Use of Computer-Assisted Methods for the Modeling of the Retention Time of a Variety of Volatile Organic Compounds: A PCA-MLR-ANN Approach

M. Jalali-Heravi* and A. Kyani

Department of Chemistry, Sharif University of Technology, P.O. Box 11365-9516, Tehran, Iran

Received October 20, 2003

A hybrid method consisting of principal component analysis (PCA), multiple linear regressions (MLR), and artificial neural network (ANN) was developed to predict the retention time of 149 C_3-C_{12} volatile organic compounds for a DB-1 stationary phase. PCA and MLR methods were used as feature-selection tools, and a neural network was employed for predicting the retention times. The regression method was also used as a calibration model for calculating the retention time of VOCs and investigating their linear characteristics. The descriptors of the total information index of atomic composition, IAC, Wiener number, W, solvation connectivity index, X1sol, and number of substituted aromatic C(sp^2), nCaR, appeared in the MLR model and were used as inputs for the ANN generation. Appearance of these parameters shows the importance of the dispersion interactions in the mechanism of retention. Comparison of the MLR and 5−2−1 ANN models indicates the superiority of the ANN over that of the MLR model. The values of 0.913 and 0.738 were obtained for the standard error of prediction set of MLR and ANN models, respectively.

1. INTRODUCTION

Volatile organic compounds (VOCs) are organic chemicals that easily vaporize at room temperature. These compounds include a very wide range of individual substances, such as hydrocarbons, halocarbons, oxygenates, etc. VOCs are one of the major groups of global atmospheric pollutants.

Hydrocarbon VOCs are usually grouped into methane and nonmethane VOCs.[1,2] There are several hundreds of different forms of VOC, and sources can be both natural and man-made. Some VOCs are quite harmful. For example, 1,3-butadiene, which its sources include the manufacturing of synthetic rubbers, petrol driven vehicles, and cigarette smoke, can cause cancer. VOCs also play a major role in the formation of various secondary pollutants through photochemical reactions in the presence of sunlight and nitrogen oxides.[3,4] Therefore, these compounds have been an important environmental issue over the last two decades and have attracted significant attention from different research groups.

There have been many attempts to provide a sensitive and specific analytical method for identifying and measuring the VOCs.[5−9] Sorbent-based or canister-based methods have been successfully developed and frequently used in the collection of ambient VOCs. However, there are some limitations for these methods. For example, sorbents are normally selective in, if not limited to, adsorbing/absorbing certain classes of VOCs. On the other hand, the conister-based technique is thought to be unsuitable for collecting polar compounds such as aldehydes and terpenes.[10] In addition, sometimes high sample volumes of air are needed for identification or measuring the VOCs. High sample volumes of humid air caused shifts in the chromatographic elution times that can degrade identification and quantification of nonspecific detectors. These are generic issues in sorbent-based air sampling that can be avoided or minimized by appropriate choices of sorbents, sample volumes, chromatographic technique, and detection method.

Despite many experimental works published on developing analytical techniques for identifying and quantifying trace levels of VOCs,[5,11] the literature contains relatively few papers on the use of computer-assisted methods for modeling of chromatographic parameters of such important class of molecules. Yan and co-workers have used ANN to predict the GC retention index of some alkylbenzenes on Carbowax-20 M.[12] Zhang et al. predicted the programmed-temperature retention values of 64 naphthas using wavelet neural networks.[13] An artificial neural network with an extended delta-bar-delta (EDBD) learning algorithm was used to predict the retention indices of 32 alkylbenzenes on three stationary phases of PEG, SE-30, and SQ.[14]

As a consequence, development of an accurate and versatile theoretical model for predicting the gas chromatographic retention time of VOCs seems to be very useful. The present paper reports on the usefulness of artificial neural network (ANN) in combination with the multiple linear regression (MLR) and principal component analysis (PCA) methods in the prediction and modeling of the retention time of 149 C_3-C_{12} volatile organic compounds obtained by GC-MS spectrometry. PCA was used as a feature selection method for clustering the descriptors and choosing the best group of them as input for the MLR and ANN models. On the other hand, there were two purposes for developing the MLR model: (1) as a feature selection technique for choosing the most suitable inputs for the neural network and (2) as a calibration model for predicting the retention time of VOCs and investigating their linear characteristics.

The present results are quite important from a point of accurate identification of VOCs. Our technique can predict the retention time of a variety of volatile organic compounds with a standard error of 0.738, and it reduces the need for different time-consuming, difficult, and expensive stages of experiment.

* Corresponding author phone: +98-600-5718; fax: +98-21-601-2983; e-mail: jalali@sharif.edu.

10.1021/ci0342270 CCC: $27.50
Published on Web 06/03/2004

Table 1. Training Set and Corresponding Observed and MLR and ANN Calculated Values of Retention Time

