

Quantitative Structure Retention Relationship Study of GC Retention Times of Complex Petroleum Compounds by Different Chemometrics Methods

A. Niazi *

Chemistry Department, Young Researcher Club, Arak Branch, Islamic Azad University, Arak, Iran

A. Yazdanipour

Chemistry Department, Arak Branch, Islamic Azad University, Arak, Iran

Abstract

Introduction: QSRR study is suggested for the prediction of GC retention times of complex petroleum compounds. Modeling of the GC retention times as a function of molecular structures was established by means of the chemometrics methods such as PLS, OSC-PLS and LS-SVM.

Aim: These models were applied for the prediction of the GC retention times of these compounds, which were not in the modeling procedure.

Material and Method: Descriptors were calculated utilizing Dragon software. These descriptors are calculated using two-dimensional representation of the molecules and therefore geometry optimization is not essential for calculating these types of descriptors.

Results: Good results were achieved in LS-SVM model with percentage error ranges from -0.14 to 0.12 for retention times. The resulted model showed high prediction ability with RMSEP of prediction of 0.0152 for LS-SVM.

Conclusion: OSC-PLS and LS-SVM models were established to predict the GC retention time of some organic compounds in petroleum sample. A proper model with high statistical quality and low prediction errors was obtained.

Keywords: QSAR/QSRR; GC Retention times; PLS; LS-SVM; OSC.

Introduction

Petroleum fractions are complex hydrocarbon mixtures containing hundreds of different hydrocarbons and hetero-compounds in widely differing concentrations. Numerous components of these complex mixtures are unavailable as standards making identification difficult. Hydrocarbon mixtures are characterized in two ways, by structural group analysis and boiling range determination by distillation or simulated distillation.^[1] Mass spectrometry (MS), high performance liquid chromatography (HPLC) and capillary GC^[2] have been used as powerful techniques to solve this analytical problem but each of these analytical techniques suffer from some limitations or disadvantages.

*Corresponding author

An important property that has been extensively studied in quantitative structure activity relationship (QSAR) is relationship (QSRR) study involves the prediction of chromatographic retention parameters using molecular structure. QSRR studies we widely investigated in gas chromatography (GC) and high-performance liquid chromatography (HPLC). The chromatographic parameters are expected to be proportional to a free energy change that is related to the solute distribution on the column. Chromatographic retention is a physical phenomenon that is primarily dependent on the interactions between the solute and the stationary phase. Molecular group contribution methods are widely employed to estimate gas chromatographic retention parameter. Among the investigation of QSAR, one of the most important factors affecting the quality of the model is the method to build the model. Many multivariate data analysis methods such as multiple linear regression (MLR),^[4] partial least squares (PLS)^[5] and artificial neural network (ANN)^[6] have been used in QSAR studies. MLR, as most commonly used chemometrics method, has been extensively applied to QSAR investigations. However, the practical usefulness of MLR in QSAR studies is rather limited, as it provides relatively poor accuracy. ANN offers satisfactory accuracy in most cases but tends to overfit the training data. The support vector machine (SVM) is a popular algorithm developed from the machine learning community. Due to its advantages and remarkable generalization performance over other methods, SVM has attracted attention and gained extensive applications.^[7,8] As a simplification of traditional of SVM, Suykens and Vandewalle^[9,10] have proposed the use of least-squares SVM (LS-SVM). LS-SVM encompasses similar advantages as SVM, but its additional advantage is that it requires solving a set of only linear equations (linear programming), which is much easier and computationally more simple.^[11]

The basic principle of the multivariate calibration is the simultaneous utilization of many independent variables, x_1, x_2, \dots, x_n , to quantify one or more dependent variables of interest, y . The partial least squares (PLS) regression analysis is the most widely used method for this purpose, and it is based on the latent variable decomposition relating two blocks of variables, matrices X and Y , which may contain spectral and concentration data, respectively. These matrices can be simultaneously decomposed into a sum of f latent variables, as follows:

$$Y = TP^T + E = \sum t_f p_f + E \quad (1)$$

$$Y = UQ^T + E = \sum u_f q_f + E \quad (2)$$

in which T and U are the score matrices for X and Y , respectively; P and Q are the loadings matrices for X and Y , respectively, E and F are the residual matrices. The two matrices are correlated by the scores T and U , for each latent variable, as follows:

$$u_f = b_f t_f \quad (3)$$

in which b_f is the regression coefficient for the f latent variable. The matrix Y can be calculated from u_f , as Eq. (4), and the concentration of the new samples can be estimated from the new scores T^* , which are substituted in Eq. (4), leading to Eq. (5)

$$Y = TBQ^T + F \quad (4)$$

$$Y_{new} = T^* BQ^T \quad (5)$$

In this procedure, it is necessary to find the best number of latent variables, which normally is performed by using cross-validation, based on determination of minimum

prediction error. Several determinations based on the application of this method to spectrophotometric and QSAR data have been reported by several workers.^[12-16]

