Comparison of Ridge Regression, Partial Least-Squares, Pairwise Correlation, Forward- and Best Subset Selection Methods for Prediction of Retention Indices for Aliphatic Alcohols

Orsolya Farkas* and Károly Héberger

Institute of Chemistry, Chemical Research Center, Hungarian Academy of Sciences, H-1525 Budapest, P.O. Box 17, Hungary

Received May 25, 2004

A quantitative structure−retention relationship (QSRR) study based on multiple linear regression (MLR) was performed for the description and prediction of Kováts retention indices (RI) of alcohol compounds. Alcohols were of saturated, linear or branched types and contained a hydroxyl group on the primary, secondary or tertiary carbon atoms. Constitutive and weighted holistic invariant molecular (WHIM) descriptors were used to represent the structure of alcohols in the MLR models. Before the model building, five variable selection methods were applied to select the most relevant variables from a large set of descriptors, respectively. The selected molecular properties were included into the MLR models. The efficiency of the variable selection methods was also compared. The selection methods were as follows: ridge regression (RR), partial least-squares method (PLS), pair−correlation method (PCM), forward selection (FS) and best subset selection (BSS). The stability and the validity of the MLR models were tested by a cross-validation technique using a leave-n-out technique. Neither RR nor PLS selected variables were able to describe the Kováts retention index properly, and PCM gave reliable results in the description but not for prediction. We built models with good predicting ability using FS and BSS as a selection method. The most relevant variables in the description and prediction of RIs were the mean electrotopological state index, the molecular mass, and WHIM indices characterizing size and shape.

1. INTRODUCTION

Quantitative structure−retention relationships (QSRR) have been used to find correlation between chromatographic data and different properties of the analytes.[1] Using these methods, it is possible to better understand the retention mechanisms of different classes of compounds and to predict their retention properties (e.g. Kováts retention indices). The nature and the number of available experimental and calculated descriptors to characterize the molecular structure are numerous.[2] Physicochemical properties, e.g., boiling point, molar volume, molar refraction, density, etc. have been used as descriptors for a long time.[3−5] The descriptors indicate the electronic, steric and hydrophobic properties of the compounds. Topological descriptors represent the two-dimensional structures of the molecules and are also widely used in QSRR studies.[6]

Weighted holistic invariant molecular (WHIM) descriptors were developed by the Chemometrics and QSAR Research Group of R. Todeschini in Milan.[7] WHIM descriptors are three-dimensional descriptors based on principal component analysis of the weighted covariance matrix obtained from the Cartesian coordinates of the molecule. They have been successfully applied in toxicity studies of chlorobenzenes[8] and of aromatic hydrocarbons[9] as well as for sorption studies of nonionic pesticides.[10] WHIM descriptors and lipophilicity were important to predict tumoricidal activity and accumulation of photosensitizers used in photodynamic therapy.[11] Similarly, anti-HIV activity was predicted from the molecular structure of diverse tetrapyrrol derivatives using WHIM descriptors.[12]

Their QSRR applications are also known: retention indices of polychlorinated biphenyls have been predicted, and the enantioselectivity of several solutes used for chromatographic separation of chiral sulfoxides has also been examined.[13]

It is not a simple task to select acceptable variables from a large pool of correlating descriptors to characterize the retention properly. A simple possibility to select from among the correlating variables is to use a correlation matrix. The large number of variables and their correlation make the model too complicated and the selection task problematic (ambiguous). Several variable selection methods, such as ridge regression (RR),[14,15] partial least squares projection of latent structures (PLS),[16,17] pairwise correlation method (PCM),[18−22] stepwise linear regression e.g. in forward selection mode (FS),[23] and best subset selection (BSS)[24] have been developed to overcome this difficulty. The main advantages of ridge regression are its ability to avoid singularity in the data matrix and its predictive ability. PLS has been primarily developed for prediction, while PCM has been developed to choose between two correlating predictor variables.[18,19] The distinction between two variables can be made using an arrangement into a 2×2 contingency table. PCM can easily be generalized for selection and ranking of more than two variables.[20,21] The comparison of factors can be made pairwise in all possible combinations. Successive

* Corresponding author phone: +36 1 438 0490; fax: +36 1 325 75 54; e-mail: ofarkas@chemres.hu.

10.1021/ci049827t CCC: $30.25
Published on Web 02/19/2005

Table 1. Kováts Retention Indices of Aliphatic Alcohols on OV-1 Stationary Phase[a]

		$RI_{calculated}$				
alcohol	$RI_{experimental}$	RR 4	PLS 4	PCM 4	FS 4	BSS 4
1-butanol	646.48	625.19	626.32	653.15	650.64	645.25
1-decanol	1256.01	1211.94	1245.19	1282.46	1275.74	1278.80
1-hexanol	852.96	931.04	900.65	871.91	867.14	865.27
1-nonanol	1158.66	1147.29	1159.67	1157.11	1162.58	1164.65
1-pentanol	750.40	802.56	775.83	763.65	762.44	759.44
2,2-dimethyl-1-pentanol	867.57	858.88	861.23	846.25	869.97	869.53
2,2-dimethyl-3-hexanol	900.39	913.30	892.26	903.71	900.57	898.71
2,2-dimethyl-3-pentanol	805.63	841.56	813.12	810.13	806.57	810.96
2,4-dimethyl-3-pentanol	821.18	831.15	851.06	818.75	811.22	815.83
2-butanol	582.51	580.65	562.26	607.56	598.22	594.63
2-ethyl-1-hexanol	1012.64	1063.63	1051.36	1022.51	1039.73	1032.75
2-heptanol	885.57	937.07	915.51	902.97	893.20	890.29
2-methyl-1-butanol	722.58	725.73	715.73	709.74	713.75	715.51
2-methyl-1-pentanol	813.35	857.74	838.16	816.62	820.72	820.25
2-methyl-2-hexanol	817.33	859.48	850.93	851.11	848.06	845.94
2-methyl-2-pentanol	717.57	764.11	761.49	761.11	755.22	755.97
2-methyl-3-pentanol	757.96	788.61	773.51	768.29	754.77	757.00
2-nonanol	1084.16	1100.09	1101.61	1120.89	1094.40	1090.17
3,3-dimethyl-1-butanol	778.77	764.50	778.72	761.72	779.03	783.77
3,5-dimethyl-3-hexanol	883.13	895.50	891.48	903.61	886.25	886.97
3-hexanol	780.36	870.41	827.55	816.77	788.10	787.46
3-methyl-1-butanol	719.03	723.79	713.70	711.35	721.16	722.24
3-methyl-2-butanol	666.02	674.33	667.49	673.23	668.07	670.34
3-methyl-3-hexanol	826.62	860.95	851.74	846.31	833.77	834.00
3-pentanol	684.21	744.55	703.66	709.27	691.36	690.78
4-ethyl-3-hexanol	953.26	997.12	985.36	957.85	957.50	958.57
4-methyl-1-pentanol	821.19	853.98	835.22	817.33	830.79	830.55
4-methyl-2-pentanol	744.14	780.97	774.44	769.82	765.69	765.85
5-methyl-3-heptanol	943.58	959.86	955.19	953.42	936.91	933.55
5-methyl-3-hexanol	838.15	873.80	864.04	857.04	849.19	848.27