no.	name	RT(EXP)	RT(MLR)	RT(ANN)	no.	name	RT(EXP)	RT(MLR)	RT(ANN)
1	1,1,1-trichloroethane	19.95	19.11	19.65	57	c-1,2-dimethylcyclohexane	25.55	25.14	24.99
2	1,1,2,2-tetrachloroclethane	26.85	26.11	26.70	58	c-3-methylpent-2-ene	19.49	18.84	18.95
3	1,1,2-trichloroethane	23.22	22.48	23.17	59	c-4-methylpent-2-ene	17.88	18.46	18.55
4	1,1-dichloroethane	17.56	16.22	16.09	60	c-hept-2-ene	22.29	21.95	21.97
5	1,2,3,5-tetramethylbenzene	32.01	31.82	31.34	61	c-hex-2-ene	19.30	19.33	19.54
6	1,2,4-trichlorobenzene	33.22	35.38	32.76	62	chloroform	18.91	17.69	18.93
7	1,2.4-trimethylbenzen	29.30	28.98	29.04	63	chlorobromomethane	18.73	17.63	18.24
8	1,2,4-trimethylcyclohexane	26.18	26.87	26.67	64	chloroethane	12.84	12.18	10.87
9	1,2-dichlorobenzene	30.15	30.82	30.54	65	chloroethene	10.55	11.92	11.06
10	1,2-dichloroethane	19.69	18.58	19.04	66	chloromethane	9.03	9.81	9.17
11	1,2-dichloropropane	21.30	20.72	20.97	67	c-pent-2-ene	15.98	16.30	16.21
12	1,2-diethylbenzene	30.80	31.70	31.27	68	cyclohexene	21.22	20.52	20.86
13	1,3,5-trimethylbenzen	28.69	28.85	28.93	69	cyclopentane	17.56	17.44	17.44
14	1,3-diethylbenzene	30.49	31.33	31.09	70	dec-1-ene	29.30	28.19	29.67
15	1,4-dichlorobenzene	29.62	30.53	30.32	71	decane	29.51	28.37	29.80
16	1,4-difluorobenzene	20.85	16.87	20.19	72	dibromochloromethane	23.99	22.67	23.69
17	1-heptene	21.57	21.95	21.97	73	dibromomethane	21.26	20.10	21.26
18	1-methylcyclopentene	20.40	19.77	20.07	74	dodecane	33.65	33.85	33.81
19	1-methylcyclohexane	22.59	22.95	22.96	75	ethylbenzene	26.07	26.09	26.29
20	2,2,3-trimethylbutane	19.98	19.56	19.29	76	freon 11	14.69	14.29	15.40
21	2,2,4-trimethylpentane	21.64	21.79	21.53	77	freon 113	16.50	13.52	14.78
22	2,2-dimethylbutane	16.64	17.33	16.86	78	freon 12	7.79	8.50	8.64
23	2,2-dimethylpropane	12.04	13.57	12.00	79	heptane	21.92	22.12	22.05
24	2,3,4-trimethylpentane	23.40	22.94	22.67	80	hexane	18.85	19.50	19.56
25	2,3-dimethylpentane	21.03	21.10	20.98	81	hexachlorobutadiene	34.16	37.84	33.35
26	2,4-dimethylpentane	19.82	20.50	20.38	82	isobutane	9.98	11.68	10.12
27	2,4-dimethylhexane	22.92	23.36	23.11	83	isobutene	11.15	11.51	10.23
28	2,5-dimethylhexane	22.85	22.83	22.61	84	isoprene	15.56	15.14	14.99
29	2-ethylbut-1-ene	19.08	18.84	18.95	85	isopropylbenzene	27.66	27.7	27.76
30	2-methylbut-2-ene	16.14	15.34	14.87	86	*m*-chlorotoluene	28.30	28.61	28.74
31	2-methylheptane	23.79	23.6	23.39	87	methylene chloride	16.06	14.53	14.85
32	2-methylhexane	20.96	21.31	21.23	88	*m*-xylene	26.30	26.11	26.44
33	2-methylpentane	17.83	18.63	18.53	89	naphthalene	33.41	31.98	31.38
34	3-chloropropene	16.21	15.82	15.57	90	nonane	27.20	26.38	26.60
35	3-ethyltoluene	28.52	28.79	28.83	91	octane	24.68	24.37	24.19
36	3-methylheptane	23.99	24.12	23.89	92	octene	24.38	24.2	24.06
37	3-methylhexane	21.22	21.77	21.68	93	*o*-xylene	26.88	26.32	26.65
38	3-methylpent-1-ene	17.41	18.84	18.95	94	*p*-chlorotoluene	28.30	28.53	28.66
39	3-methylpentane	18.29	19.02	18.95	95	pent-1-ene	15.01	16.30	16.21
40	4-ethyltoluene	28.58	28.65	28.71	96	pentane	15.42	16.47	16.14
41	4-isopropyltoluene	29.98	30.08	30.21	97	propylbenzene	28.37	28.29	28.34
42	4-methylheptane	23.84	24.19	23.96	98	propylene	6.78	9.06	8.45
43	4-methylpent-1-ene	17.39	18.46	18.55	99	*p*-xylene	26.30	26.03	26.36
44	benzene	20.43	20.01	20.72	100	*sec*-buthylbenzene	29.72	30.24	30.30
45	benzyl chloride	29.51	29.10	29.08	101	styrene	26.74	25.81	26.05
46	bromodichloromethane	21.50	20.4	21.34	102	t-1,3-dichloropropene	23.02	21.87	22.41
47	bromoethane	15.71	14.96	14.59	103	t-3-methylpent-2-ene	19.18	18.84	18.95
48	bromofluorobenzen	27.46	26.58	26.88	104	t-but-2-ene	11.97	12.85	11.91
49	bromomethane	12.13	13.75	13.85	105	tertbuthylbenzene	29.30	28.79	29.00
50	bromoform	26.35	24.52	25.72	100	tetrachloromethane	20.60	19.70	21.21
51	buta-1,3-diene	11.29	12.65	12.07	107	tetrachloroethene	24.82	25.10	25.84
52	but-1-ene	11.15	12.85	11.91	108	t-hept-3-ene	21.86	21.95	21.97
53	butane	11.50	13.02	11.79	109	t-hex-2-ene	19.00	19.33	19.54
54	buthylbenzene	30.68	30.25	30.52	110	toluene	23.54	23.19	23.93
55	butyne	12.29	12.65	12.07	111	t-pent-2-ene	15.71	16.30	16.21
56	c-1,2-dichloroethene	18.56	18.19	19.04					

2. METHODS

The main aim of the present work was development of an artificial neural net to predict the retention time of different varieties of VOCs. One of the main problems in developing these types of models is choosing the proper inputs (descriptors) for them. There are two different methods of feature selection technique: objective and subjective methods. Using the former method, selects the relation between the descriptors themselves, whereas the latter method defines the relation between the descriptors and the dependent variable i.e., retention time.

We have employed a PCA, which discriminates descriptors with different features, as an objective feature selection method to classify the descriptors. Therefore, this technique was applied for choosing a suitable set of generated descriptors for developing a multiple linear regression model. The best generated MLR model was used to prepare a calibration model, which predicts the retention time of VOCs and illustrates the extension of the linear characteristics of the retention behavior of these compounds. We also have used this method as a subjective feature selection method. Therefore, the descriptors that appeared in the MLR model were used as inputs for developing the ANN.

