correlated the relative GLC retention time respect to hexane expressed like a logarithm of these magnitude (log t_{rr})[8)] with the same set of indices, this characteristic indicated that the idea involved in Vc (definition of Vc with using appropriate physical-chemistry properties) is interesting for evaluate topological index, The second categories is automatic established in function of the first ordering.

All regression function log t_{rr} vs.(x_i, y_i, z_i); (x_i, y_i); (x_i, z_i); (y_i, z_i) presented similar R^2 values and similar R^2 value respect to regression function log t_{rr} vs Vc

Is necessary to point out that the interpretation of the parameters of multivariable regression is valid if these parameters are in orthogonal form[9)] and the number of independent variable used must be in accordance with the number of cases treatise, if not,

R^2 value is false by excess[10)], these limitation are not present in all types of regression used in the present study (y = mx + n) where Vc, log t_{rr} and the topological indices were used for the 35 saturated hydrocarbons model.

The results obtained in this issue indicated that is possible used the idea of Vc parameter to the evaluation of topological indices applied to other organic homologue series. And to reduced multiple regression till to three independent variable to linear regression of the type y = mx + n

PROCEDURE

The Vc parameter is obtained through the distance (D) between the set ordered (x_i , $y_{i,}$ z_1) of a particular saturated hydrocarbon and the pair ordered (x_o, $y_{o,}$ z_o) corresponding to ethane, particularly (11,48, 50.299, 147.5) distance D is obtained by Euclidian formulae, equation 1

$$D = [(x_i - x_o)^2 + (y_i - y_o)^2 + (z_i - z_o)^2]^{0.5} = Vc \quad (1)$$

This equation (1) was applied to 35 saturated hydrocarbons with the values expressed in columns 5-7 presented in Table 1 (For i = 0 using equation (1) the ethane distance is equal 0). Results for Vc values are expressed in Table 2, column 9. In column 10 are presented the calculated Vc from equation (2) and at column 11, the absolute error percent of Vc respect to $Vc_{\ calculated}$ by equation (3)

The regression of Vc parameter respect to $^1\chi^h$ are defined by equation 2

$$Vc = -94.732(\pm 2.416)+149.532(\pm 1.010)*{}^1\chi^h \quad (2)$$
R^2 = 99.84%
r = 0.9992
s.d = 3.3994
F = 21896.8

Table 1. Relative retentions time respected to n-hexane and the physicochemical properties of hydrocarbons.

	Compounds	trr Relative	Log trr	M.R.	Pcr bar	Vcr cm^3/mol
1	Etane	0,02	-1,6990	11,48	50,299	147,5
2	Propane	0,07	-1,1549	16,08	44,091	203,5
3	2-Methylpropane	0,17	-0,7696	20,85	39,357	243,5
4	n-Butane	0,21	-0,6778	20,68	38,965	259,5
5	2,2-Dimethylpropane	0,23	-0,6383	25,25	35,473	304,5
6	2-Methylbutane	0,36	-0,4437	25,45	35,013	309,5
7	n-Pentane	0,44	-0,3565	25,27	34,684	315,5
8	2,2-Dimethylbutane	0,59	-0,2291	29,85	31,740	360,5
9	2,3-Dimethylbutane	0,75	-0,1249	30,23	31,633	359,5
10	2-Methylpentane	0,77	-0,1135	30,05	31,350	365,5
11	3-Methylpentane	0,87	-0,0605	30,05	31,350	365,5
12	n-Hexane	1,00	0,0000	29,87	31,070	371,5
13	2,2-dimethylpentane	1,23	0,0899	34,45	28,566	416,5
14	2,4-Dimethylpentane	1,29	0,1106	34,83	28,475	415,5
15	2,2,3-trimetilbutano	1,37	0,1367	34,63	28,812	410,5
16	3,3-Dimethylpentane	1,61	0,2068	34,45	28,566	416,5
17	2-Methylhexane	1,77	0,2480	34,65	28,233	421,5
18	2,3-Dimethylpentane	1,81	0,2577	34,83	28,475	415,5
19	3-Methylhexane	1,91	0,2810	34,65	28,233	421,5
20	3-Etilpentano	2,09	0,3201	34,65	28,233	421,5
21	2,2,4-Trimethylpentane	2,16	0,3345	39,23	26,057	466,5
22	n-Heptane	2,35	0,3711	34,47	27,995	427,5
23	2,2,3,3-Tetramethylbutane	2,81	0,4487	39,03	26,353	461,5
24	2,2-Dimethylhexane	2,84	0,4533	39,05	25,846	472,5
25	2,5-Dimethylhexane	3,09	0,4900	39,43	25,767	471,5
27	2,2,3-Trimetilpentano	3,14	0,4969	39,43	25,767	471,5
28	2,4-Dimetilhexano	3,14	0,4969	39,43	25,767	471,5
26	3,3-Dimethylhexane	3,34	0,5237	39,05	25,846	472,5
29	2,3,4-Trimetilpentano	3,60	0,5563	39,61	25,978	465,5
30	2,3,3-Trimetilpentano	3,73	0,5717	39,43	25,765	471,5
31	2,3-Dimetilhexano	3,93	0,5944	39,43	25,767	471,5
32	3-etil-2-Metilpentano	3,98	0,5999	39,05	25,846	472,5
33	2-Methtylheptane	4,14	0,6170	39,25	24,559	477,5
34	4-Methylheptane	4,19	0,6222	39,25	25,559	477,5
35	3,4-Dimethylhexane	4,26	0,6294	39,43	25,767	471,5

