Journal of Chromatography, 648 (1993) 389-393
Elsevier Science Publishers B .V., Amsterdam

CHROM. 25 347

Prediction of gas chromatographic retention indices of some benzene derivatives

M. Jalali-Heravi* and Z. Garkani-Nejad

Chemistry Department, Shahid Bahonar University of Kerman, Kerman 76169 (Iran)

(First received February 22nd, 1993; revised manuscript received May 26th, 1993)

ABSTRACT

Gas chromatographic retention indices for some benzene derivatives on Apiezon MH were successfully modelled with the aid of a computer. Numerical descriptors were calculated and multiple linear regression analysis methods were used to generate model equations relating structural features to Kováts retention indices. These descriptors encode topological, geometric, electronic and calculated physical properties of the molecules. A model with $R = 0.998$ and $SE = 10.067$ was generated. Calculations of retention indices for a prediction set show that this model has a good predictive ability.

INTRODUCTION

The Kováts retention index in gas chromatography (GC) represents the retention behaviour of a compound relative to a standard set of hydrocarbons, utilizing a logarithmic scale. The retention index, I_A, for compound A is defined as

$$I_{\mathrm{A}} = 100\,N + 100 \cdot \frac{\log \mathrm{MA)} - \log t_{\mathrm{R}}(N)}{\log t_{\mathrm{R}}(N+1) - \log t_{\mathrm{R}}(N)} \quad (1)$$

where $t_R(A)$ is the adjusted retention time for compound A and $t_R(N + 1)$ and $t_R(N)$ are the adjusted retention times for *n*-alkanes of carbon number N + 1 and N that are larger and smaller, respectively, than the adjusted retention time for the unknown.

The identification of many compounds is often accomplished on the basis of GC peak comparisons with a standard sample of the suspected material. However, it is not always possible to obtain samples of pure standard materials for such comparisons. Therefore, the development of a theoretical model for estimating the retention index seems to be necessary. In this study, computer-assisted methods were employed to generate a statistical relationship between molecular-based structural parameters (descriptors) and the observed retention indices for some benzene derivatives. These techniques are based on the construction of linear mathematical models relating the observed retention indices to numerically encoded structural parameters called ***descriptors. These*** models have the general form

$$S = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

where S is the predicted retention index for the molecule of interest, the X_i are numerical descriptors, the b_i ***are*** coefficients determined from a linear regression analysis of a set of observed retention indices and n denotes the number of descriptors in the model.

EXPERIMENTAL

The methodology used in this study consists of three fundamental stages: (a) selection of data set, (b) molecular descriptor generation and (c)

* Corresponding author.

regression analysis. Computations of descriptors were performed by using some FORTRAN programs developed in our laboratory. The SPSS/PC package [1] was used for regression calculations.

Data set

The experimental data used in this study were reported by Khorasani [2]. The Kováts retention indices were determined on a stainless-steel column (2 m x 1.8 mm I.D.) packed with 10% (w/w) of hydrogenated Apeizon M (Apiezon MH) coated on acid-washed, dimethylchlorosilane-treated Chromosorb W (100-120 mesh) [2]. The retention indices for monosubstituted benzenes were normalized to 150°C by using least-squares plots of retention index against temperature. The other compounds were measured in the range 90°C (for fluorotoluene)–180°C (for bromochlorobenzene) [2]. Retention indices for 38 benzene derivatives studied ranged from 664.1 to 1287.7 index units (i.u.), with a mean retention index of 965.2 i.u. These compounds were divided randomly into a set used for constructing the equations (training set) and a set used for testing the validity of the generated model (prediction set) (Table I).

Descriptor generation

A total of 58 separate molecular structure descriptors were calculated for each compound in the data set. These descriptors can be classified into four major groups: topological, geometric, electronic and physico-chemical.