2.1. Brief Description of Neural Networks. A detailed description of the theory behind a neural network has been adequately described elsewhere.[15−18] Presently the most

widely used ANN type is a multilayer feed-forward network, which is trained by the back-propagation (BP) learning algorithm.[19] An artificial neural network consists of a number of "neurons" or "hidden units" that receive data from the outside, process the data, and output a signal. A "neuron" is essentially a regression equation with a nonlinear output. When more than one of these neurons are used, nonlinear models can be fitted. The back-propagated network receives a set of inputs, which are multiplied by each neuron's weight. These products are summed for each neuron, and a nonlinear transfer function is applied. In this investigation, the log-sigmoid function, i.e., $f(x) = 1/(1+\exp(-x))$, was used as a transfer function. The transformed sums are then multiplied by the output weights where they are summed a final time, transformed, and interpreted. Since a back-propagation network is a supervised method, the desired output must be known for each input vector so an error can be calculated. This error is propagated backward through the network, adjusting the weights so that the next time the network sees the same input patterns, it will come closer to the desired output. The patterns are repeated many times until the network learns the relationship.

In the present work, the retention time of a variety of VOCs obtained from a GC-MS system using the nonpolar column of DB-1 was calculated by using the ANNs.

3. EXPERIMENTAL SECTION

Data of 143 C_3-C_{12} volatile organic compounds taken from ref 5 were used as the data set. We have also added the three compounds of bromochloromethane, 1,4-difluorobenzene, and bromofluorobenzene that were used as internal standards in the GC-MS analysis[5] to the data making a total of 149 VOCs listed in Tables 1 and 2. VOCs in the standards or samples were dried with the aid of a Nafion permeable membrane dryer and then concentrated cryogenically prior to gas chromatographic separation and detection by a mass spectrometer.[5] The 143 VOC peaks were monitored separately in 26 retention time windows in the SIM mode using a Hewlett-Packard HP 6890 gas chromatograph and a model 5973 mass spectrometer fitted with a 60 m $\times$ 0.32 mm i.d., 1 μm film thickness DB-1 capillary column.[5] In the present work, these compounds were randomly divided into two groups. A training set consisted of 111 molecules (Table 1), which was used for model generation, and a prediction set consisted of 38 molecules for the evaluation of the generated models (Table 2). As can be seen from this table, the prediction set was chosen in a way that adequately represents the training set.

3.1. Descriptor Generation. The next step in developing a model is generation of the numerical description of the molecular structures. The numerical descriptors are responsible for encoding important features of the structure of the molecules and can be categorized as geometric, electronic, and topological properties. A total of 150 descriptors were calculated for each compound in the data set, from which 127 parameters were calculated from the Dragon software[20] and the remaining 23 were obtained using the WinMopac package.[21] Since there were a large number of descriptors for each compound, we used a PCA to make a qualitative model, which discriminates descriptors with different features. Since a large number of different descriptors (150 parameters) were calculated for each compound, the PCA method was used as an objective feature selection technique to classify them into different groups.

Table 2. Prediction Set and Corresponding Observed and MLR and ANN Values of Retention Time

no.	name	RT(EXP)	RT(MLR)	RT(ANN)
1	1,1-dichloroethene	15.77	15.83	16.25
2	1,2,3-trimethylbenzene	29.98	29.25	29.27
3	1,2,4,5-tetramethylbenzene	32.08	31.76	31.31
4	1,2-dibromoethane	24.26	24.16	25.11
5	1,3-dichlorobenzene	29.51	30.61	30.38
6	1,4-dichlorobutane	26.74	25.34	25.55
7	1,4-diethylbenzene	30.65	31.09	31.00
8	1-methylcyclohexene	23.84	22.75	22.85
9	2,2,5-methylhexane	24.31	23.86	23.66
10	2,2-dimethylhexane	22.59	22.54	22.31
11	2,3-dimethylbutane	17.67	17.89	17.59
12	2-ethyltoluene	28.96	29.06	29.06
13	2-methylbutane	14.44	15.51	14.77
14	2-methylbut-1-ene	15.27	15.34	14.87
15	3,6-dimethyloctane	28.22	27.89	28.58
16	bromotrichloromethane	23.40	22.18	23.23
17	c-1,3-dichloropropene	22.47	21.87	22.41
18	c-but-2-ene	12.55	12.85	11.91
19	c-hept-3-ene	21.92	21.95	21.97
20	chlorobenzene	25.60	25.61	26.12
21	cyclohexane	20.73	20.71	20.91
22	cyclopentene	17.21	17.24	17.49
23	freon 114	10.30	7.53	8.50
24	freon 22	7.42	4.47	7.03
25	hexylbenzene	34.78	34.77	33.68
26	indan	30.27	30.26	30.05
27	isobuthylbenzene	29.66	29.52	29.82
28	methylcyclopentane	19.96	19.97	20.10
29	nonene	26.93	26.20	26.45
30	*o*-chlorotoluene	28.4	28.82	28.93
31	propane	7.00	9.22	8.38
32	propyne	8.51	8.84	8.55
33	t-1,2-dichloroethene	17.33	18.19	19.04
34	t-1,2-dimetcyclohexane	24.74	25.14	24.99
35	t-4-methylpent-2-ene	17.78	18.46	18.55
36	t-hept-2-ene	22.05	21.95	21.97
37	trichloroethene	21.57	21.90	22.83
38	undecane	31.65	30.69	32.78

3.2. Regression Analysis. The stepwise multiple linear regression procedure was used for model generation. The stepwise addition method implemented in the software package of SPSS[22] was used for choosing the descriptors contributing to the retention time of VOCs. As a first step, a correlation matrix was performed for all 150 descriptors calculated for each molecule. Inspection of this matrix did not show a considerable correlation ($R \geq 0.90$) between them. In each stage of the next step, a group of descriptors that was in the same area of the PCA plot was used as input for the regression analysis using the stepwise procedure. In this step, each group of descriptors was introduced to the SPSS separately, and then this procedure was repeated by the combining of every two groups and finally all three groups together. The more suitable models obtained in each stage were compared, and among them the best MLR model was chosen for further evaluation. The best MLR model consists of five descriptors, and all of them were located on the topological area in the PC1−PC2 loading plot (Figure 1). The five parameters appearing in this model were the total information index of atomic composition, IAC, the Wiener number, W, the square of the Wiener number, W^2, the solvation connectivity index, X1sol, and the number of substituted aromatic C (sp^2), nCaR. The main goals of

