Estimation of Kováts Retention Indices Using Group Contributions

Stephen E. Stein,* Valeri I. Babushok, Robert L. Brown, and Peter J. Linstrom

Physical and Chemical Properties Division, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899-8380

Received December 9, 2006

We have constructed a group contribution method for estimating Kováts retention indices by using observed data from a set of diverse organic compounds. Our database contains observed retention indices for over 35 000 different molecules. These were measured on capillary or packed columns with polar and nonpolar (or slightly polar) stationary phases under isothermal or nonisothermal conditions. We neglected any dependence of index values on these factors by averaging observations. Using 84 groups, we determined two sets of increment values, one for nonpolar and the other for polar column data. For nonpolar column data, the median absolute prediction error was 46 (3.2%). For data on polar columns, the median absolute error was 65 (3.9%). While accuracy is insufficient for identification based solely on retention, it is suitable for the rejection of certain classes of false identifications made by gas chromatography/mass spectrometry.

INTRODUCTION

The method of gas chromatography/mass spectrometry (GC/MS) is routinely used to identify volatile compounds by matching acquired spectra with spectra in a reference library. Such mass spectral reference libraries have been developed for many years and are often integral parts of instrument data systems. While it is evident that matching against reference retention values in addition to reference mass spectra would increase the reliability of compound identification, such reference values are not available for a large proportion of compounds in modern comprehensive reference libraries, though some specialized libraries contain this information. For example, only 9% of the molecules in the recent release of the NIST/EPA/NIH mass spectral reference library[1] have observed retention index (RI) data. In an effort to provide relevant data, we have undertaken a program to build a comprehensive, structure-based collection[2] of RIs and to develop methods for estimating indices of compounds for which RI measurements are not available.

The retention of molecules on stationary phases involves a variety of intermolecular forces.[3] Because of this complexity, no general prediction method founded on molecular properties is currently feasible. On the other hand, owing to the widespread interest in retention values for compound identification, many empirical procedures for estimating RI values from molecular structure or correlations with other molecular properties have been suggested.[4−14] In general, these contain a number of adjustable parameters fitted to measured RI values of a limited number of compounds related to a specific research problem. The most versatile methods are based on additive contributions from specific groups of atoms in a molecule. Since atomic forces extend over short distances, it is often reasonable to assume that only nearest neighbors will affect a particular atom or small group of atoms. Although suitable primarily for estimating properties of isolated molecules such as heats of formation, this approach has also been used to predict many properties that involve intermolecular forces.[15]

Here, we suggest a simple linear group increment model for RI estimation. It uses measured RI values to determine increment values for 84 different groups. The approach is similar to that used earlier by us[16] to predict normal boiling points. We have found this method useful for identifying possible errors while building a retention index collection based on RI literature data. These predictions are also of direct use for enhancing confidence in compound identifications made by matching spectra to those in a comprehensive reference library, such as the AMDIS system.[17] It is important to note that this approach is very approximate. At present, we can only estimate prediction errors for classes of compounds, not for individual compounds. However, we do show histograms of the observed errors for all compounds in our database as well as the observed average errors for a variety of compound classes.

DETERMINATION OF GROUP INCREMENT VALUES

For this work, we used the same groups as employed for boiling point estimations.[16] We assumed that the retention index of a molecule could be calculated using a linear model

$$\mathrm{RI} = \sum_n f_n g_n + h \qquad (1)$$

where f_n is the number of times group n appears in the molecule; g_n is its increment value. We calculated group increment values by fitting observed retention indices to this equation via minimization of absolute deviations. To correct for any uniform prediction error, we included an adjustable parameter h. Its value was chosen to be that which made the median deviation (MED) zero.

Nonpolar Column Data. Our database contains RI values for 25 728 distinct molecules. Of these, 25 296 could be represented by our current set of groups. These values were measured on columns having either nonpolar or slightly polar

* Corresponding author e-mail: steve.stein@nist.gov.

10.1021/ci600548y CCC: $37.00
Published on Web 03/17/2007

Table 1. GC Stationary Phases Represented in NIST RI Database

type of phase	phase	examples of trade specification
non-polar	dimethylpolysiloxane	OV-101, HP-1, DB-1, SE-30
	methyl silicone, 5% phenyl groups	DB-5, SE-54, HP-5, Ultra-2
	apiezons (L, M, N)	
	squalane	
	apolan (branched hydrocarbons, $C_{87}H_{176}$)	
polar	polyethylene glycol, acid- and base-modified polyethylene glycols	Carbowax 20M, Innowax, CP-Wax 52 CB

stationary phases (see Table 1). Within each phase type, we averaged any multiple measurements for a particular compound (see below) by choosing their median value. A set of 4472 compounds had median RI values measured on both phase types. For these, the average absolute difference (AAD) between the medians was 25 and the average deviation (AD), nonpolar − slightly polar, was −11. Since this difference was small, we avoided deriving two sets of increment values by ignoring any such phase differences and simply averaged medians. The database contained three types of RI measurements: isothermal Kováts indices,[18] nonisothermal Kováts indices (from temperature-programming, using the definition of Van den Dool and Kratz[19]), and Lee indices[20] (LIs; isothermal and nonisothermal). The last type was converted to Kováts indices. For conversion of isothermal retention indices, the earlier derived relationship[21] was used. Nonisothermal LIs were converted using the correlation

$$\mathrm{RI} = 127.7 + 4.5269 \times \mathrm{LI} + 2.6193 \times 10^{-3} \times \mathrm{LI}^2 + 5.00 \times 10^{-7} \times \mathrm{LI}^3$$

The correlation was derived using database values of Kováts indices for reference compounds of Lee scale neglecting temperature dependencies and differences in RI values for nonpolar stationary phases considered in this work.