[a] Experimental and calculated RI values of the training set (combination A).

completion of PCM is recommended to preserve maximal diversity of the already ranked descriptors while using the same data set.[22] Stepwise linear regression has been developed to provide good models with much less calculations as compared to all possible regressions.[23] The best subset selection or all possible regression method is able to select the best combination of descriptors from the group of independent variables, as all combinations are examined.[24] The disadvantage of this method is that the calculation time increases exponentially with the increasing number of descriptors.

In our work, constitutive and WHIM descriptors have been used to build models for describing and predicting the retention properties of 44 saturated aliphatic alcohols on an apolar OV-1 stationary phase. The dipole moment of the alcohols is much higher than that of hydrocarbons, so the interaction between the stationary phase and the alcohol molecules is more complex. Polarity and polarizability have to be considered in the model, in addition to the size and shape parameters. Because of their retention properties, alcohols are good model compounds for QSRR studies. Bermejo et al. used boiling point, molar refractivity and connectivity indices to predict Kováts retention indices of alcohols.[25] Zinn et al. described the retention properties of alcohols using similar parameters complemented with the dipole moment.[26] Physical properties such as boiling point are suitable for an accurate prediction of retention indices, but it is difficult to access these data. Calculated descriptors are easily available, e.g. Guo et al. have successfully employed molecular connectivity indices[27] and Heinzen et al. have applied a semiempirical topological index to build predictive models for retention indices of saturated alcohols.[28]

Our principal aim was to compare the efficiency of several variable selection methods. The methodology for model comparison studies is not well established. We used all methods in a well-defined way solely for variable selection. The same model building technique (multiple linear regression) was used to compare the methods in a fair way. Our second goal was to build a model for description and prediction of Kováts retention indices for a wide variety of linear and branched saturated alcohols replacing the measured properties by easily calculated descriptors. In addition, we would like to know the usefulness of the WHIM descriptors in the description of gas chromatographic retention relationships. Principal component analysis was applied for preselection to ensure highly correlating descriptors in a pool. Ridge regression, partial least-squares method, pairwise correlation method, multiple linear regression (forward selection) and best subset selection have been used to reduce the number of descriptors to three or four. The fixed number of descriptors is, again, expedient for a fair model comparison. Then, models were built using multiple linear regression in a standard mode. Their predictive efficiencies have been compared.

2. EXPERIMENTAL SECTION

2.1. Retention Data. The Kováts retention indices (RI) of 44 saturated alcohols measured on an OV-1 stationary phase were taken from Pias et al.[29] and Zhang et al.[30] The data set of the Kováts retention indices of all alcohols studied has been shown in Tables 1 and 2. The compounds are linear

Table 2. Kováts Retention Indices of Aliphatic Alcohols on OV-1 Stationary Phase[a]

		$RI_{calculated}$				
alcohol	$RI_{experimental}$	RR 4	PLS 4	PCM 4	FS 4	BSS 4
1-heptanol	955.05	993.40	977.66	948.10	961.19	960.71
1-octanol	1057.34	1066.95	1064.81	1041.09	1056.76	1057.68
2,2-dimethyl-1-propanol	657.34	624.89	663.87	660.15	669.93	679.73
2,4-dimethyl-2-pentanol	775.91	816.38	811.16	811.31	807.82	811.64
2-ethyl-1butanol	825.94	958.29	897.45	816.95	806.23	824.27
2-hexanol	782.18	862.35	827.32	816.95	799.33	796.70
2-methyl-2-heptanol	916.43	945.43	939.83	951.73	949.78	943.28
2-methyl-3-hexanol	852.71	880.29	864.17	856.90	847.85	847.48
2-pentanol	682.66	731.73	704.22	711.01	701.77	698.85
3-heptanol	880.52	941.40	915.57	899.61	875.70	874.48
3-methyl-1-pentanol	828.82	858.14	839.21	816.68	824.47	823.97
3-octanol	981.75	1026.20	1013.64	1001.01	971.41	969.87
4-heptanol	875.42	947.89	915.71	899.72	871.69	871.38
4-octanol	975.50	1030.12	1013.69	1000.72	960.90	960.48

[a] Experimental and calculated RI values of the test set (combination A).

or branched and contain a hydroxyl group on the primary, secondary or tertiary carbon atoms. The number of carbon atoms varies from 4 to 10.