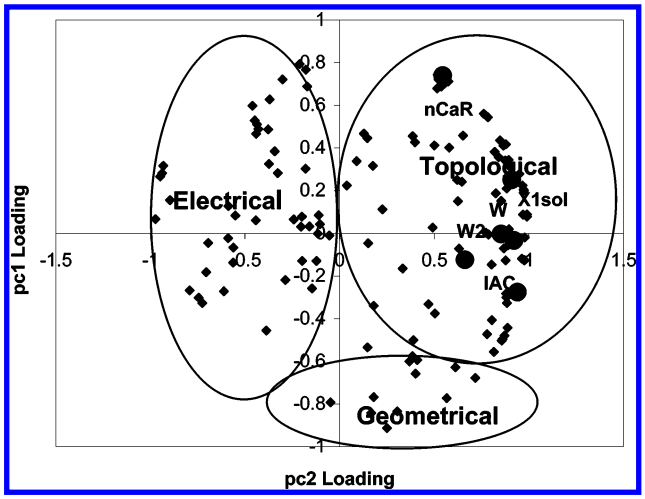

Figure 1. Principal component analysis of 150 descriptors. Highlighted circles represent the descriptors appearing in the models.

generating the MLR model were developing a calibration model for the prediction of VOCs retention time as well as choosing a set of suitable descriptors as inputs for developing the artificial neural network models.

3.3. The Neural Networks Model. A three-layer network with a sigmoidal transfer function was designed. Since the ANNs are not able to select the most suitable descriptors that would be used as their inputs, PCA and stepwise multiple regression feature-selection methods were used for this purpose. A set of five descriptors appearing in the MLR model were used as input parameters for generation of the networks. The signal of the output represents the retention time values for VOC compounds obtained using a DB-1 chromatographic column, and the number of the nodes in the hidden layer would be optimized. Such an ANN may be designed as 5-y-1 net to indicate the number of nodes in input, hidden, and output layers, respectively. The initial weights were randomly selected between −0.3 and +0.3, and the biases values of weights were set to be one. These values were optimized during the network training. The number of neurons in the hidden layer, learning rate, and momentum were optimized. To evaluate the performance of the ANN, standard error of calibration (SEC) and standard error of prediction (SEP) were used.[23] The number of neurons of the hidden layer with the minimum value of SEC was selected as the optimum number. Then learning rate and momentum were optimized in a similar way. We have also used a cross-validation method to test the validity of the ANN model.[24]

3.4. Software Used. The Dragon software[20] and the semiempirical molecular orbital package of WinMopac (version 7.21)[21] were used for calculating the descriptors. Some of the geometric and electronic descriptors were calculated using the AM1 Hamiltonian implemented in the molecular orbital package of HyperChem (version 6.01).[25] The standard version of SPSS software for windows was used to perform the regression analysis.[22] Win NN 32, free downloadable from the Internet, was used for developing the neural networks.[26]

4. RESULTS AND DISCUSSION

The main goals of the present work were as follows: (1) to accurately predict the retention time of VOCs on a nonpolar stationary phase of DB-1, (2) to achieve a better understanding of the physicochemical basis of retention and its mechanism, and (3) to compare the ability of the linear (MLR) and nonlinear (ANN) chemometric techniques in predicting the retention behavior of a diverse set of VOCs for the GC-MS method. To fulfill these goals one needs a very diverse data set. As can be seen in Tables 1 and 2, a data set consisting of 149 C_3-C_{12} VOCs with very different structures was chosen to develop the appropriate models.

4.1. Principal Component Analysis of Descriptors. Developing a general model requires a diverse set of data, and, thereby, a large number of descriptors have to be considered. Descriptors are numerical values that encode different structural features of the molecules. Selection of a set of appropriate descriptors from a large number of them requires a method which is able to discriminate between the parameters. We have performed a PCA on all 150 descriptors calculated for each molecule. The analysis of the 111×150 matrix revealed 18 components for the loading matrix using the Scree plot criteria. For which, PC1, PC2, and PC3 made 41.6, 16.9, and 8.6% contributions to the total components, respectively. In fact PC1 and 2 have made a total of 60% of the variances and, therefore, play the major role in the importance of the descriptors. It should be noted that all loading plots show similar trends; therefore, only the PC1−PC2 loading plot is shown in this paper for the sake of brevity. Figure 1 illustrates a loading plot of PCA factor 1 and factor 2 for all different types of descriptors. Inspection of this figure reveals that almost all topological descriptors are grouped together in one area. A similar pattern can be seen for the electronic and geometric descriptors. Highlighted (with a closed circle) in Figure 1 are the descriptors that have appeared in the MLR model. As one would expect, there should be no serious electronic interactions between the nonpolar VOCs and a stationary phase such as DB-1, and, therefore, the absence of the electronic descriptors in the MLR model is justified. It is noteworthy that the five descriptors appearing in the MLR model show the largest

Table 3. Best MLR Model for the Prediction of VOCs Retention Time Together with the Mean Effects of the Descriptors Appearing in the Model

descriptors	notation	regression coeficient	mean effect
(1) total information index of atomic composition	IAC	0.156 (±0.037)	2.57
(2) Wiener number	W	-9.781×10^{-2} (±0.013)	−4.88
(3) square of Wiener number	W^2	1.570×10^{-4} (0.000)	0.70
(4) solvation connectivity index	X1sol	7.885 (±0.264)	25.80
(5) number of substituted aromatic C(sp^2)	nCaR	0.916 (±0.143)	0.43
constant		−4.491 (±0.569)	

[a] The statistics for the model are as follows: $n = 111$, $R = 0.986$, SE = 1.068, $F = 755$.