Three types of RI measurements were treated as equivalent, and temperature dependencies were ignored. Most indices (85%) were measured on capillary columns. No distinction was made between capillary and packed columns. As a result of all these approximations, our RI values are in effect averages over typical ranges of temperature, column type, and stationary phase. Some idea of the errors introduced by these approximations may be obtained from a subset of molecules each having more than one observed RI value. There were 9975 such compounds having an average of slightly over 10 observations each. For these, the AAD of their median RI values was 11 (0.9%). Figure 1 is a histogram of the percent deviations of multiple RI observations from their median values.

In preliminary work, we observed that the error distribution between observed RI and those predicted by eq 1 could be approximately represented by a two-sided exponential function $y = y_0 \exp(-|x|/s)$ and $y_0 = N/(2s)$, where s is the observed AAD and N is the total number of molecules in the data set. For this distribution, the maximum likelihood estimator is calculated by varying the increment values g_k so as to minimize the sum of the absolute errors.[22] As a first approximation to the increment values, we minimized the sum of the errors squared (LSQ) and then used these as starting values for the absolute error minimization calculation (ABS).[23]

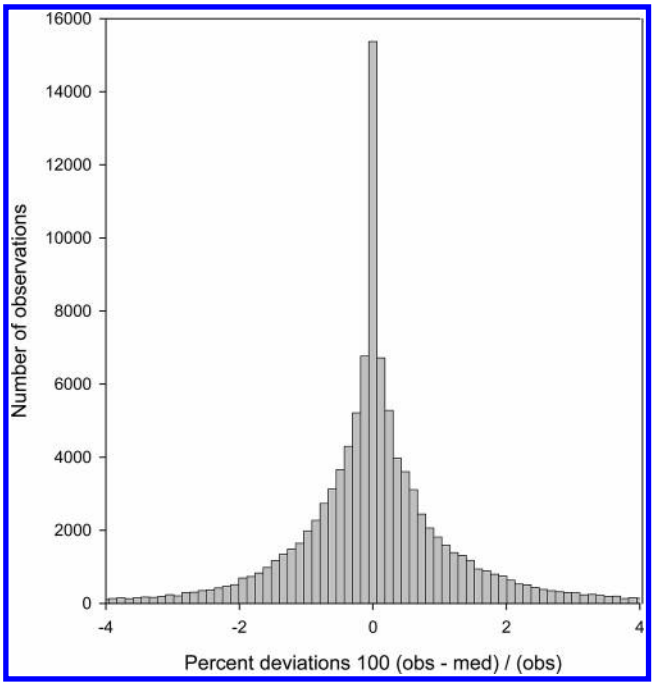

Figure 1. Percent deviations of multiple RI measurements for the same molecule with nonpolar columns. There are 9975 individual compounds represented with a total of 105 230 observations averaging approximately 10 observations per compound.

A fit using the data from the complete set of 25 296 compounds gave poor RI estimates for some compounds. These fell roughly into three classes: highly fluorinated compounds, cyclic siloxanes, and large or complex ring systems (steroids, adamantane analogs, and polynuclear aromatic compounds). We employed simple, convenient criteria to identify such molecules. Silicon- and fluorine-containing compounds are tagged if the number of either atom is greater than one-fifth the total number of all atoms in the molecule. For ring systems, we use a ring perception algorithm that calculates the total possible number of rings (for example, it returns six for anthracene). If this number is more than a factor of 2.5 times the minimum number of rings (three for anthracene), we flag the predicted RI as a likely poor estimate. These screening methods selected 2301 such compounds. Exclusion of these left a total of 22 995 that we used to fit the group increment parameters.

Table 2 contains parameter values from the ABS fit. With these increments, the AAD for the whole data set was 70 (4.4%) and the median absolute deviation (MAD) was 45 (3.1%). A value of $h = 1.9$ was used for the adjustable parameter (eq 1). A histogram of these errors is shown in Figure 2. In addition, we also list in Table 2 the number of database compounds from the 22 995-compound set in which a particular group appears. Increment values g_n for groups that appear in few compounds will be more sensitive to database modifications than those groups appearing in many. To demonstrate this, we used the bootstrap method[24,25] to generate 100 synthetic data sets. Each set was generated by randomly selecting *with replacement* 22 995 data points from the original set. Fits of each set yielded somewhat different increment values. Table 2 shows these variations in g_n as average absolute deviations e_n taken over the 100 samples. Group increment variations are roughly proportional to the square root of the reciprocal of the number of compounds containing the group.