2.2. Computation Methodology. The geometry of alcohols was optimized by the HyperChem 4.0 program package (HyperCube Inc., Canada) using the AM 1 semiempirical method. We used this procedure to represent the three-dimensional structure of the molecules. Constitutional and WHIM descriptors were calculated using the Dragon program package.[31] The program omitted one of the two (or more) variables automatically that showed a correlation higher than 0.99. The number of calculated descriptors was 109. In the preselection step, 109 descriptors have been reduced to 17 using principal component analysis. The number of independent variables has been reduced by the PCA method using principal component loadings. A circle was defined around the Y variable (Kováts retention index) with a radius of 0.4 (when analyzing the correlation matrix). Those variables have been selected for which the loading points lie within the circle. In such a way, 17 highly correlated descriptors were selected. These 17 descriptors represent a frequent case in modeling studies: only several factors influence the system, but we cannot choose from among them as no appropriate theoretical (or physical) model exists. The real task is to choose the relevant descriptors from the highly correlated available ones.

These 17 descriptors served as a starting pool for the comparison of variable selection methods. Five different variable selection methods reduced the descriptors even further. A multiple linear regression (MLR) method was carried out to describe the linear relation between the retention index (RI) and independent variables (descriptors). Statistical calculations were performed by Statistica 5.5 software package.[32] The five different variable selection methods were the following: ridge regression (RR), partial least-squares method (PLS), pairwise correlation method (PCM), forward selection (FS) and best subset selection (BSS). By increasing the number of independent variables, the statistical parameters of the models such as coefficient of determination (R^2) and standard deviation (SD) became better. The 17 descriptors were ranked in five different ways using RR, PLS, PCM, FS and BSS. Then, the models were built with multiple linear regression using the best three and four descriptors selected by RR, PLS, PCM, FS and BSS, respectively. In such a way, we obtained comparable models. The maximum number of PLS components has been employed in the case of the PLS method. The number of the independent variables in a subset has been maximized to four in the case of BSS. The variable selection methods were compared from two points of view. The *description* ability was characterized using the following statistical parameters: coefficient of determination (R^2), Fisher ratio (F) and residual error (SD). The *predictive* ability was tested using retention index data not involved in model building: leave-n-out cross-validation[33] tests were performed using sixty seven percent of the compounds as training and thirty three percent as a test set. This procedure has been repeated four times in various combinations. The combinations have been denoted with capital letters A, B, C and D. The four combinations allow a relatively good mapping of the data structure. This means that linear and branched alcohols from the entire retention index range have been properly represented in the test set(s). The predictive ability has been measured by the R^2_{CV} (coefficient of determination for the cross-validation) values using different training and test sets in each case. $\bar{R}^2_{CV}$ is the average of the four R^2_{CV} values obtained above. Predictive error sum of squares (PRESS) was also calculated in all four cases, and their values were also averaged. Further examination of model applicability was undertaken by reviewing the residual plots and the plots of the predicted versus observed retention data for the entire retention index set.

3. RESULTS AND DISCUSSION

Kováts retention indices have been predicted for aliphatic saturated alcohols containing carbon atoms 4 to 10 using RR, PLS, PCM, FS and BSS for the variable selection process and using MLR for the model building. The calculated RI values of the alcohols using all variable selection methods and combination A (training and prediction set separation) can be seen in Tables 1 and 2. As it can be seen from the tables, very good results are mixed with not acceptable values. For example, the retention index of 2-nonanol is always overestimated. The reason for this might be that the structure encoding is not acceptable for this compound using these descriptors or a systematic experi-

Table 3. Training and Test Combinations of Alcohols in B, C and D Sets[a]

alcohols	set B	set C	set D	alcohols	set B	set C	set D
1-butanol	1	2	1	3-methyl-2-butanol	1	2	1
1-decanol	2	1	2	3-methyl-3-hexanol	2	1	1
1-hexanol	1	2	1	3-pentanol	1	2	1
1-nonanol	2	1	1	4-ethyl-3-hexanol	2	1	1
1-pentanol	1	2	1	4-methyl-1-pentanol	1	2	1
2,2-dimethyl-1-pentanol	2	1	1	4-methyl-2-pentanol	2	1	1
2,2-dimethyl-3-hexanol	1	2	1	5-methyl-3-heptanol	1	2	1
2,2-dimethyl-3-pentanol	2	1	1	5-methyl-3-hexanol	2	1	1
2,4-dimethyl-3-pentanol	1	2	1	1-heptanol	1	1	2
2-butanol	2	1	1	1-octanol	1	1	2
2-ethyl-1-hexanol	1	2	1	2,2-dimethyl-1-propanol	1	1	2
2-heptanol	2	1	1	2,4-dimethyl-2-pentanol	1	1	2
2-methyl-1-butanol	1	2	1	2-ethyl-1-butanol	1	1	2
2-methyl-1-pentanol	2	1	1	2-hexanol	1	1	2
2-methyl-2-hexanol	1	2	1	2-methyl-2-heptanol	1	1	2
2-methyl-2-pentanol	2	1	1	2-methyl-3-hexanol	1	1	2
2-methyl-3-pentanol	1	2	1	2-pentanol	1	1	2
2-nonanol	2	1	1	3-heptanol	1	1	2
3,3-dimethyl-1-butanol	1	2	1	3-methyl-1-pentanol	1	1	2
3,5-dimethyl-3-hexanol	2	1	1	3-octanol	1	1	2
3-hexanol	1	2	1	4-heptanol	1	1	2
3-methyl-1-butanol	2	1	1	4-octanol	1	1	1

[a] A set is shown in Tables 1 and 2. 1: training set, 2: test set.

mental error is suspected in the retention index measurements. The split of data into training and test sets is shown in Table 3. The statistical parameters of the MLR models have been shown in Table 4, i.e. the descriptive and predictive performances together. The models have been denoted by the abbreviation of the applied variable selection method and the number of the descriptors in the model (e.g. RR 3 means that the variable selection has been performed by ridge regression and the number of the descriptors included in the MLR model is three).