Topological descriptors include fragment, substructure and environment descriptors [3] and molecular connectivity indices [4]. Geometric descriptors include principal moments of inertia [5], shadow areas [6], Van der Waals volume [7], surface area [8], principal axes of the molecules and kappa index [9]. Electronic descriptors consist of dipole moments, molar refraction [10], electron density and partial charges of atoms with the most negative and positive charges and distance between atoms with the most positive and negative charges. Calculated physical property descriptors include molecular polarizability [11] and the logarithm of the partition coefficient in octanol-water (log P) [12].

Geometric and electronic descriptors depend on the three-dimensional coordinates of atoms. Therefore, in order to calculate these types of descriptors one needs to optimize the molecular structure of each molecule. In this work, MNDO [13], which is a semi-empirical molecular orbital method, was used for such an optimization.

TABLE I

DATA SET

No.	Compound	No.	Compound	No.	Compound
Training set		14	p-Fluoroanisole	28	o-Fluorotoluene
1	Benzene	15	m-Chloroanisole	29	o-Fluoroanisole
2	Fluorobenzene	16	m-Methylanisole	30	o-Xylene
3	Chlorobenzene	17	m-Xylene	31	o-Bromochlorobenzene
4	Bromobenzene	18	m-Chlorobromobenzene	32	o-Chlorofluorobenzene
5	Toluene	19	m-Bromotoluene		
6	Anisole	20	m-Fluorotoluene	***Prediction set***	
7	p-Chloroanisole	21	m-Chlorotoluene	1	p-Chlorofluorobenzene
8	p-Xylene	22	m-Chlorofluorobenzene	2	p-Methylanisole
9	p-Fluorotoluene	23	m-Dibromobenzene	3	o-Chlorotoluene
10	p-Bromotoluene	24	m-Dichlorobenzene	4	o-Bromotoluene
11	p-Bromofluorobenzene	25	o-Methylanisole	5	m-Fluoroanisole
12	p-Chlorobromobenzene	26	o-Chloroanisole	6	m-Bromofluorobenzene
13	p-Chlorotoluene	27	o-Bromofluorobenzene		

Regression analysis

Some of the *58* descriptors generated for each compound encoded similar information about the molecules of interest (they were highly correlated). It was therefore desirable to test each descriptor and eliminate those with high correlation coefficients. Correlations between two descriptors can be easily obtained from the correlation matrix. When a high correlation was detected ($R > 0.95$), one or more of the descriptors were removed from consideration. By using this criterion, thirteen of the original 58 descriptors were eliminated.

Linear models were formed by a stepwise addition of terms [14]. A deletion process was then employed where each variable in the model was held out in turn and a model was generated by using the remaining descriptors. A final set of selected equations were then tested for stability and validity through a variety of statistical methods. The choice of which equation to consider further was made by using four criteria: multiple correlation coefficient (R), standard deviation (SD), F statistic and the number of descriptors in the model. An ideal model is one that has high R and F values, low standard deviation, and least number of independent variables (descriptors).

RESULTS AND DISCUSSION

A number of good models for modelling GC retention indices of the benzene derivatives given in Table I were developed by using the descriptors available. The best equation found was

$$
\begin{aligned}
z = {} & (137.114 \pm 3.359) XV_0 \\
& - (55.414 \pm 3.791) NOCH_3 \\
& + (2.321 \pm 0.344) VOL \\
& + (9.462 \pm 2.675) DIMO - 5.726 \pm 25.084
\end{aligned}
$$

$$(n = 32, R = 0.998, F = 1835, SD = 10.067) \quad (2)$$

where Z = retention index, XV,, = zero-order valence term, $NOCH_3$ = number of methyl groups in the molecule, VOL = Van der Waals volume of the molecule and $DIMO$ = dipole moment of the molecule. The variables are listed in the order in which they were selected. The high values of R and the F statistic and low standard deviation indicate that this equation represents a very good model for calculating retention indices of benzene derivatives.