Orthogonal signal correction (OSC) was introduced by Wold et al. to remove systematic variation from the response matrix X that is unrelated, or orthogonal, to the property matrix Y . Therefore, one can be certain that important information regarding the analyte is retained. Since then, several groups have published various OSC algorithms in an attempt to reduce model complexity by removing orthogonal components from the signal.^[17, 18]

Theory of LS-SVM has also been described clearly by Suykens et al. and application of LS-SVM in quantification, classification and QSAR reported by some of the workers.^[19-21] So, we will only briefly describe the theory of LS-SVM. The LS-SVM is capable of dealing with linear and nonlinear multivariate calibration and resolves multivariate calibration problems in a relatively a fast way. In LS-SVM a linear estimation is done in kernel-induced feature space ($y = w^T \phi(x) + b$). In the present paper, the PLS and LS-SVM methods were applied in QSAR/QSRR for modeling the relationship between the GC retention times of 36 compounds by using structural molecular descriptors.

Materials and methods

Hardware and software

The computations were made with an AMD 2000 XP (1 Gb RAM) microcomputer with the Windows XP operating system and with Matlab (version 6.5, Mathwork, Inc.). The PLS evaluations were carried out by using the PLS program from PLS-Toolbox Version 2.0 for use with Matlab from Eigenvector Research Inc. The LS-SVM optimization and model results were obtained using the LS-SVM lab toolbox (Matlab/C Toolbox for Least-Squares Support Vector Machines). ChemDraw Ultra version 9.0 (ChemOffice 2005, CambridgeSoft Corporation) software was used to draw the molecular structures and optimization by the AM1.

Descriptors were calculated utilizing Dragon software (Milano Chemometrics and QSAR research group, <http://www.disat.unimib.it/chm/>). These descriptors are calculated using two-dimensional representation of the molecules and therefore geometry optimization is not essential for calculating these types of descriptors.

Data set

The QSRR model fro the estimation of the GC retention times of complex petroleum compounds is established in the following steps: the molecular structure input and generation of the files containing the chemical structures is stored in a computer-readable format; quantum mechanics geometry is optimized with a semi-empirical (AM1) method; structural descriptors are computed; and the structural-retention time model is generated by the chemometrics methods and statistical analysis. The retention time of 36 organic compounds was collected from.^[1]

Results and discussion

GC retention times of 36 of complex petroleum compounds taken from the literature,^[1] and is presented in Table 1. A major step in constructing QSAR/QSRR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number. A wide variety of descriptors have been reported to be used in QSAR/QSRR analysis.^[22]

A pool containing molecular descriptors is derived to property characterize the chemical structure of the these compounds, involving variables of the type Constitutional, Topological Geometrical, Charge, GETAWAY (GEometry, Topological, Atoms-Weighted

Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic molecular profiles, Radial Distribution Functions, Functional Groups and Atom-Centered Fragments. These variables are calculated by means of the software Dragon version 5.4. For the evaluation of the predictive ability of a different model, the root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP) can be used.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^n (y_{pred} - y_{obs})^2}{n}} \quad (6)$$

$$RSEP(\%) = 100 \times \sqrt{\frac{\sum_{i=1}^n (y_{pred} - y_{obs})^2}{\sum (y_{obs})^2}} \quad (7)$$

where y_{pred} is the predicted concentration in the sample, y_{obs} is the observed value of the concentration in the sample and n is the number of samples in the validation set.^[13] the chromatographic retention time.^[3] A quantitative structure retention

Table 1-GC retention times of complex petroleum compounds in the present study [1].