3.1. Models RR 4 and RR 3. The following MLR models have been obtained with descriptors selected by RR 4 and RR 3. Ridge regression was performed 9 times using 9 different ridge regression parameters (lambda) between 0.001 and 0.6. The slopes of the 9 regression equations were plotted versus lambda values in case of all independent variables. The name of this plot is "ridge trace".[34] The four descriptors, which showed the maximal effect on the ridge trace plot, have been used in the equation denoted RR 4: The following

$$\mathrm{RI} = -302.2\ (\pm 177.8) + 5.966\ (\pm 0.3721)\ \mathrm{MW} + 543.2\ (\pm 180.8)\ \mathrm{P1u} + 856.6\ (\pm 340.4)\ \mathrm{P2u} - 446.8\ (\pm 221.0)\ \mathrm{P2m} \quad \text{(RR 4)}$$

three descriptors have been used in the equation denoted RR 3:

$$\mathrm{RI} = 209.4\ (\pm 56.03) + 6.121\ (\pm 0.4038)\ \mathrm{MW} + 82.34\ (\pm 243.6)\ \mathrm{P2u} - 456.9\ (\pm\ 242.1)\ \mathrm{P2m} \quad \text{(RR 3)}$$

(The statistical parameters are summarized in Table 4).

The value of the R^2 and the standard deviations are nearly equal to those obtained in the PLS models, but the average coefficients of determination for cross-validation ($\bar{R}^2_{CV}$) are low, and the average cross-validation errors (PRESS) are considerably high. The four-descriptor model has been selected for illustration (Figure 1). These models are not good

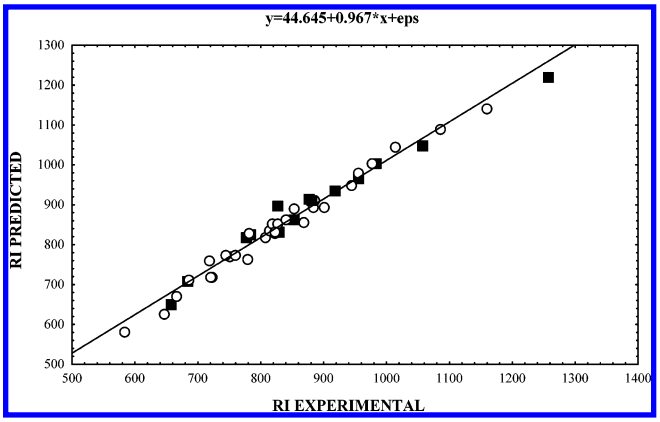

Figure 1. Model RR 4: Cross-validation. Plot of the predicted versus experimental retention indices (Combination D) RI: Kováts retention index, ○: training set of alcohols ■: test set of alcohols.

Table 4. Statistical Parameters of MLR Models[a]

method for variable selection	descriptors	R^2	SD	F	$\bar{R}^2_{CV}$	PRESS	b
RR 4	MW, P2u, P2m, P1u	0.9442	33.95	164.92	0.8193	50.23	0.68
RR 3	MW, P2u, P2m	0.9433	33.76	222.10	0.8406	44.65	0.76
PLS 4	AMW, Ms, P1u, G1m	0.9457	33.47	169.97	0.9455	33.52	1.00
PLS 3	AMW, Ms, P1u	0.9456	33.06	232.21	0.9426	31.84	1.04
PCM 4	MW, AMW, Ms, Vm	0.9819	19.29	531.56	0.9398	36.99	0.52
PCM 3	MW, AMW, Ms	0.9816	19.23	712.85	0.9397	35.17	0.55
FS 4	Ms, Vu, L1s, Tv	0.9883	15.53	824.11	0.9796	22.94	0.68
FS 3	Ms, Vu, L1s	0.9826	18.71	753.42	0.9659	29.56	0.63
BSS 4	Ms, L1u, L1e, Vu	0.9902	14.22	985.60	0.9801	22.40	0.63
BSS 3	Ms, As, Vm	0.9838	18.05	810.34	0.9707	29.49	0.61

[a] R^2: coefficient of determination; SD: standard deviation; F: Fisher ratio; $\bar{R}^2_{CV}$: average of coefficients of determination for the cross-validation R^2_{CV}; PRESS: predicted error sum of squares; b = SD/PRESS, the balance parameter. $n_{training}$ = 29 or 30: number of the alcohol molecules in the training set; n_{test} = 14 or 15: number of the alcohol molecules in the test set; MW: molecular mass; AMW: average molecular mass AMW = $\frac{1}{A}\sum_{i=1}^{A} m_i$ where m is the atomic mass and i runs over the A number of atoms of the molecule; Ms: mean electrotopological state index; Vm: V total size index weighted by atomic masses; P1u: 1st component size directional WHIM index, unweighted; P2u: 2nd component shape directional WHIM index, unweighted; P2m: 2nd component shape directional WHIM index, weighted by atomic masses; G1m: 1st component symmetry directional WHIM index, weighted by atomic masses; L1u: 1st component shape directional WHIM index, unweighted; L1e: 1st component shape directional WHIM index, weighted by atomic Sanderson electron-negativities; L1s: 1st component shape directional WHIM index, weighted by atomic electrotopological states; Tv: T total size index, weighted by atomic van der Waals volumes;Vu: V total size index, unweighted; As: A total size index, weighted by atomic electrotopological states.

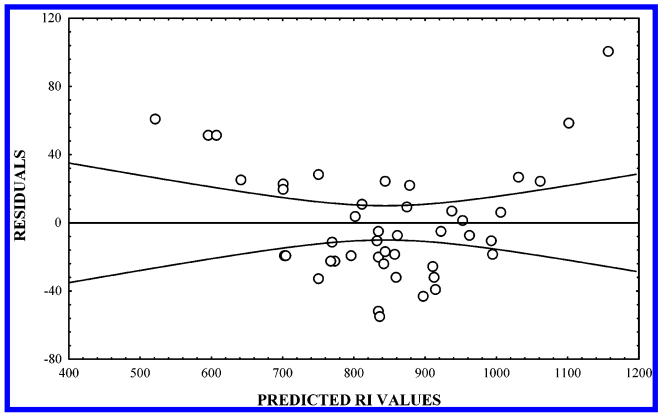

Figure 2. Model RR 4. Plot of predicted retention indices versus residuals, (without cross-validation) RI: Kováts retention index.