Table 2. Group Increment Contributions to Kováts Retention Indices. Nonpolar Stationary Phase Data[a]

group	*n*	*g*	*e*
Hydrocarbon Increments			
(1) −CH_3	19 726	112	1
(2) >CH_2	14 920	99	0
(3) >C*r*H_2	5080	121	1
(4) >CH−	6938	22	2
(5) >C*r*H−	4334	69	1
(6) >C<	2254	−14	3
(7) >C*r*<	1619	32	2
(8) =CH_2	1850	98	2
(9) =CH−	4082	102	1
(10) =C*r*H−	2371	110	1
(11) =C<	1886	67	2
(12) =C*r*<	2573	90	2
(13) *aa*CH	9081	114	1
(14) *aa*C−	9167	114	2
(15) *aaa*C	574	161	1
(16) ≡CH	207	101	5
(17) ≡C−	628	106	2
Oxygen Increments			
(18) −OH	71	106	13
(19) 1−OH	645	255	3
(20) 2-OH	693	239	3
(21) 3−OH	409	189	7
(22) ϕ−OH	492	221	5
(23) −O−	4360	75	1
(24) −O*r*−	2235	112	3
(25) −O−OH	10	372	53
Carboxyl Increments			
(26) −HCO	598	299	3
(27) >CO	992	235	3
(28) >C*r*O	730	291	4
Carboxyl Increments			
(29) −CO−O−	6389	266	2
(30) −C*r*O−O*r*−	266	465	9
(31) −CO−OH	188	461	30
(32) −CO−NH_2	41	514	24
(33) −CO−NH−	1034	497	4
(34) −C*r*O−N*r*H−	222	424	7
(35) −CO−N<	710	286	4
(36) −C*r*O−N*r*<	244	397	10
Nitrogen Increments			
(37) −NH_2	185	254	6
(38) ϕ−NH_2	175	303	6
(39) >NH	706	198	4
(40) >N*r*H	218	268	7
(41) >N−	738	38	4
(42) >N*r*−	931	118	4
(43) >N−NO	13	466	15
(44) >N*r*−NO	5	564	30
(45) *aa*N	887	105	2
(46) =NH	4	−9	103
(47) =N−	787	209	5
(48) =N*r*−	333	116	6
(49) =N*r*−N*r*H−	21	496	34
(50) −N*r*=C*r*R−N*r*R−	129	552	12
(51) −N*r*=C*r*R−N*r*H−	18	566	11
(52) −N*r*=C*r*H−N*r*R−	154	450	22
(53) −N*r*=C*r*H−N*r*H−	16	767	43
(54) −N=N−	25	167	10
(55) −N=C=S	37	571	10
(56) −NO	6	123	32
(57) −NO_2	532	393	5
(58) −CN	169	354	8
(59) ϕ−CN	124	276	7
Halogen Increments			
(60) −F	506	−12	3
(61) ϕ−F	539	−24	2
(62) −Cl	212	189	3
(63) 1−Cl	616	236	2
(64) 2−Cl	614	217	2
(65) 3−Cl	207	172	2
(66) ϕ−Cl	1740	179	2
(67) −Br	433	306	3
(68) ϕ−Br	619	320	3
(69) −I	92	425	4
(70) ϕ−I	86	400	6
Sulfur Increments			
(71) −SH	168	316	4
(72) ϕ−SH	4	317	23
(73)−S−	1029	251	1
(74) −S*r*−	843	263	2
(75) >SO	11	458	54
(76) >SO_2	53	506	24
(77) >CS	34	480	20
(78) >C*r*S	10	436	59
Silicon Increments			
(79) >SiH−	33	39	8
(80) >Si<	2578	−115	3
(81) >Si*r*<	192	−128	8
Phosphorus Increments			
(82) >P−	3	98	39
(83) >PO−	179	246	8
(84) >PS−	51	244	25

[a] *n* is the number of database compounds in which a particular group appears. *g* is the retention index increment value for particular group. *e* is the average absolute deviation from the median increment values obtained from fits of 100 random samples drawn with replacement from the nonpolar RI database. The symbol *a* denotes an aromatic bond. Atoms having the postscript *r* are in rings. Symbols > and < denote two single bonds. The symbol ϕ denotes an aromatic system.

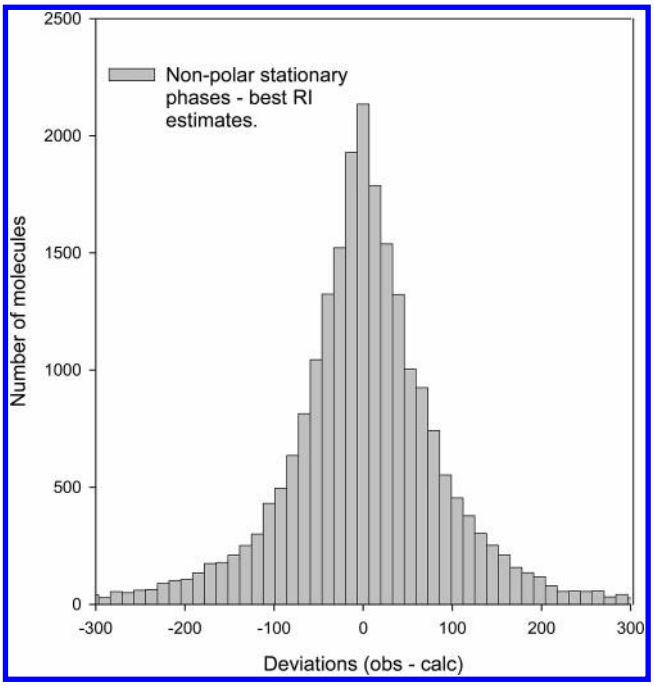

Figure 2. Prediction errors for RI measured on nonpolar stationary phase columns. Data set consists of compounds for which the model gives best estimates. Number of molecules was 22 995.