No.	Substance	Retention time (s)	No.	Substance	Retention time (s)
1 ^t	Methane	8.15	31 ^p	n-Nonane	15.42
2 ^t	Ethane	8.21	32 ^t	n-Decane	18.25
3 ^t	Propane	8.31	33 ^p	n-Undecane	20.72
4 ^t	iso-Butane	8.43	34 ^t	n-Dodecane	23.14
5 ^p	n-Butane	8.52	35 ^t	n-Tridecane	25.06
6 ^t	n-Pentane	8.91	36 ^t	n-Tetradecane	26.72
7 ^t	iso-Pentane	8.79	37 ^t	n-Pentadecane	28.14
8 ^t	2,2-DiMeBu	9.11	38 ^t	n-Hexadecane	29.54
9 ^t	2-Methylpentane	9.32	39 ^t	Benzene	10.38
10 ^p	3-Methylpentane	9.45	40 ^t	Toluene	12.05
11 ^t	n-Hexane	9.56	41 ^t	Ethylbenzene	14.44
12 ^t	Cyclohexane	10.48	42 ^t	p-Xylene	14.65
13 ^t	n-Heptane	10.84	43 ^t	m-Xylene	15.21
14 ^t	2,4-Dimethylpentane	10.84	44 ^t	o-Xylene	14.65
15 ^t	2,5-Dimethylhexane	11.71	45 ^p	n-Propylbenzene	16.38
16 ^p	2,3,4-Trimethylpentane	11.68	46 ^t	Ethanol	8.43
17 ^t	Methylcyclohexane	11.40	47 ^t	Propanol	8.97
18 ^t	n-Octane	12.78	48 ^t	2-Propanol	8.43

^t training set

^p prediction set

In order to detect the homogeneities in the data set and identify possible outliers and clusters, PCA was performed within the calculated structure descriptors space for the whole data set. PCA is a useful multivariate statistical technique in which new variables (called principal components, PCs) are calculated as linear combinations of the old ones. These PCs are sorted by decreasing information content (i.e. decreasing variance) so that most of the information is preserved in the first few PCs. An important feature is that the obtained PCs are uncorrelated, and they can be used to derive scores which can be used to display most of

the original variations in a smaller number of dimensions. These scores can also allow us to recognize groups of samples with similar behavior (Fig. 1).

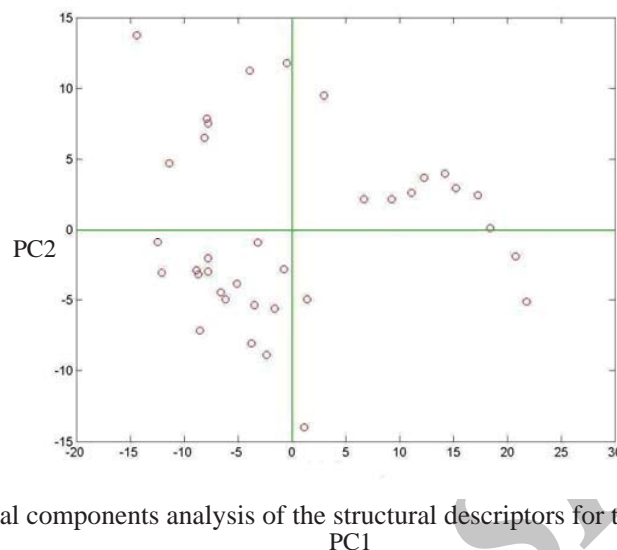


Fig. 1- Principal components analysis of the structural descriptors for the data set.

PLS analysis

The factor-analytical multivariate calibration method is a powerful tool for modeling, because it extracts more information from the data and allows building more robust models. According to GC retention time data (Table 1), data randomly classified to training and prediction sets. The optimum number of factors to be included in the calibration model was determined by computing the prediction error sum of squares (PRESS) from cross-validated models using a high number of factors (half of the number of total training set + 1). The cross-validation method employed was to eliminate only one compound at a time and then PLS calibrated the remaining of training set. The retention time of the left-out sample was predicted by using this calibration. This process was repeated until each compound in the training set had been left out once. According to Haaland suggestion,^[23] the optimum number of factor was selected (Fig. 2).