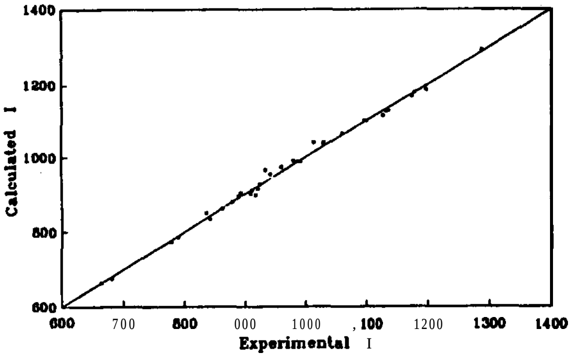

Fig. 1. Plot of calculated *versus* experimental retention indices.

The calculated and observed retention indices and structural descriptors employed in eqn. 2 are given in Table II for all the compounds studied. The plot of calculated *versus* observed retention indices is shown in Fig. 1 and reveals no deviation from linearity. Examination of the residuals (Fig. 2) indicates that they are normally distributed. The correlation matrix (Table III) for the four descriptors used in eqn. 2 shows no correlation between the parameters.

The variables in eqn. 2 encode different aspects of the molecular structures. The zero-order valence term (XV_0) is a topological descriptor

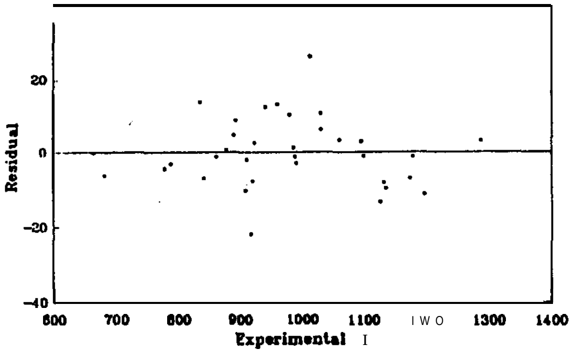

Fig. 2. Plot of residuals versus experimental retention indices.

TABLE II

EXPERIMENTAL AND CALCULATED RETENTION INDICES AND DESCRIPTORS EMPLOYED IN THE SELECTED MODEL

Compound”	Descriptor				Calculated retention index (i.u.)	Experimental retention index (i.u.)
	NOCH$_3$	*XV*$_0$	*VOL*	*DIMO*		
Training set						
1	0	3.464	88.618	0.000	674.9	681.3
2	0	3.163	93.558	1.995	664.0	664.1
3	0	4.591	102.304	1.837	878.6	877.9
4	0	5.371	105.972	1.392	989.8	979.6
5	1	4.387	105.096	0.068	784.9	788.2
6	0	4.795	113.870	1.072	926.2	923.6
7	0	5.921	127.585	2.242	1123.5	1131.7
8	2	5.309	121.642	0.001	893.7	889.2
9	1	4.086	110.077	1.960	773.1	777.7
10	1	6.294	122.492	1.384	1099.3	1096.3
11	0	5.070	110.923	0.667	953.2	940.9
12	0	6.497	119.634	0.463	1167.2	1174.4
13	1	5.513	118.808	1.826	987.8	989.2
14	0	4.494	118.855	2.353	908.6	910.6
15	0	5.921	127.506	1.096	1112.4	1126.0
16	1	5.718	130.453	1.534	1040.2	1029.6
17	2	5.309	124.332	0.061	900.5	892.0
18	0	6.497	119.596	1.585	1177.7	1179.0
19	1	6.294	122.387	1.378	1098.9	1100.0
20	1	4.086	110.058	1.998	773.5	778.0
21	1	5.513	118.779	1.827	987.7	990.9
22	0	4.290	107.228	1.873	849.1	835.4
23	0	7.278	123.271	1.315	1290.7	1287.7
24	0	5.717	115.941	1.744	1063.8	1060.5
25	1	5.718	130.495	1.459	1039.6	1013.5
26	0	5.922	127.615	2.482	1125.9	1135.6
27	0	5.070	110.900	2.743	972.8	959.6
28	1	4.086	110.085	1.952	773.1	777.4
29	0	4.494	118.910	2.674	911.8	919.7
30	2	5.309	121.534	0.073	894.2	916.2
31	0	6.497	119.609	2.490	1186.3	1197.6
32	0	4.290	107.171	3.123	860.8	862.0
Prediction set						
1	0	4.290	107.289	0.204	833.4	840.5
2	1	5.718	130.396	1.077	1035.7	1029.5
3	1	5.513	118.782	1.786	987.4	986.3
4	1	6.294	122.415	1.345	1098.7	1095.7
5	0	4.494	118.780	1.263	898.1	908.5
6	0	5.070	110.889	1.771	963.6	932.8