A comparison of some of these increment values with those reported in the literature is shown in Table 3. It demonstrates a reasonable agreement of calculated increments with increment values available in the literature. Note that literature values were determined mostly from the substitution effect based on small limited sets of compounds. Thus, Voelkel[26] determined group increments based on the linear relationship of RI on the number of oxyethylene units for nonionic surfactants with the general formula RX-$(CH_2H_2O)_n$R′. Cook and Raushel[28] estimated increment values from the RI of monosubstituted benzene derivatives assuming that the retention index consists of two additive

Table 3. Comparison of Calculated Increments with Literature Data. Non-Polar Phase.

group	this work	SE-30[26] [a]	SE-30[27] [b]	SE-30[28] [c]	OV-101[29] [d]
−SH	316	329			
−NH_2	254	211	244	297−303	
−O−	75	92			
−S−	251	330			
>NH	198	155			
−CH_3	112	100	68	103−104	90
−CH_2	99	100			
−I	400−425			367−390	
−Br	306−320			267−282	
−Cl	172−236			186−194	151−159
−CN	354			303−313	
−CHO	299			284−294	
−NO_2	393			403−421	

[a] Isothermal data, 443 K. [b] Isothermal data, 200 C. [c] Isothermal data, 100−160 C. [d] Temperature-programmed data.

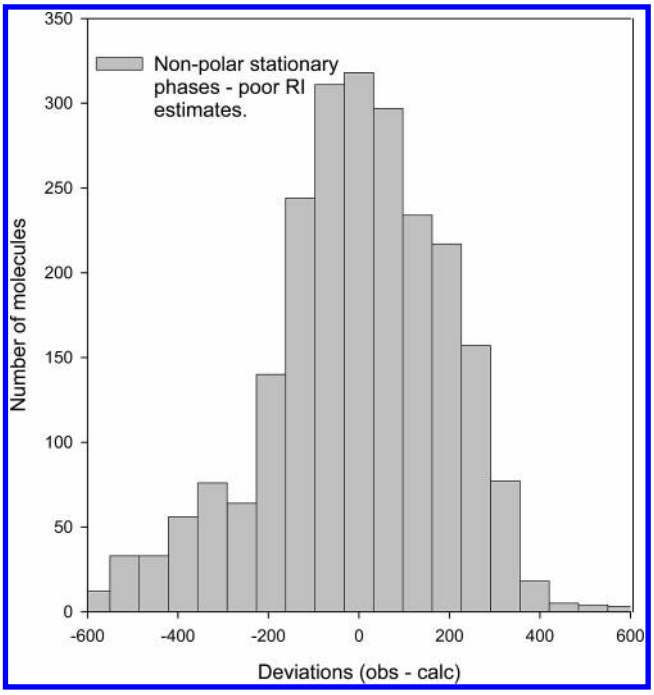

Figure 3. Prediction errors for RI measured on nonpolar stationary phase columns. Data set consists of compounds whose RI values are poorly predicted by the model. Number of molecules was 2301.

parts: the aromatic ring contribution and the substituent contribution.

For the set of 2301 compounds with poorly predicted RIs, the group increment values listed in Table 2 produced a large underprediction of RI values. A large value of the correction term $h = 150$ was required to reduce the MED to zero. The AAD value for the prediction error was 174 (8.3%) with a MAD value of 127 (5.3%) for the set. Figure 3 contains a histogram of these deviations.

Polar Column Data. Our polar column data set contained measured RI values for 9100 distinct molecules. Of these, 8769 could be predicted with our current set of group increment values. We ignored any temperature dependencies and assumed that all polar phases in the database (Table 1) were equivalent. Preliminary calculations showed that our model gave poor predictions for the same types of molecules as in the nonpolar case. There were 206 such compounds. We used the observed RI values from the remaining 8563 compounds to calculate a set of group increment values. Increments derived from an ABS fit of eq 1 appear in Table 4. With these values, and using $h = 2.6$, we obtained an AAD for these molecules of 101 (5.7%) and a MAD of 64 (3.9%). Figure 4 contains a histogram of these errors. As for the nonpolar data, we also did a bootstrap simulation using 100 samples taken from the 8563-compound data set. Results are shown in Table 4 as average absolute deviations e_n in the group increments.