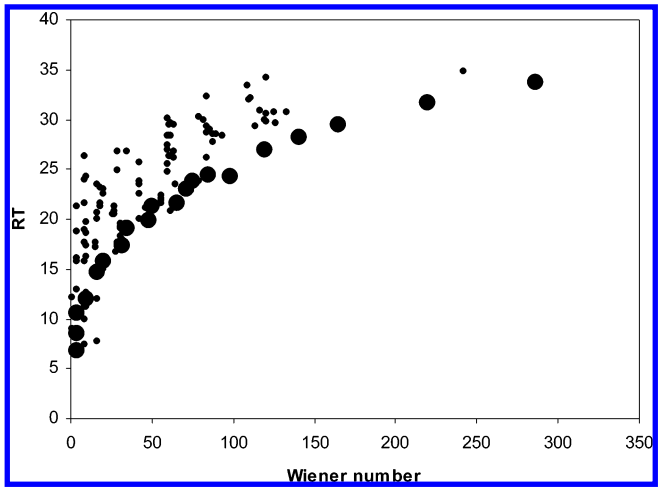

Figure 2. Plot of retention time against Wiener number of compounds in the data set. The arbitrary highlighted large circles show the nonlinearity clearer.

variances in the loading matrix of PCA analysis. Therefore, selection of these descriptors using the MLR technique is also confirmed by the PCA analysis.

4.2. Multiple Regression Analysis as Feature-Selection and Calibration Model. Linear models were formed by a stepwise addition of terms. 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 parameters.[27] However, since the number of descriptors calculated for each compound was large, we have clustered them by using PCA. Each group of descriptors was chosen as input for the software package of SPSS and then the stepwise addition method implemented in the software was used for choosing the descriptors contributing to the retention time of VOCs. The specifications for the best selected MLR model are shown in Table 3. Also, the mean effect for each descriptor is included in this table. The calculated values of retention time using this model are presented in Tables 1 and 2 for the training and prediction sets, respectively.

The parameters of IAC, W, X1sol, and nCaR appearing in the MLR model mainly show the topological characteristics. This indicates that dispersion interactions and the extent of branching of the molecules affect the retention behavior of VOCs on the DB-1 column. The solvation connectivity index shows a mean effect of 25.80, which is the largest among the descriptors appearing in the model. This parameter can be considered as entropy of solvation and somehow indicates the dispersion interactions occurring in the solutions. X1sol also is a measure of branching of the molecules. The large contributions of this parameter in the retention behavior of VOCs is in agreement with the contribution that one would expect for the interaction of a nonpolar stationary phase of DB-1 with the nonpolar VOCs. The presence of W as a topological descriptor in the model indicates that the retention time depends on the degree of branching and compactness of the molecules. It can be seen from Table 3 that the square of Wiener number (W^2) has also appeared in the MLR model. The presence of this parameter improves the ability of the model compared with that of a simple MLR. Figure 2 shows a plot of VOCs retention time versus the values of Wiener number for the molecules of the data set. It is obvious that there is no collinearity between the retention time and the Weiner number. Some points are arbitrary highlighted (with a closed circle) to show the nonlinearity clearer. Inspection of Table 3 reveals that the overall effect of Weiner number on the retention time is negative. This is in contradiction with what one may expect from the compactness of the molecules. The authors believe that there may be some interactions between this parameter and some other descriptors. Indeed, we observed a considerable improvement in statistics of the model when we added the interaction between W and the important parameter of X1sol, i.e., X1sol×W. IAC results from the chemical formula of the compounds and shows the variety of the molecules due to the differences of their atoms. This parameter shows a mean effect of 2.57. nCaR which represents the number of aromatic C(sp^2) in the molecules also shows a small contribution to the retention time of VOCs. The calculated values of the five descriptors appearing in the best MLR model are given in Table 4 for all molecules included in the training and prediction sets.

We have used two strategies for testing the validity of the predictive power of the selected MLR model. As a first strategy, the retention times for the 38 molecules included in the prediction set was calculated by the generated MLR model. The resulting correlation coefficient (R_p) of 0.992 and standard error (SE_p) of 0.913 show a good consistency with their counterparts of the training set (see Table 5) and indicate the predictive ability of this model. As a second strategy, a sort of leave-some-out cross-validation method was used.[24] We have randomly divided 149 VOCs into 10 different test sets, each consisting of 15 molecules. Then, 15 molecules have been removed each time from the data set, and a model was developed with the remaining molecules. Then the retention time of the 15 molecules was predicted by this model. This process was repeated until each molecule had a chance to be predicted once. The results of the cross-validation method for the MLR model are given in Table 6. The average R^2 value for this model is in agreement with what one would expect.

4.3. Artificial Neural Network Analysis. There were two purposes for developing the MLR model: (1) as a feature selection tool for choosing inputs for the neural network and (2) providing the possibility for comparison of the ability of linear and nonlinear models in predicting the VOCs retention time. For the sake of comparison, the descriptors used in the MLR model should be the same as the input parameters for generating the network. Therefore, a BP-ANN was generated by using the five descriptors appearing in the MLR model as its inputs.

It is noteworthy that including the parameter of W^2 as input for BNN is not very justified but comparison of the ability of the MLR and ANN models in predicting the retention times of VOCs requires the equality of the descriptors appearing in the MLR model and the ANN inputs.

The neural network methodology has several empirically determined parameters. These include the following: (1) when to stop training (i.e. the number of iterations or the convergence criterion), (2) the number of hidden nodes, and (3) learning rate and momentum terms. The values of constructed ANNs parameters were optimized with the procedure that was reported in our previous works.[17,28] A 5−2−1 neural network was developed with the optimum momentum and learning rates of 0.9 and 0.1, respectively.