“The compounds are numbered as in Table I.

that encodes the size and degree of branching of the molecules. It contains corrections for the difference in the type of halogen in the molecules. The number of methyl groups in the molecule (*NOCH*$_3$) is also a topological descriptor. The Van der Waals volume *(VOL)* is a geometric descriptor. The presence of this descriptor in the model reveals the importance of

TABLE III

CORRELATION COEFFICIENTS BETWEEN THE DESCRIPTORS OF THE SELECTED MODEL

Descriptor	XV,	*$NOCH_3$*	*VOL*	*DIMO*
XV_0	1.000			
$NOCH_3$	0.003	1.000		
VOL	0.761	-0.292	1.000	
DIMO	*-0.076*	*-0.483*	*0.025*	1.000

the size of the molecules in the retention mechanism. The dipole moment of the molecules (*DIMO*) is an electronic descriptor. This is in agreement with the idea that polarity can play an important role in the retention behaviour of molecules.

In order to illustrate the predictive ability of eqn. 2, the retention indices of six compounds outside the original data set were calculated by using this model. These compounds were not included in the procedure of model generation. The predicted and experimental retention indices for these compounds are compared in Table II. Except for *m*-bromofluorobenzene, the predicted values agree well with the observed retention indices. Dewar and Rzepa [15] have shown that the MNDO method is unable to calculate the heats of formation and molecular structures of fluorine-contaning molecules. Therefore, the discrepancies obtained for these types of compounds may be due to the *VOL* and *DIMO* descriptors, which depend on the optimized structures. In general, the predicted values agree well with the observed retention indices, confirming the validity of the model.

REFERENCES

1 *SPSSJPC, The Statistical Package for IBM PC,* Quiad Software, Ontario, 1986.

2 J.H. Khorasani, *Ph.D. Thesis,* University of Salford, Salford, 1989.

3 M.N. Hasan and P.C. Jurs, *Anal. Chem., 55 (1983) 263.*

4 L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research,* Academic Press, New York, 1976.

5 K.R. Symon, *Mechanics,* Addison-Wesley, Ontario, 1971, p. 402.

6 R.N. Rohrbaugh and P.C. Jurs, *Anal. Chim. Acta,* 199 (1987) 99.

7 T.R. Stouch and P.C. Jurs, *J. Chem. Znf. Comput. Sci., 26 (1986) 26.*

8 R.S. Pearlman, *Quantum Chem.* Program *Exch. Bull.,* 1 (1981) 15.

9 L.B. Kier, *Quant. Struct. Act. Relat., 4* (1985) 109.

10 A.I. Vogel, *Elementary Practical Organic Chemistry, Part 2, Qualitative Organic Analysis,* Wiley, New York, 1966, p. 24.

11 K.J. Miller and J.A. Savchik, *J. Am. Chem. Soc.,* 101 (1979) 7206.

12 C. Hansch and A. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology,* Wiley-Interscience, New York, 1979, p. 18.

13 M.J.S. Dewar and W. Thiel, *J. Am. Chem. Soc.,* 199 *(1977) 4899.*

14 N. Draper and H. Smith, *Applied Regression Analysis,* Wiley-Interscience, New York, 2nd ed., 1981, p. 307.

15 M.J.S. Dewar and H.S. Rzepa, *J. Am. Chem. Soc.,* 100 (1978) 58.