Increment values (Table 4) underpredicted the RI values for the set of 206 poorly predicted compounds. To correct this, we used $h = 117$ for the correction term. As noted above, its value was chosen to make MED zero. For this set of compounds, our RI predictions had an AAD of 101 (6.1%)

Table 4. Group Increment Contributions to Kováts Retention Indices. Polar Stationary Phase Data[a]

group	*n*	*g*	*e*
hydrocarbon increments			
(1) −CH_3	7789	113	3
(2) >CH_2	5583	99	0
(3) >C*r*H_2	2305	128	2
(4) >CH−	2581	6	4
(5) >C*r*H−	1972	96	3
(6) >C<	518	−65	10
(7) >C*r*<	1141	39	5
(8) =CH_2	1155	125	4
(9) =CH−	2150	133	1
(10) =C*r*H−	1345	159	2
(11) =C<	853	91	5
(12) =C*r*<	1493	122	4
(13) *aa*CH	1819	166	2
(14) *aa*C−	1849	145	4
(15) *aaa*C	60	253	7
(16) ≡CH	45	207	14
(17) ≡C−	102	219	4
Oxygen Increments			
(18) −OH	46	397	34
(19) 1−OH	446	747	5
(20) 2-OH	519	645	5
(21) 3−OH	343	561	10
(22) ϕ−OH	184	715	21
(23) −O−	647	180	6
(24) −O*r*−	699	202	5
(25) −O−OH			
Carboxyl Increments			
(26) −HCO	474	602	5
(27) >CO	559	524	5
(28) >C*r*O	319	626	11
carboxyl increments			
(29) −CO−O−	2369	515	3
(30) −C*r*O−O*r*−	168	1126	15
(31) −CO−OH	119	1383	15
(32) −CO−NH_2	16	1606	77
(33) −CO−NH−	18	1100	91
(34) −C*r*O−N*r*H−	14	1325	60
(35) −CO−N<	24	1047	28
(36) −C*r*O−N*r*<	9	1112	97
Nitrogen Increments			
(37) −NH_2	75	511	10
(38) ϕ−NH_2	39	744	26
(39) >NH	126	378	14
(40) >N*r*H	103	506	27
(41) >N−	73	128	15
(42) >N*r*−	152	196	17
(43) >N−NO	1	901	4
(44) >N*r*−NO			
(45) *aa*N	372	201	5
(46) =NH	2	670	19
(47) =N−	15	281	73
(48) =N*r*−	134	227	13
(49) =N*r*−N*r*H−	7	1178	102
(50) −N*r*=C*r*R−N*r*R−	1	470	16
(51) −N*r*=C*r*R−N*r*H−	1	1761	5
(52) −N*r*=C*r*H−N*r*R−	3	919	230
(53) −N*r*=C*r*H−N*r*H−	2	1855	21
(54) −N=N−			
(55) −N=C=S	22	981	22
(56) −NO	14	206	22
(57) −NO_2	90	577	16
(58) −CN	41	781	29
(59) ϕ−CN	14	627	48
halogen increments			
(60) −F	13	−29	8
(61) ϕ−F	4	−59	27
(62) −Cl	60	261	11
(63) 1−Cl	308	456	3
(64) 2−Cl	570	396	4
(65) 3−Cl	126	239	4
(66) ϕ−Cl	228	230	7
(67) −Br	105	526	8
(68) ϕ−Br	17	391	19
(69) −I	36	685	5
(70) ϕ−I	7	608	10
Sulfur Increments			
(71) −SH	113	561	11
(72) ϕ−SH	3	642	120
(73)−S−	304	395	5
(74) −S*r*−	243	447	6
(75) >SO	2	720	200
(76) >SO_2	4	1107	176
(77) >CS	3	1149	191
(78) >C*r*S	2	785	110
Silicon Increments			
(79) >SiH−			
(80) >Si<	130	−308	7
(81) >Si*r*<			
Phosphorus Increments			
(82) >P−			
(83) >PO−	43	288	29
(84) >PS−			

[a] *n* is the number of database compounds in which a particular group appears. *g* is the retention index increment value for particular group. *e* is the average absolute deviation from the median increment values obtained from fits of 100 random samples drawn with replacement from the nonpolar RI database. The symbol *a* denotes an aromatic bond. Atoms having the postscript *r* are in rings. Symbols > and < denote two single bonds. The symbol ϕ denotes an aromatic system.

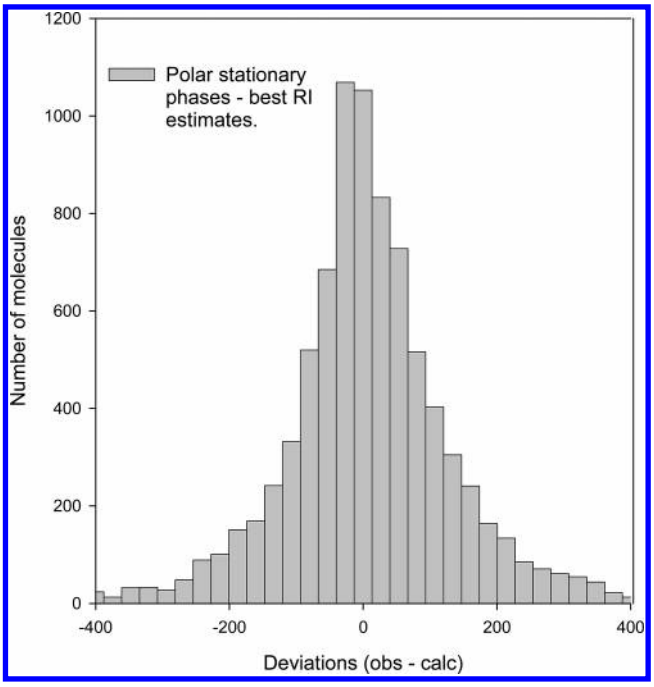

Figure 4. Prediction errors for RI measured on polar stationary phase columns. Data set consists of compounds for which the model gives best estimates. Number of molecules was 8563.