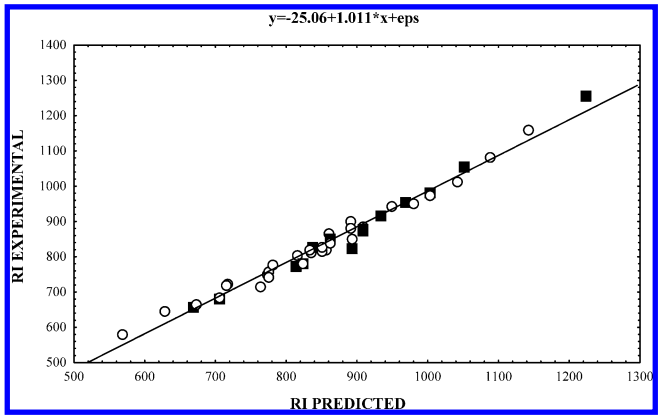

Figure 3. Model PLS 4: Cross-validation. Plot of the predicted versus experimental retention indices (Combination D), RI: Kováts retention index, ○: training set of alcohols ■: test set of alcohols.

for predicting Kováts retention indices, and these sets of descriptors do not provide a linear model. The same deviation in the plot of the predicted versus experimental retention indices could be observed as in the PLS models. This deviation can better be seen from the plot of the predicted retention indices versus residuals (Figure 2).

3.2. Models PLS 4 and PLS 3. Variable selection by PLS has been carried out in the following way: A training set has been used for variable selection and a PLS model has been built. Descriptors having the greatest regression coefficients in the PLS model were selected. These descriptors (four or three) have been used in the MLR models as independent variables. The following MLR models have been obtained with descriptors selected by PLS denoted PLS 4 and PLS 3:

$$\mathrm{RI} = -4801\ (\pm 2822) + 1891\ (\pm 684.7)\ \mathrm{AMW} - 1505\ (\pm 188.1)\ \mathrm{Ms} - 113.4\ (\pm 74.56)\ \mathrm{P1u} - 471.5\ (\pm 205.3)\ \mathrm{G1m} \quad \text{(PLS 4)}$$

$$\mathrm{RI} = -4510\ (\pm 2334) + 1820\ (\pm 563.4)\ \mathrm{AMW} - 1518\ (\pm 173.1)\ \mathrm{Ms} - 115.0\ (\pm 73.16)\ \mathrm{P1u} \quad \text{(PLS 3)}$$

It can be seen from Table 4 that the values of the average coefficients of determination for cross-validation are fairly high. However, a deviation in the plot of the predicted versus the experimental retention indices shows that these models are inadequate (see PLS 4 model in Figure 3). The coefficients of determination are smaller than in case of PCM, FS and BSS models, and the standard deviations are relatively high, so PLS 4 and PLS 3 models are not able to give a reliable representation of the retention behavior of alcohol compounds.

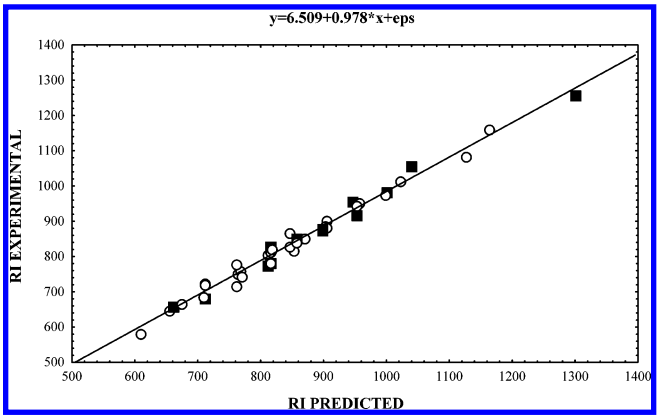

Figure 4. Model PCM 4: Cross-validation. Plot of the predicted versus experimental retention indices (Combination D), RI: Kováts retention index, ○: training set of alcohols ■: test set of alcohols.

3.3. Models PCM 4 and PCM 3. The following MLR models have been obtained with descriptors selected by PCM using probability-weighted ordering. The variables were ordered according to the subtraction of the confidence probability of losses from the confidence probability of wins. The confidence level was fixed at 95%. As a selection criterion, the Chi square test was used. The best four or three descriptors were included in the MLR models:

$$\mathrm{RI} = -15030\ (\pm 3309) + 6.150\ (\pm 1.647)\ \mathrm{MW} + 3558\ (\pm 685.8)\ \mathrm{AMW} - 944.9\ (\pm 106.4)\ \mathrm{Ms} + 3.849\ (\pm 4.385)\ \mathrm{Vm} \quad \text{(PCM 4)}$$

$$\mathrm{RI} = -17350\ (\pm 1955) + 7.350\ (\pm 0.7890)\ \mathrm{MW} + 4045\ (\pm 393.2)\ \mathrm{AMW} - 1011\ (\pm 74.14)\ \mathrm{Ms} \quad \text{(PCM 3)}$$

It can be concluded that these models describe the retention mechanism properly. The models are statistically stable and fit the data well, which can be observed from the high values of the coefficient of determination and Fisher statistics, whereas the standard deviations are relatively small. The difference in the statistical parameters between the two models (PCM 4 contains the four and PCM 3 the three best independent variables) is not significant. The plot of the predicted versus experimental retention indices for PCM 4 model has been shown in Figure 4. The plot is linear; the value of the slope is practically 1. The plot of the predicted retention indices versus residuals of the PCM models looks normal (Figure 5). There is a relatively small additive error in the intercept; the same phenomenon can be seen in the other models. The only weak point of the PCM models is the not good performance for cross-validation. $\bar{R}^2_{CV}$ and PRESS values show that the predictive ability of PCM models is not sufficient compared to other methods. No wonder, since prediction lies outside of the scope of this method.