Table 4. Values of the Descriptors Appearing in the Models Studied in This Work

no.	IAC	W	W2	X1sol	nCaR	no.	IAC	W	W2	X1sol	nCaR	no.	IAC	W	W2	X1sol	nCaR
Training Set																	
1	12.49	16	256	2.75	0	38	16.53	31	961	2.81	0	75	17.84	64	4096	3.93	1
2	12.00	29	841	3.80	0	39	17.63	31	961	2.81	0	76	6.85	16	256	2.25	0
3	12.49	18	324	3.20	0	40	20.69	90	8100	4.33	2	77	12.49	58	3364	2.50	0
4	12.00	9	81	2.31	0	41	23.52	120	14400	4.70	2	78	7.61	16	256	1.50	0
5	23.52	110	12100	4.61	4	42	23.15	75	5625	3.81	0	79	20.39	56	3136	3.41	0
6	18.00	84	7056	5.06	3	43	16.53	32	1024	2.77	0	80	17.63	35	1225	2.91	0
7	20.69	84	7056	4.20	3	44	12.00	27	729	3.00	0	81	9.71	121	14641	6.20	0
8	24.80	84	7056	4.20	0	45	19.30	64	4096	4.29	1	82	12.08	9	81	1.73	0
9	17.51	60	3600	4.38	2	46	9.61	9	81	2.89	0	83	11.02	9	81	1.73	0
10	12.00	10	100	2.62	0	47	10.39	4	16	2.12	0	84	12.50	18	324	2.27	0
11	15.79	18	324	2.91	0	48	19.51	60	3600	3.81	2	85	20.69	88	7744	4.31	1
12	23.52	117	13689	4.88	2	49	6.85	1	1	2.00	0	86	19.30	61	3721	4.08	2
13	20.69	84	7056	4.18	3	50	6.85	9	81	3.46	0	87	7.61	4	16	2.12	0
14	23.52	121	14641	4.86	2	51	9.71	10	100	1.91	0	88	17.84	61	3721	3.79	2
15	17.51	62	3844	4.37	2	52	11.02	10	100	1.91	0	89	17.84	109	11881	4.97	2
16	17.51	62	3844	2.63	2	53	12.08	10	100	1.91	0	90	25.91	120	14400	4.41	0
17	19.28	56	3136	3.41	0	54	23.52	133	17689	4.93	1	91	23.15	84	7056	3.91	0
18	15.27	26	676	2.89	0	55	9.71	10	100	1.91	0	92	22.04	84	7056	3.91	0
19	19.28	42	1764	3.39	0	56	9.51	10	100	2.62	0	93	17.84	60	3600	3.81	2
20	20.39	42	1764	2.94	0	57	22.04	60	3600	3.81	0	94	19.30	62	3844	4.08	2
21	23.15	66	4356	3.42	0	58	16.53	31	961	2.81	0	95	13.77	20	400	2.41	0
22	17.63	28	784	2.56	0	59	16.53	32	1024	2.77	0	96	14.86	20	400	2.41	0
23	14.86	16	256	2.00	0	60	19.28	56	3136	3.41	0	97	20.69	94	8836	4.43	1
24	23.15	65	4225	3.55	0	61	16.53	35	1225	2.91	0	98	8.26	4	16	1.41	0
25	20.39	46	2116	3.18	0	62	6.85	9	81	2.60	0	99	17.84	62	3844	3.79	2
26	20.39	48	2304	3.13	0	63	9.61	4	16	2.48	0	100	23.52	121	14641	4.84	1
27	23.15	71	5041	3.66	0	64	10.39	4	16	1.77	0	101	16.00	64	4096	3.93	1
28	23.15	74	5476	3.63	0	65	8.75	4	16	1.77	0	102	13.77	20	400	3.12	0
29	16.53	31	961	2.81	0	66	6.85	1	1	1.50	0	103	16.53	31	961	2.81	0
30	13.77	18	324	2.27	0	67	13.77	20	400	2.41	0	104	11.02	10	100	1.91	0
31	23.15	79	6241	3.77	0	68	15.27	27	729	3.00	0	105	23.52	114	12996	4.61	1
32	20.39	52	2704	3.27	0	69	13.77	15	225	2.50	0	106	3.61	16	256	3.00	0
33	17.63	32	1024	2.77	0	70	27.55	165	27225	4.91	0	107	5.51	29	841	3.80	0
34	12.16	10	100	2.27	0	71	28.67	165	27225	4.91	0	108	19.28	56	3136	3.41	0
35	20.69	88	7744	4.33	2	72	9.61	9	81	3.18	0	109	16.53	35	1225	2.91	0
36	23.15	76	5776	3.81	0	73	7.61	4	16	2.83	0	110	14.95	42	1764	3.39	1
37	20.39	50	2500	3.31	0	74	34.19	286	81796	5.91	0	111	13.77	20	400	2.41	0
Prediction Set																	
1	9.51	9	81	2.31	0	14	13.77	18	324	2.27	0	27	23.52	126	15876	4.79	1
2	20.69	82	6724	4.22	3	15	28.67	141	19881	4.70	0	28	16.53	26	676	2.89	0
3	23.52	111	12321	4.61	4	16	6.85	16	256	3.25	0	29	24.79	120	14400	4.41	0
4	12.00	10	100	3.33	0	17	13.77	20	400	3.12	0	30	19.30	60	3600	4.09	2
5	17.51	61	3721	4.37	2	18	11.02	10	100	1.91	0	31	9.30	4	16	1.41	0
6	19.30	35	1225	3.62	0	19	19.28	56	3136	3.41	0	32	6.90	4	16	1.41	0
7	23.52	125	15625	4.86	2	20	15.90	42	1764	3.68	1	33	9.51	10	100	2.62	0
8	18.04	42	1764	3.39	0	21	16.53	27	729	3.00	0	34	22.04	60	3600	3.81	0
9	25.91	98	9604	3.92	0	22	12.50	15	225	2.50	0	35	16.53	32	1024	2.77	0
10	23.15	71	5041	3.56	0	23	12.00	58	3364	1.75	0	36	19.28	56	3136	3.41	0
11	17.63	29	841	2.64	0	24	9.61	9	81	0.87	0	37	8.75	18	324	3.20	0
12	20.69	86	7396	4.34	2	25	29.13	242	58564	5.93	1	38	31.43	220	48400	5.41	0
13	14.86	18	324	2.27	0	26	18.96	79	6241	4.47	2						

Table 5. Comparison between the Different Statistical Parameters of MLR and ANN Models

method	R_t	R_p	R^2_t	R^2_p	SE_t	SE_p
MLR	0.986	0.992	0.973	0.984	1.068	0.913
ANN	0.995	0.990	0.989	0.990	0.654	0.738

We used early stopping to optimize learning iteration size and avoid overtraining. The SEC and SEP were used as the error functions and were reported every 500 trainings. These error functions were computed as follows

$$SE = \sqrt{\frac{\sum_{i=1}^{n}(t_i - y_i)^2}{n}} \quad (1)$$

where t_i is the teaching output (desired output) in the analyzed set, y_i is the actual output (target) in the analyzed set, and n indicates the number of training or prediction patterns. The calculated values of SEP and SEC were plotted against the number of iterations (Figure 3). The overfitting will start after 5000 trainings of the network. Therefore, the optimum number of iterations was chosen to be 5000 for the generated ANNs.