Table 5. Summary of RI Prediction Errors[a]

column	nonpolar		polar	
data set	G	P	G	P
n	22955	2301	8563	206
h	1.9	150	2.6	117
AD	−2.1	−21	5.7	8.6
SD	107	257	154	139
AAD	70 (4.4)	174 (8.3)	101 (5.7)	101 (6.1)
MAD	45 (3.1)	127 (5.3)	64 (3.9)	76 (4.7)

[a] Column: Type of column stationary phase. Data set: Molecules having G (good) and P (poor) RI predictions. *n*: Number molecules in the data set. *h*: Correction for underprediction of RI values. Its value is adjusted to make the median deviation zero. AD: Average deviation. SD: Standard deviation. AAD: Average absolute deviation (%). MAD: Median absolute deviation (%).

and a MAD of 76 (4.7%) for the set. Table 5 contains a summary of prediction errors for both nonpolar and polar data.

DISCUSSION AND CONCLUSIONS

The method we have used in this work is essentially a curve-fitting procedure for interpolating between values of some complex multivariable function. Although using the same numerical methods, it is unrelated to normal regression theory. Errors arise primarily from the inadequacy of the linear interpolation function rather than from measurement errors in the observed database RI.[25] As a result, it is not possible to estimate errors in linear combinations of group increments with standard statistical methods. This kind of error has been called "method error".[15] Method errors are commonly reported as the average prediction errors for specific types of compounds. Tables 6 and 7 contain observed AADs and MADs for several classes of compounds from our nonpolar and polar data sets, respectively. Groups that may be included in a particular class are identified (in parenthesis following each class name) by their group numbers as denoted in Tables 2 and 4. Asterisks denote required groups. For example, to be included in the class "aromatic HCs" a molecule may contain *any* of the groups in the range 1−12 or 16 or 17 but *must* contain at least one group from the range 13−15. The multifunctional class contains molecules not assigned to any of the listed classes.

The prediction accuracy we have observed in this work appears consistent with what one might expect from a simple group method. One can extend the method to additional types of compounds by devising additional groups. For some types, simple rearrangements or combinations of existing groups can increase prediction accuracy. One obvious limitation of the model is its failure to distinguish isomers. Molecules with the same groups, no matter how arranged, will have the same

Table 6. Observed Average Kováts Retention Index Prediction Errors for Different Classes of Compounds—Nonpolar Stationary Phases[a]

class (groups in class)	size	AAD	MAD
aliphatic HCs (1−12, 16,17)	3185	39 (3.3)	26 (2.2)
aromatic HCs (1−12, 13*−15*, 16,17)	603	57 (3.9)	45 (3.5)
alcohols (1−17, 18*−21*)	898	41 (3.2)	27 (2.3)
phenols (1−17, 22*)	104	68 (4.5)	43 (3.5)
aldehydes (1−17, 26*)	337	46 (3.6)	31 (2.7)
ketones (1−17, 27*, 28*)	467	56 (4.5)	40 (3.4)
esters (1−17, 29*, 30*)	1935	46 (3.0)	29 (2.0)
ethers (1−17, 23*, 24*)	612	66 (5.6)	44 (3.8)
carboxylic acids (1−17, 31*)	82	51 (4.3)	33 (2.4)
nitrogen (1−17, 37*−59*)	1339	70 (5.1)	56 (3.9)
sulfur (1−17, 71*−78*)	929	46 (3.7)	34 (3.0)
fluorides (1−17, 60*,61*)	24	33 (3.3)	30 (1.8)
chlorides (1−17, 62*−66*)	566	71 (4.6)	50 (3.9)
bromides (1−17, 67*, 68*)	101	61 (5.7)	50 (4.6)
iodides (1−17, 69*, 70*)	70	45 (4.4)	37 (3.3)
multifunctional	11 743	88 (4.4)	62 (3.3)
all molecules	22 995	70 (4.4)	45 (3.2)

[a] Size is the number of molecules in a particular class. AAD: Average absolute deviations. MAD: Median absolute deviations. Numbers in parentheses are percent AADs and MADs. Note that the set of compounds with poorly predicted RI values are not included in this table.