3.4. Models FS 4 and FS 3. The following MLR models have been obtained with descriptors selected by FS. The number of the descriptors that exceeded the 5% significance level was 7. The most significant four or three descriptors

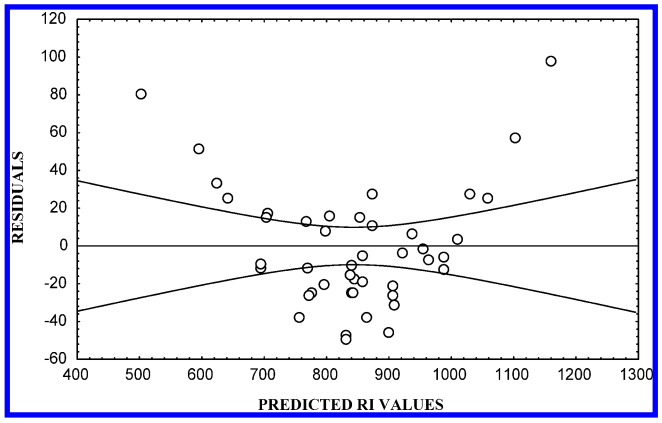

Figure 5. Model PCM 4. Plot of the predicted retention indices versus residuals, (without cross-validation) RI: Kováts retention index.

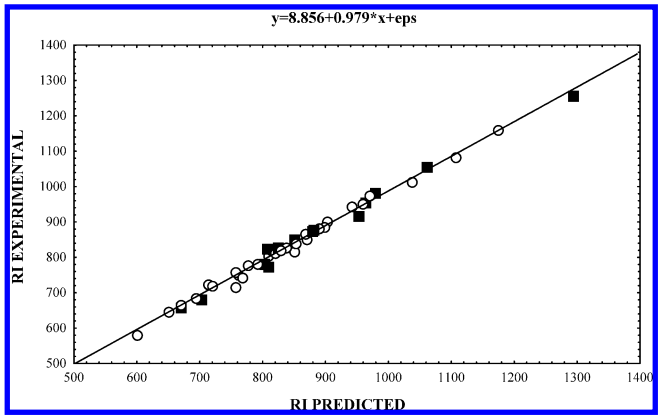

Figure 6. Model FS 4: Cross-validation. Plot of predicted versus experimental retention indices (Combination D), RI: Kováts retention index, ○: training set of alcohols ■: test set of alcohols.

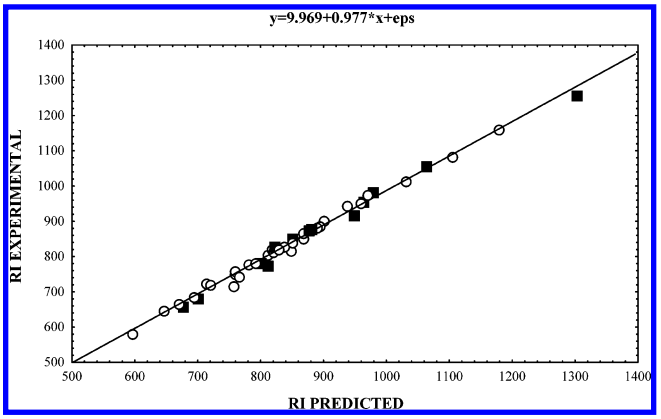

Figure 7. Model BSS 4: Cross-validation. Plot of predicted versus experimental retention indices (Combination D), RI: Kováts retention indices, ○: training set of alcohols ■: test set of alcohol

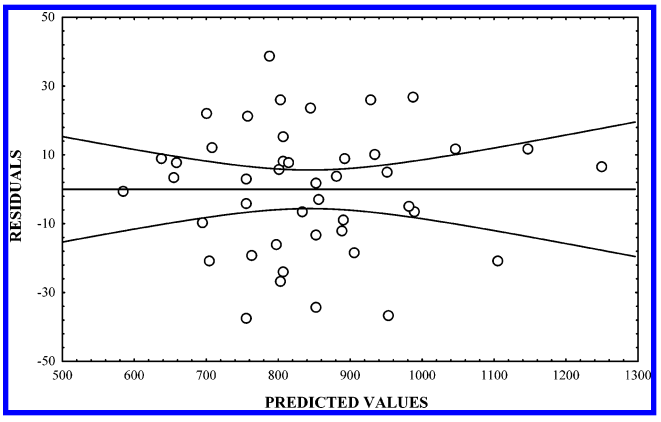

Figure 8. Model BSS 4. Plot of the predicted retention indices versus residuals, (without cross-validation) RI: Kováts retention index.

were included in the MLR models respectively:

$$RI = 1624\ (\pm 150.4) - 431.7\ (\pm 57.92)\ Ms + 21.18\ (\pm 2.674)\ L1s - 22.08\ (\pm 5.066)\ Tv + 10.10\ (\pm 1.269)\ Vu \quad (FS\ 4)$$

$$RI = 1590\ (\pm 180.9) - 424.2\ (\pm 69.71)\ Ms + 10.58\ (\pm 1.347)\ L1s + 6.773\ (\pm 1.221)\ Vu \quad (FS\ 3)$$

The values of the R^2 and $\bar{R}^2_{CV}$ are very high, and the SD and PRESS values are acceptable. It can be concluded from the statistical indices that FS resulted in reliable and applicable models for description and prediction of retention indices. Only a small additive error can be observed in the intercept. The plot of the predicted versus experimental retention indices of FS 4 model has been shown in Figure 6.

3.5. Models BSS 4 and BSS 3. The following MLR models have been obtained with descriptors selected by BSS:

$$RI = 1560\ (\pm 138.1) - 430.7\ (\pm 53.23)\ Ms - 208.7\ (\pm 23.68)\ L1u + 213.2\ (\pm 23.04)\ L1e + 8.302\ (\pm 0.968)\ Vu \quad (BSS\ 4)$$

$$RI = 1703\ (\pm 163.9) - 466.9\ (\pm 63.70)\ Ms - 14.89\ (\pm 2.189)\ As + 21.05\ (\pm 2.024)\ Vm \quad (BSS\ 3)$$

This method should provide the best linear models as all combinations of up to four descriptors have been examined. Interesting to note that the FS 4 model is only the 6th and FS 3 is only the 4th on the ranking lists of all models, with four or three descriptors, respectively.