M.R indicated molar refractions index. Pcr, critical pressure and Vcr indicated critical volumn

Table 2. Logarithm of retention time and diverses topological and parameter index (Vc)

Nº	log trr	W	Z	$^{1}\chi_{v}$	MTI	Xu	$^{1,h}\chi$	Vc	Vc. calc.	% error
1	-1,6990	1	2	1,0000	4	0,0000	0,6666	0,0000	4,9456	*
2	-1,1549	4	3	1,4142	16	0,7188	1,0000	56,3430	54,7996	2,7
3	-0,7696	9	4	1,7321	36	1,2568	1,3093	96,6220	101,0513	4,6
4	-0,6778	10	5	1,9142	38	1,3728	1,3819	112,5720	111,9059	0,6
5	-0,6383	16	5	2,0000	64	1,7252	1,6330	157,6980	149,4534	5,2
6	-0,4437	18	7	2,2701	68	1,8681	1,7140	162,7200	161,5654	0,7
7	-0,3565	20	8	2,4142	74	1,9948	1,7566	168,7240	167,9355	0,5
8	-0,2291	28	9	2,5607	106	2,3052	2,0517	213,8070	212,0624	0,8
9	-0,1249	29	10	2,6427	108	2,3429	2,0542	212,8200	212,4362	0,2
10	-0,1135	32	11	2,7701	118	2,4753	2,0836	218,8220	216,8325	0,9
11	-0,0605	31	12	2,7542	114	2,4381	2,1142	218,8220	221,4082	1,2
12	0,0000	35	13	2,9142	128	2,5923	2,1316	224,8240	224,0100	0,4
13	0,0899	46	14	3,0607	170	2,9015	2,4166	269,8760	266,6266	1,2
14	0,1106	48	15	3,1251	176	2,9472	2,4080	268,8870	265,3407	1,3
15	0,1367	42	13	2,9434	156	2,7652	2,3961	263,8760	263,5612	0,1
16	0,2067	44	16	3,1213	162	2,8423	2,4670	269,8760	274,1630	1,6
17	0,2480	52	18	3,2701	190	3,0645	2,4589	274,8870	272,9518	0,7
18	0,2577	46	17	3,1807	168	2,8963	2,4515	268,8870	271,8453	1,1
19	0,2810	50	19	3,3081	182	3,0167	2,4840	274,8870	276,7051	0,7
20	0,3245	48	20	3,3461	174	2,9613	2,5095	274,8870	280,5182	2,0
21	0,3345	66	19	3,4165	242	3,3668	2,7393	319,9200	314,8806	1,6
22	0,3711	56	21	3,4142	204	3,1691	2,5066	280,8870	280,0845	0,3
23	0,4487	58	17	3,2500	214	3,1762	2,7402	314,9120	315,0152	0,0
24	0,4533	71	23	3,5607	260	3,4859	2,7924	325,9190	322,8208	1,0
25	0,4900	74	25	3,6259	270	3,5321	2,7866	324,9270	321,9535	0,9
26	0,4969	63	22	3,4814	230	3,3071	2,7908	324,9270	322,5815	0,7
27	0,4969	71	26	3,6639	258	3,4783	2,8086	329,9270	325,2432	1,4
28	0,5237	67	25	3,6213	244	3,4289	2,8320	325,9190	328,7422	0,9
29	0,5563	65	24	3,5534	236	3,3464	2,7875	318,9290	322,0881	1,0
30	0,5717	62	23	3,5040	226	3,2849	2,8092	324,9280	325,3329	0,1
31	0,5944	70	27	3,6807	254	3,4643	2,8214	324,9270	327,1572	0,7
32	0,5999	67	28	3,7187	242	3,3992	2,8435	325,9190	330,4618	1,4
33	0,6170	79	29	3,7701	288	3,6805	2,8339	331,0020	329,0263	0,6
34	0,6222	75	30	3,8081	272	3,5699	2,8537	330,9260	331,9871	0,3
35	0,6204	68	29	3,7187	246	3,4227	2,8489	324,9270	331,2693	2,0

The meanings of the column titles are defined in the next,* indicated reference substance

Where r is the correlation coefficient, s.d is standard error of estimate and F is Fisher-ratio. Analysis of variance of the above correlation is in Table 3

Table 3. Analysis of variance of equation number 2 correlation

Source	Sum of Squares	Df	Mean Square	F-Ratio	p-Value
Model	253043.0	1	253043.0	21896.79	0.000
Residual	381.354	33	11.5562		
Total (Corr)	253425.0	34			