Table 7. Observed Average Kováts Retention Index Prediction Errors for Different Classes of Compounds—Polar Stationary Phases[a]

class (groups in class)	size	AAD	MAD
aliphatic HCs (1−12, 16,17)	994	71 (4.9)	47 (3.7)
aromatic HCs (1−12, 13*−15*, 16,17)	311	80 (4.7)	66 (4.4)
alcohols (1−17, 18*−21*)	854	72 (3.8)	46 (2.6)
phenols (1−17, 22*)	50	275 (13)	233 (11)
aldehydes (1−17, 26*)	324	70 (4.0)	44 (2.8)
ketones (1−17, 27*, 28*)	430	106 (6.5)	76 (4.6)
esters (1−17, 29*, 30*)	1209	60 (3.4)	38 (2.2)
ethers (1−17, 23*, 24*)	453	127 (8.1)	98 (6.5)
carboxylic acids (1−17, 31*)	96	74 (3.1)	45 (1.9)
nitrogen (1−17, 37*−59*)	554	134 (8.2)	98 (6.4)
sulfur (1−17, 71*−78*)	248	99 (6.6)	72 (4.5)
fluorides (1−17, 60*, 61*)	2	172 (19)	172 (18)
chlorides (1−17, 62*−66*)	160	62 (5.0)	43 (3.3)
bromides (1−17, 67*, 68*)	54	53 (3.8)	43 (3.7)
iodides (1−17, 69*, 70*)	42	28 (2.0)	23 (1.8)
multifunctional	2780	134 (2.0)	98 (1.8)
all molecules	8563	101 (5.7)	64 (3.9)

[a] Size is the number of molecules in the class. AAD: Average absolute deviations. MAD: Median absolute deviations. Numbers in parentheses are percent AADs and MADs. The set of compounds with poorly predicted RI values are not included in this table.

predicted RI. For example, our database contains 38 isomers of tetrabromo-dibenzofuran with RI values ranging from 2688 to 2852. Since all have the same groups, our model predicts an RI of 2763 for each. The 22 913 compounds in our nonpolar data set contained 3792 such isomer sets that include 13 503 compounds. Isomer sets contained an average of 3.6 molecules per set with an average AAD of 22 per set. This represents a significant error. Furthermore, it reduces the diversity of our nonpolar database since the model in effect distinguishes only 13 202 structures. Similarly, 7273 structures were distinguishable in the polar data set containing 8255 different molecules. To predict RI differences between isomers, additional adjustable parameters are required. For example, a study[30] using a linear model required a set of 18 parameters to accurately predict RI values for all 135 polychlorinated dibenzofurans. One might imagine building a very general and accurate prediction scheme that incorporates many such special methods. For a given molecule, this would require automatic selection of the best particular method from what would probably be a very large database of procedures and parameters. This could be routinely updated as new correlations become available.

While the accuracy of the present general method is clearly insufficient to distinguish compounds with comparable retention by estimated retention properties alone, it is well-suited for the rejection of many false-positive results appearing in "hit lists" from mass spectrometry library searches.[31,32] Since this means of identification is based only on fragment peaks, and it is not uncommon for high-molecular-weight compounds to generate primarily low-molecular-weight fragments, similar spectra can result from different high- and low-molecular-weight compounds. The use of even approximate retention indices can reliably eliminate such false identifications from automated library search systems. Further, these estimates can generally assist analysts in verifying identifications as long as the degree of prediction errors is understood.

REFERENCES AND NOTES

(1) Ausloos, P.; Clifton, C. L.; Lias, S. G.; Mikaya, A. I.; Stein, S. E.; Tchekhovskoi, D. V.; Sparkman, O. D.; Zaikin, V.; Zhu, D. The Crtitical Evaluation of a Comprehensive Mass Spectral Library. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 287−299 (NIST Standard Reference Database 1A; NIST/EPA/NIH Mass Spectral Library with Search Program; Data Version: NIST 05, 2005).

(2) Babushok, V. I.; Linstrom, P. J.; Reed, J. J.; Zenkevich, I. G.; Brown, R. L.; Mallard, W. G.; Stein, S. E. Development of Database on Gas-Chromatographic Retention Properties of Organic Compounds. In preparation.

(3) Peng, C. T. Prediction of Retention Indices. V. Influence of Electronic Effects and Column Polarity on Retention Index. *J. Chromatogr., A* **2000**, *903*, 117−143.

(4) Budahegyi, M. V.; Lombosi, E. R.; Lombosi, T. S.; Meszaros, S. Y.; Nyiredy, Sz.; Tarjan, G.; Timar, I.; Takacs, J. M. 25th Anniversary of the Retention Index System in Gas−Liquid Chromatography. *J. Chromatogr.* **1983**, *271*, 213−307.

(5) Tarjan, G.; Nyiredy, Sz.; Gyor, M.; Lombosi, E. R.; Lombosi, T. S.; Budahegyi, M. V.; Meszaros, S. Y.; Takacs, J. M. Thirtieth Anniversary of the Retention Index According to Kovats in Gas−Liquid Chromatography. *J. Chromatogr.* **1989**, *472*, 1−92.

(6) Evans, M. B.; Haken, J. K. Recent Developments in the Gas-Chromatographic Retention Index Scheme. *J. Chromatogr.* **1989**, *472*, 93−127.

(7) Buryan, P.; Nabivach, V. M.; Dmitrikov, V. P. Structure Retention Correlations of Isomeric Alkylphenols in Gas−Liquid Chromatography. *J. Chromatogr.* **1990**, *509*, 3−14.