The high value of the R^2 and small SD by the four-descriptor model indicates that this model describes RI very well. The R^2 and the SD values of the three-descriptor model are acceptable. There is a remarkable difference between the standard deviation and PRESS of the two BSS models. Besides the good descriptive properties, the best subset selection resulted in the most reliable predictive ability (see also Figure 7). The plot of the predicted retention indices versus residuals of the BSS models looks normal (Figure 8).

3.6. Comparison of Variable Selection Methods. There was a discussion in the 1990s on which method is better for prediction. Frank and Friedman[35] have proved with simulations that RR has a better performance (smaller prediction error) and shrinks the regression parameters more optimally than PLS, Principal Component Regression (PCR) or BSS. They stated that the difference in performance of RR, PLS and PCR seems not to be great, but the statistical properties of PLS and PCR "remain largely a mystery". In addition, BSS gave the worst performance results in that study, and they suggested using RR or PLS instead of "old methods" such as BSS or FS. Later Friedman et al.[36] came to the same conclusion in a prostate cancer example: PLS and RR tend to behave similarly, but RR shrinks a data set smoothly while PLS shrinks in discrete steps. This can make PLS a little

unstable. BSS shrinks also in discrete steps and overshoots often the solution. Wold[37] paraphrased this in his answer on that latent variable methods such as PLS are more useful to resolve a lot of real chemical and other scientific problems because of the presence of correlated independent variables. We have used RR, PLS and BSS not for model building but for variable selection, hence our results are not directly comparable with the above-mentioned ones. Our example shows that PLS selects from among correlated descriptors somewhat better than RR, but BSS gave the best results. PCM is not a regression method; it has been developed to rank variables and not for prediction. Thus, it is remarkable that the efficiency of PCM to select suitable descriptors for prediction is near equal to that of PLS.

The F-ratio (stopping rule) by FS provides only local control of the model search and does not attempt to find the best model. Therefore, we applied FS in a stepwise model predefining a 5% significance level. Forward selection gives stable and reliable models in the case of these descriptors. These data show that those methods, which were developed for prediction, such as PLS or RR, are not necessarily the best ones.

In the last column of Table 4, we defined a new parameter, which shows the balance between description and prediction. Values close to one show that the method is well balanced; i.e. prediction is expected to be similar to description. Results near to 0.66 are close to the expectation values as the training and test sets were shared in a ratio of 67 to 33%. Values significantly lower than 66% suggest that the description is more or less independent from the predictive ability of the method.

3.7. Descriptors. Table 4 contains the descriptors included into the models. Molecular mass and mean electrotopological state (characterizing size and polarity of the molecules, respectively) play important role in the retention behavior of alcohols. This result is not surprising because it is well-known that volatility depends on the molecular mass. In addition, solely the molecular mass is not sufficient to describe the retention; the model should be refined e.g. with the shape and electronic properties of the molecules. The appearance of the size- (Vm, Vu, L1s, L1e, Tv, As), shape- (P1u, P2u, P2m) and symmetry-related (G1m) WHIM descriptors in the models represents the role of the degree of branching and the compactness of the molecules. It is interesting that FS and BSS models do not contain MW or AMW, but contain the "size" WHIM indices. The F-ratio (stopping rule) may cause this result. The most frequently found descriptor in the models is the mean electrotopological state.

3.8. Comparison of our Models with Literature Models. Although our aim was not to reach the best models but to test the usefulness of WHIM descriptors and to compare various variable selection methods, the results are comparable with literature ones.

Direct comparison of our RI models with other RI models from the literature is difficult. Guo et al. (ref 27) used 25 alcohols in their MLR model. The descriptive ability of their combined MLR model is good (R^2=0.9932). However, no such important parameter as R^2_{CV} (correlation coefficient for cross-validation) has been given. We calculated the missing R^2_{CV} value for stationary phase SE-30 (OV-1 and SE-30 are practically the same) using Guo's equation. Their test set contained 6 alcohols only. The value of R^2_{CV} was 0.9920 and the SD_{CV} value was 21.9. The statistical parameters of our BSS 4 model were not significantly worse than the above-mentioned results. It is worth noting that it is more difficult to give a reliable prediction for 14 or 15 molecules than for 6 molecules.

Heinzen et al. built an "excellent" linear predictive model (R^2_{CV} = 0.9980) with empirical "topological" descriptors. However, they employed the experimental retention index values to calculate their empirical "topological" index. It is not surprising that their model is able to "predict" known retention indices. How they can predict "not-yet-measured" indices is unknown.

4. CONCLUSIONS

Shape- and size-related WHIM descriptors in conjunction with constitutive descriptors are suitable for describing and predicting retention properties of alcohols.

BSS was the best method for variable selection. The efficiency of FS approaches that of BSS.

RR and PLS selections were not able to describe the retention mechanism and to produce good predictive models for Kováts retention indices. Contrary to literature findings, PLS is superior to ridge regression in this particular example. On the other hand, PLS is the best balanced method for description and prediction, on the basis of a statistical balance parameter.

PCM provides good variables for description, but it is the least balanced from among the examined methods.

ACKNOWLEDGMENT

This work was supported by the Hungarian Research Foundation: OTKA T 037684. The authors thank Dr. Judit Jakus for reading the manuscript.