In this correlation, since the p-value in the ANOVA Table 3 is less than 0.01 there is a statistically significant relationship between both variables at the 99% confidence level. The R-Squared statistic indicates that the model explains 99.84% of the variability in Vc, r indicate a strong relationship between the variables, s.d error shows the standard deviation of the residual to be 3.399

The relation Vc calculated by means of equation (2) versus Vc values is defined by equation Nº 3

$Vc_{calculated} = 0.39046\ (\pm 1.798) + 0.9985\ (\pm 0.0067)*Vc$ (3)
$R^2 = 0.9985$
R = 0.9993
s.d = 3.39
F = 21896.12

The analysis of this correlation is made by the same way that equation (2) but without ANOVA analysis, not necessary, because little percentage of errors existents between calculated Vc respect to Vc

Tabla 4. Correlation matriz of topological indices and parametrix index

		Log trr	Z	W	MTI	$^1\chi$	Xu	$^1\chi^h$	Vc
		1	2	3	4	5	6	7	8
Log trr	1	**1,0000**	0,9150	0,9401	0,9408	0,9906	0,9911	0,9912	0,9890
Z	2	0,9150	**1,0000**	0,9711	0,9679	0,9451	0,9105	0,9197	0,9132
W	3	0,9401	0,9711	**1,0000**	0,9999	0,9651	0,9545	0,9604	0,9599
MTI	4	0,9408	0,9679	0,9999	**1,0000**	0,9648	0,9561	0,9618	0,9617
$^1\chi$	5	0,9906	0,9451	0,9651	0,9648	**1,0000**	0,9924	0,9902	0,9883
Xu	6	0,9911	0,9105	0,9545	0,9561	0,9924	**1,0000**	0,9932	0,9954
$^1\chi^h$	7	0,9912	0,9197	0,9604	0,9618	0,9902	0,9932	**1,0000**	0,9992
Vc	8	0,9890	0,9132	0,9599	0,9617	0,9883	0,9945	**0,9992**	**1,0000**

In Table 4 the matrix of all possible combinations of regression were present, each a_{ij} matrix term represent the correlation index (r) where the linear relation Vc f ($^1\chi^h$) is the biggest one (0.9992) In function of the r values matrix (terms $a_{8,2}$ to $a_{8,7}$) evaluated by Vc f (TIs) studied, is possible to ordered all considered (TIs), the order is: [Z < W< MTI< $^1\chi$ < Xu < $^1\chi^h$] that is the same order considering log t_{rr} f (TIs) (terms $a_{2,1}$ to $a_{7,1}$) This transitivity property is useful to evaluated the correlation of an experimental relation (log t_{rr}) with topological indices knowing a priori the matrix of r values related to Vc f (TIs) function.

The linear regression (y = mx + n) log t_{rr} versus Vc is statistically very similar to the multiple regression (y = a + bx + cy + dz) log t_{rr} versus independent variables MR, Pcr and Vcr but the great F ratio value indicated a more predictability capacity for linear model. See Table 5

Table 5. Statistical results of linear and multiple regression model.

Regression model	R- squared	Standard error of estimate	F- ratio
Linear model	97.8039	0.0830	1469.65
Multiple regression	98.416	0.0727	642.03.40

These results, indicate that using the concept of Euclidian distance in space E^3 it is possible to reduced multiple regression with three independent variables to a linear regression of the form y = mx + n and in this way solved the problem of orthogonal procedure of factors or

to depend of the number points analyzed[9, 10)] these problems was mentioned in the introduction.

Note. Df: Means liberty grade, f indicated function. Regressions were made by Stat-Graphic Plus 4 Software.

CONCLUSSIONS

1. Vc is useful parameter for ordered (TIs) indices in function of its regressions values r respect to GLC relative retention times.

2. Any multiple regressions are possible to reduced to a linear expression by means of Vc parametric idea, only a maximum of three independent variables are permit

3. A very significant linear correlation exist between Vc and $^{1}\chi^{h}$ this implies a great dependence between $^{1}\chi^{h}$ with critical pressure and critical volume of the hydrocarbons. In fact, this implies a very good significant topological criteria to defined $^{1}\chi^{h}$

REFERENCES

1. Z. Mihalic, N. Trinajstic. J. Chem Educ. 69, 701-712 (1992)

2. M. Randic. J. Amer. Chem. Soc. 97,6609-6615 (1975).

3. H. Hosaya. Boll. Chem. Soc. Japan 44, 2332-2339 (1971)

4. H. P. Schultz. J. Chem. Inform. Comput. Sci. 29, 227-228 (1989).

5. B. Ren. J. Chem. Inform. Comput. Sci. 39, 139-143 (1999).

6. C. Yang., C. Zhong. J. Chem. Inform. Comput. Sci. 43, 1998-2004 (2003).

7. E. Cornwell. J. Chil. Chem. Soc. 50, 483-487 (2005).

8. G. Zweig, J. Sherma "Handbook of Chromatography" CRC Press (1976). page 50.

9. M. Randic. J. Chem. Inform. Comput. Sci. 37, 672-687 (1997).

10. J. C. Toplis., R. P. Edwards. J. Med. Chem. 22, 1238-1244 (1979)