To evaluate the neural network, a leave-some-out cross-validation technique, similar to that used for the MLR model, was performed. The results are summarized in Table 6. Inspection of the results reveals the stability of the network. These results also show a good predictive ability for the ANN model.

To compare the chemometric methods of MLR and ANN in predicting the retention time of VOCs, some statistics for

Table 6. Results of the Cross-Validation Procedure for the MLR and ANN Models

no. of model	MLR shrinking cross-validation			ANN shrinking cross-validation		
	R^2_t	R^2_p	S.C.V[a]	R^2_t	R^2_p	S.C.V[a]
1	0.977	0.986	−0.009(15)	0.990	0.981	0.009(15)
2	0.979	0.985	−0.006(15)	0.991	0.935	0.055(15)
3	0.976	0.985	−0.009(15)	0.990	0.990	0.000(15)
4	0.976	0.985	−0.009(15)	0.892	0.927	−0.036(15)
5	0.978	0.986	−0.008(15)	0.990	0.990	0.000(15)
6	0.975	0.986	−0.010(15)	0.990	0.982	0.007(15)
7	0.978	0.981	−0.002(15)	0.949	0.892	0.057(15)
8	0.978	0.986	−0.008(15)	0.990	0.989	0.002(15)
9	0.976	0.985	−0.009(15)	0.990	0.995	−0.006(15)
10	0.976	0.985	−0.009(14)	0.990	0.984	0.007(14)

[a] S.C.V = $R^2_t - R^2_p$; numbers in the parentheses represent the number of the molecules that have been removed each time in model development and were predicted by the generated model.

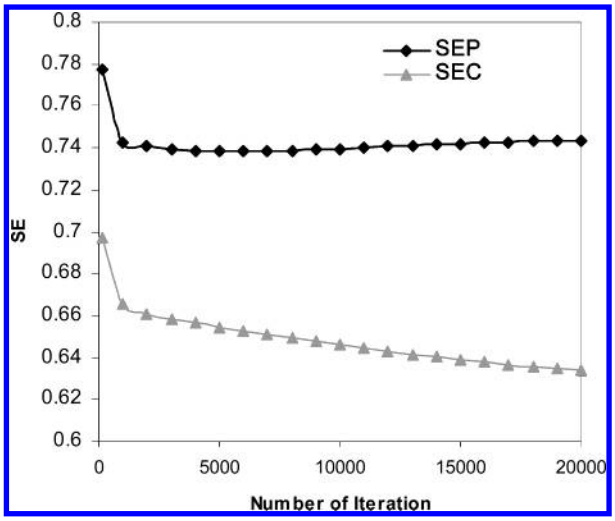

Figure 3. Variations of SE versus the number of iterations for the training and prediction sets.

these models are included in Table 5. It can be seen from this table that despite the similarity between the correlation coefficients for the training and prediction sets, the SEs have improved considerably in the case of ANN. The values of 1.068 and 0.913 for the standard error of MLR training and prediction sets, respectively, have to be compared with the values of 0.654 and 0.738 for their counterparts in the ANN model. Figure 4 shows the plot of the ANN predicted versus the experimental values of the retention time for the data set.

To investigate the existence of a systematic error in developing the ANN model, the residuals of ANN predicted values of DB-1 retention times were plotted against the experimental values in Figure 5. The propagation of the residuals on both sides of zero indicates that no systematic error exists in the development of the neural network. It is noteworthy that a similar plot for the MLR model shows some contradictions. Most of the MLR calculated retention times for the molecules eluted before around 20 min seem to be underestimated and show a negative residual. Therefore, one may conclude that the VOCs studied in the present work show different retention characteristics. For the molecules eluted before 20 min a nonlinear characteristic can be expected for their retention behaviors on a DB-1 stationary phase.

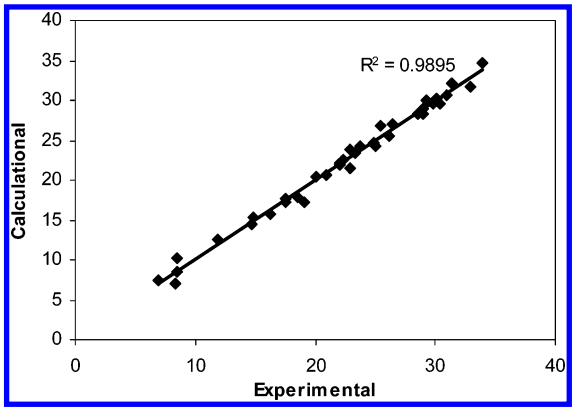

Figure 4. Plot of the ANN calculated retention times against the experimental values for the prediction set.

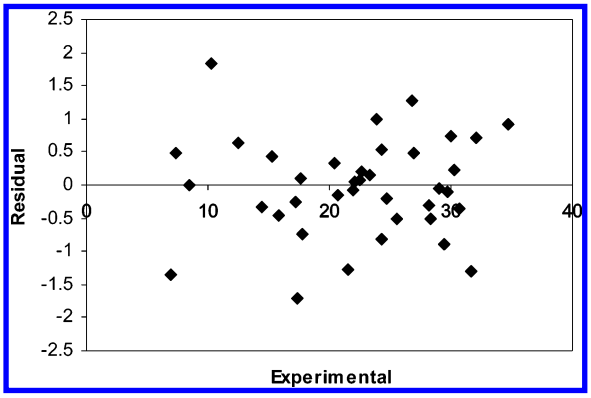

Figure 5. Plot of residuals versus experimental values of retention time for the ANN model.

5. CONCLUSION

An accurate and versatile artificial neural network was developed for predicting the retention time of some C_3-C_{12} volatile organic compounds on a nonpolar DB-1 stationary phase. The problem of the inability of the ANNs in selecting the appropriate descriptors as their inputs was overcome by using the PCA and linear regression techniques. PCA can be used as a powerful tool for clustering the descriptors when a large number of them with different features are available. Appearance of topological descriptors in the linear regression model demonstrated that the dispersion interactions are mainly responsible for the retention of a volatile organic molecule on the DB-1 stationary phase. A large contribution from the solvation connectivity index to the retention time revealed the importance of the entropy of solvation in retention behavior. Comparison of the linear (MLR) and nonlinear (ANN) methods showed the superiority of the ANN over that of the MLR models for the prediction of the retention time of VOCs. The results suggest that choosing the appropriate inputs is the main key in developing an ANN model, and PCA and MLR techniques can be considered as powerful feature-selection tools for this purpose.