REFERENCES AND NOTES

(1) Kaliszan, R. Structure and Retention in Chromatography: A Chemometric Approach. In *Chromatography: Principles and Practice*; Ravindranath, B., Ed.; Harwood Academic Publishers: Amsterdam, 1997; Vol. 1.
(2) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.
(3) Héberger, K. Empirical correlation equations describing retention data of hydrocarbons on dinonylphthalate and poly(ethylene glycol) 4000. *Chromatographia* **1998**, *25*, 725−730.
(4) Héberger, K. Empirical correlations between gas-chromatographic retention data and physical or topological properties of solute molecules. *Anal. Chim. Acta* **1989**, *223*, 161−174.
(5) Héberger, K. Discrimination between linear and nonlinear models describing retention data of alkylbenzenes in gas chromatography. *Chromatographia* **1990**, *29*, 375−384.
(6) Pompe, M.; Razinger, M.; Novic, M.; Veber, M. Modelling of gas chromatographic retention indices using counterpropagation neural networks. *Anal. Chim. Acta* **1997**, *348*, 215−221.
(7) Todeschini, R.; Lasagni, M.; Marengo, E. New molecular descriptors for 2D and 3D structures − Theory. *J. Chemom.* **1994**, *8*, 263−272.
(8) Gramatica, P.; Navas, N.; Todeschini, R. 3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physicochemical properties of polychlorinated biphenyls (PCBs). *Chemom. Intell. Lab. Syst.* **1998**, *40*, 53−63.9.
(9) Di Marzio, W.; Galassi, S.; Todeschini, R.; Consolaro, F. Traditional versus WHIM molecular descriptors in QSAR approaches applied to fish toxicity studies. *Chemosphere* **2001**, *44*, 401−406.
(10) Gramatica, P.; Corradi, M.; Consonni, V. Modelling and prediction of soil sorption coefficients of nonionic pesticides by molecular descriptors. *Chemosphere* **2000**, *41*, 763−777.
(11) Vanyúr, R.; Héberger, K.; Kövesdi, I.; Jakus, J. Prediction of tumoricidal activity and accumulation of photosensitizers in photo-

dynamic therapy using multiple linear regression and artificial neural networks. *Photochem. Photobiol.* **2002**, *75*, 471−478.
(12) Vanyúr, R.; Héberger, K.; and Jakus, J. Prediction of Anti-HIV-1 Activity of a Series of Tetrapyrrole Molecules *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1829−1836.
(13) Montanari, A. C.; Cass Q. B.; Tiritan M. E.; De Souza A. L. S. A QSRR study on enantioselective separation of enantiomeric sulphoxides. *Anal. Chim. Acta* **2001**, *419*, 93−100.
(14) Hoerl, A. E.; Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55−67.
(15) Hoerl, A. E.; Kennard, R. W. Ridge regression: Applications to nonorthogonal problems. *Technometrics* **1970**, *12*, 69−82.
(16) Geladi, P.; Kowalski, B. R.; Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(17) Wold, S. PLS for Multivariate Linear Modeling. In *Chemometric Methods for Molecular Design*; Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp195−218.
(18) Héberger, K.; Rajkó, R. Discrimination of Statistically Equivalent Variables in Quantitative Structure−Activity Relationships. In *QSAR in Environmental Sciences* (7th ed.) Chen, F. G.; Schüürmann, G., Ed.; SETAC Special Publication Series, SETAC Press: 1997; pp 425−433.
(19) Rajkó, R.; Héberger, K. Conditional Fisher's exact test as a selection criterion for pair-correlation method. Type I and Type II errors. *Chemom. Intell. Lab. Syst.* **2001**, *57*, 1−14.
(20) Héberger, K.; Rajkó, R. Variable selection using pair−correlation method. Environmental applications. *SAR QSAR Environ. Res.* **2002**, *13*, 541−554.
(21) Héberger, K.; Rajkó, R. Generalization of pair−correlation method (PCM) for nonparametric variable selection. *J. Chemom.* **2002**, *16*, 436−443.
(22) Héberger, K.; Andrade, J. M. Procrustes rotation and pairwise correlation: a parametric and a nonparametric method for variable selection. *Croat. Chem. Acta* **2004**, *77*, 117−125.
(23) Draper, N. R.; Smith, H. In *Applied Regression Analysis*, 2nd ed.; John Wiley & Sons Inc.: New York, 1981; pp 294−379.
(24) Miller, A. J. *Subset Selection in Regression*; Chapman and Hall: London, 1990; pp 43−82.
(25) Bermejo, J.; Guillén, M. D. Prediction of Kováts retention index of saturated alcohols on stationary phases of different polarity. *Anal. Chem.* **1987**, *59*, 94−97.
(26) Bergman, G.; Götze, H. J.; Hermann, A.; Zinn, P. Application of target factor analysis to gas chromatography. Reproduction, prediction and classification. *Chromatographia* **1991**, *32*, 259−264.
(27) Guo, W.; Li, Y.; Zheng, X. M. The predicting study for chromatographic retention index of saturated alcohols by MLR and ANN. *Talanta* **2000**, *51*, 479−488.
(28) Junkes, B. S.; Amboni, R. D. M. C.; Yunes, R. A.; Heinzen, V. E. F. Prediction of the chromatographic retention of saturated alcohols on stationary phases of different polarity applying the novel semiempirical topological index. *Anal. Chim. Acta* **2003**, *477*, 29−39.
(29) Pias, J. B.; Gasco, L. *J. Chromatogr*. **1975**, *104*, 1 D 14 Table 885
(30) Zhang, X.; Lu, P. Unified equation between Kovats indices on different stationary phases for select types of compounds *J. Chromatogr. A* **1996**, *731/1−2*, 187−199.
(31) Todeschini, R.; Consonni, V.; Pavan, M. Dragon software version 2.1 2002.
(32) Statistica 5.5 software package StatSoft Inc., Tulsa, OK, USA.
(33) Livingstone, D. Multiple regression−robustness, chance effects and comparison of models. In *Data analysis for chemists*; Oxford University Press: 1995; pp 134−135.
(34) Draper, N. R.; Smith, H. Ridge regression. In *Applied regression analysis*; John Wiley & Sons: 1981; pp 313−319.
(35) Frank, I. E.; Friedman, J. H. A Statistical view of some chemometrics regression tools. *Technometrics* **1993**, *35*, 109−135.
(36) Hastie, T.; Tibshirani, R.; Friedman, J. Discussion: A comparison of the selection and shrinkage methods. In *The Elements of statistical learning. Data mining, inference and prediction*; Springer−Verlag: New York, 2001; pp 68−75.
(37) Wold, S. Discussion: PLS in chemical practice. *Technometrics* **1993**, *35*, 136−139.

CI049827T