(8) Jalali-Heravi, M.; Fatemi, M. H. Artificial Neural Network Modeling of Kovats Retention Indices for Noncyclic and Monocyclic Terpenes. *J. Chromatogr., A* **2001**, *915*, 177−183.

(9) Randic, M.; Basak, S. C.; Pompe, M.; Novic, M. Prediction of Gas Chromatographic Retention Indices Using Variable Connectivity Index. *Acta Chim. Slov.* **2001**, *48*, 169−180.

(10) Price, G. J.; Dent, M. R. Prediction of Retention in Gas−Liquid-Chromatography using the UNIFAC Group Contribution Method. 2. Polymer Stationary Phases. *J. Chromatogr.* **1991**, *585*, 83−92.

(11) Peng, C. T.; Ding, S. F.; Hua, R. L.; Yang, Z. C. Prediction of Retention Indexes. 1. Structure Retention Index Relationship on Apolar Columns. *J. Chromatogr.* **1988**, *436*, 137−172.

(12) Kaliszan, R. *Structure and Retention in Chromatography. A Chemometric Approach*; Harwood: Amsterdam, The Netherlands, 1997.

(13) Tulasamma, P.; Reddy, K. S. Quantitative Structure and Retention Relationships for Gas Chromatographic Data: Application to Alkyl Pyridines on Apolar and Polar Phases. *J. Mol. Graphics Modell.* **2006**, *25*, 507−513.

(14) Skrbic, B.; Onjia, A. Prediction of the Lee Retention Indices of Polycyclic Aromatic Hydrocarbons by Artificial Neural Network. *J. Chromatogr., A* **2006**, *1108*, 279−294.

(15) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; American Chemical Society: Washington, DC, 1990.

(16) Stein, S. E.; Brown, R. L. Estimation of Normal Boiling Points from Group Contributions. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 581−587.

(17) **A**utomatic **M**ass Spectral **D**econvolution and **I**dentification **S**ystem (AMDIS). This is a NIST-maintained computer program which allows one to automatically locate any of a set of target compounds in a GC/MS data library. Stein, S. E. An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 770−781.

(18) Kováts, E. Gas-Chromatographische Charakterisierung Organischer Verbindungen. 1. *Helv. Chim. Acta* **1958**, *41*, 1915−1932.

(19) Van den Dool, H.; Kratz, P. D. A Generalization of Retention Index System including Linear Temperature Programmed Gas Liquid Partition Chromatography. *J. Chromatogr.* **1963**, *11*, 463−471.

(20) Lee, M. L.; Vassilaros, D. L.; White, C. M.; Novotny, M. Retention Indexes for Programmed-Temperature Capillary-Column Gas-Chromatography of Polycyclic Aromatic Hydrocarbons. *Anal. Chem.* **1979**, *51* (6), 768−773.

(21) Babushok, V. I.; Linstrom, P. J. On the Relationship between Kovats and Lee Retention Indices. *Chromatographia* **2004**, *60*, 725−728.

(22) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed; Cambridge University Press: New York, 1992; p 701.

(23) We used *singular value decomposition* for the LSQ and the *downhill simplex method* for the ABS minimization calculations. See ref 18 for details.

(24) Flury, B. *A First Course in Multivariate Statistics*; Springer-Verlag: New York, 1997.

(25) Mood, A. M. *Introduction to the Theory of Statistics*; McGraw-Hill: New York, 1950; p 309.

(26) Voelkel, A. Retention Indices and Thermodynamic Functions of Solution for Model Non-Ionic Surfactants in Standard Stationary Phases Determined by Gas Chromatography. *J. Chromatogr.* **1987**, *387*, 95−104.

(27) Rybkina, T. I.; Kirichenko, E. A.; Kochetov, V. A. Identification of Nitrogen-Containing Organosilicon Compounds in Gas Chromatography. *Zh. Anal. Khem.* **1989**, *44*, 88−91.

(28) Cook, L. E.; Raushel, F. M. Calculation of Retention Indices for Benzene and Benzene Derivatives on the Basis of Molecular Structure. *J. Chromatogr.* **1972**, *65*, 556−559.

(29) Mihara, S.; Masuda, H. Correlation between Molecular Structures and Retention Indices of Pyrazines. *J. Chromatogr.* **1987**, *402*, 309−317.

(30) Liang, X.; Wang, W.; Schramm, K. W.; Zhang, Q.; Oxynos, K.; Henkelmann, B.; Kettrup, A. A New Method of Predicting of Gas Chromatographic Retention Indices for Polichlorinated Dibenzofurans (PCDFs). *Chemosphere* **2000**, *41*, 1889−1895.

(31) Eckel, W. P. Making Sense of Nontarget Compound Data from GC-MS Library Searches. *Am. Lab.* **2000**, *32* (6), 17−18.

(32) Eckel, W. P.; Kind, T. Use of Boiling Point−Lee Retention Index Correlation for Rapid Review of Gas Chromatography−Mass Spectrometry Data. *Anal. Chim. Acta* **2003**, *494*, 235−243.

CI600548Y