Supporting Information Available: A total of 150 descriptors calculated for each molecule in the data set and the value of these descriptors for 149 molecules of the training and prediction sets (a 149 × 150 matrix). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Guenther, A.; Hewitt, C. N.; Erickson, D.; Fall, R.; Geron, C.; Gradel, T.; Harley, P.; Klinger, L.; Lerdau, M.; McKay, W. A.; Pierce, T.;

Scholes, B.; Steinbrecher, R.; Tallamrauju, R.; Taylor, J.; Zimmerman, P. A Global Model of Natural Volatile Organic Compound Emissions. *J. Geophys. Res.* **1995**, *100*, 8873.
(2) Scientific Assessment of Ozone Depletion: 1994, Report No. 37, World Meteorological Organization, Geneva, 1995.
(3) Derwent, R. G.; Jenkin, M. E.; Saunders, S. M. Photochemical Ozone Creation Potentials for a Large Number of Reactive Hydrocarbons under European Conditions. *Atmos. Environ.* **1996**, *30*, 181.
(4) Atkinson, R. Atmospheric Chemistry of VOCs and NO_x. *Atmos. Environ.* **2000**, *34*, 2063.
(5) Sin, D. W.-M.; Wong, Y.-C.; Sham, W.-C.; Wang. D. Development of an Analytical Technique and Stability Evaluation of 143 C_3-C_{12} Volatile Organic Compounds in Summa Canisters by Gas Chromatography Mass Spectrometry. *Analyst* **2001**, *126*, 30.
(6) Oliver, K. D.; Pleil, J. D.; McClenny, W. A. Sample Integrity of Trace Level Volatile Organic Compounds in Ambient Air Stored in SUMMA Polished Canisters. *Atmos. Environ.* **1986**, *20*, 1403.
(7) Gholson, A. R.; Jayant, R. K. M.; Storm. J. F. Evaluation of Aluminum Canisters for the Collection and Storage of Air Toxics. *Anal. Chem.* **1990**, *62*, 1899.
(8) Hsu, J. P.; Miller, G.; Moran. V. Analytical Method for Determination of Trace Organics in Gas Samples Collected by Canister. *J. Chromatogr. Sci.* **1991**, *29*, 83.
(9) Pankow, J. F.; Luo, W.; Isabelle, L. M.; Bender, D. A.; Baker. R. J. Determination of a Wide Range of Volatile Organic Compounds in Ambient Air Using Multisorbent Adsorption/Thermal Desorption and Gas Chromatography/Mass Spectrometry. *Anal. Chem.* **1998**, *70*, 5313.
(10) Batterman, S. A.; Zhang, G.-Z.; Batterman. M. Analysis and Stability of Aldehydes and Terpenes in Electropolished Canisters. *Atmos. Environ.* **1998**, *32*, 1647.
(11) Peng, C.-Y.; Batterman. S. Performance Evaluation of a Sorbent Tube Sampling Method Using Short Path Thermal Desorption for Volatile Organic Compounds. *J. Environ. Monit.* **2000**, *2*, 313.
(12) Yan, A.; Jiao, G.; Hu, Z.; Fan, B. T. Use of Artificial Neural Networks to Predict the Gas Chromatographic Retention Index Data of Alkylbenzenes on Carbowax-20M. *Comput. Chem.* **2000**, *24*, 171.
(13) Zhang, X.; Qi, J.; Zhang, R.; Liu, M.; Hu, Z.; Xue, H.; Fan, B. Prediction of Programmed-Temperature Retention Values of Naphthas by Wavelet Neural Networks. *Comput. Chem.* **2001**, *25*, 125.
(14) Zhang, R.; Yan, A.; Liu, M.; Liv, H.; Hu, Z. Application of Artificial Neural Networks for Prediction of the Retention Indices of Alkylbenzenes. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 113.
(15) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; VCH: Weinheim, 1999.
(16) Bose, N. K.; Liang, P. *Neural Network, Fundamentals*; McGraw-Hill: New York, 1996.
(17) Jalali-Heravi, M.; Garkani-Nejad, Z. Prediction of Electrophoretic Mobilities of Sulfonamides in Capillary Zone Electrophoresis Using Artificial Neural Networks. *J. Chromatogr. A* **2001**, *927*, 211.
(18) Jalali-Heravi, M.; Parastar, F. Development of Comprehensive Descriptors for Multiple Linear Regression and Artificial Neural Network Modeling of Retention Behaviors of a Variety of Compounds on Different Stationary Phases. *J. Chromatogr. A* **2000**, *903*, 145.
(19) Patterson, D. W. *Artificial Neural Networks: Theory and Applications*; Simon and Schuster: New York, 1996; Part III, Chapter 6.
(20) Todeschini, R.; Consonni V.; Pavan, M. Dragon Software version 2.1, via Pisani, 13-20124 Milano, Italy, 2002.
(21) www.psu.ru/science/soft/winmopac/index_e.html.
(22) SPSS for Windows, version 10.05, Standard version, SPSS Inc., 1999.
(23) Blank, T. B.; Brown, S. T. Nonlinear Multivariate Mapping of Chemical Data Using Feed-forward Neural Networks. *Anal. Chem.* **1993**, *65*, 3084.
(24) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from Similarity Matrixes. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, *36*, 2929.
(25) HyperChem, Molecular Modeling System, Hyper Cube, Inc. and Auto esk, Inc., 1993.
(26) www.geocities.com/sciware/winnn.htm.
(27) Draper, N.; Smith. H. *Applied Regression Analysis*, 2nd ed.; Wiley-Interscience: New York, 1981; p 307.
(28) Jalali-Heravi, M.; Parastar, F. Use of Artificial Neural Networks in a QSAR Study of Anti-HIV Activity for a Large Group of HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 147.

CI0342270