Preprocessing by orthogonal signal correction

For calibration set five OSC components were used for filtering. Evaluation of the prediction errors for the validation set reveals that the OSC treated data give substantially lower RMSEP values than original data. Also, the OSC-filtered data give much simpler calibration models with fewer components than the ones based on original data. The results imply that the OSC method indeed removes information from descriptor data that is not necessary for fitting of the Y-variables. In some cases the OSC method also removes non-linear relationships between X and Y. The score plots for the PLS and OSC-PLS are shown in Fig. 2. As score plots reveal the geometrical placement of the solutions in principal components space. The experimental noise can destroy this relation but by removing the noise using OSC filtering, the OSC-PLS score plots depicted in a more clear way the location of the solutions in the scores map which are the same as square experimental design was used in preparation of calibration set.

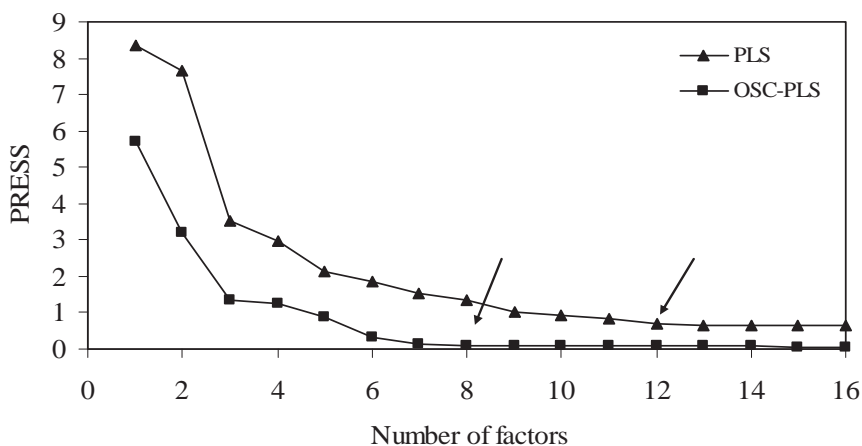


Fig. 2- Plots of PRESS versus number of factors by PLS and OSC-PLS.

LS-SVM analysis

The all descriptors were used as the input to develop nonlinear model by LS-SVM. The quality of LS-SVM for regression depend on γ and σ^2 parameters. In this work, LS-SVM was performed with radial basis function (RBF) as a kernel function. To determine the optimal parameters, a grid search was performed based on leave-one-out cross-validation on the original training set for all parameter combinations of γ and σ^2 from 1 to 1000 and 1 to 1000, respectively, with increment steps of 1. Table 2 shows the optimum γ and σ^2 parameters for the LS-SVM and RBF kernel, using the calibration sets.

Prediction of GC retention times of complex petroleum compounds

The predictive ability of these methods (PLS, OSC-PLS and LS-SVM) were determined using 6 retention time (their structure are given in Table 1). The results obtained by PLS, OSC-PLS and LS-SVM methods are listed in Table 2. Table 2 also shows RMSEP, RSEP and the percentage error for prediction of GC retention time of these compounds. As can be seen, the percentage error was also quite acceptable only for OSC-PLS and LS-SVM.

Table 2-Actual and predicted values of GC retention times of complex petroleum compounds using PLS, OSC-PLS and LS-SVM models.

Substance	Experimental retention time (s)	Predicted retention time (s)					
		PLS	Error (%)	OSC-PLS	Error (%)	LSSVM	Error (%)
2,3,4-Trimethylpentane	11.68	10.56	-9.58	11.23	3.85	11.67	-0.09
3-Methylpentane	9.45	8.13	-13.96	9.32	1.38	9.46	0.11
n-Butane	8.52	7.32	-14.08	8.34	2.11	8.53	0.12
n-Nonane	15.42	11.36	-26.33	15.69	-1.75	15.41	-0.06
n-Propylbenzene	16.38	14.68	-10.38	16.02	2.19	16.39	0.06
n-Undecane	20.72	18.63	-10.09	20.24	2.32	20.69	-0.14
N.F. ^a		12		8			
PRESS		0.6817		0.0981			
γ						10	
σ^2						110	
RMSEP		2.1671		0.3378		0.0152	
RSEP (%)		15.1130		2.3558		0.1065	

^a Number of factor

Good results were achieved in LS-SVM model with percentage error ranges from -0.14 to 0.12 for retention times. Also, it is possible to see that LS-SVM presents excellent prediction abilities when compared with other regression. According to the results, structural descriptors are suitable descriptors for describing the retention time of these compounds. When LS-SVM method with all descriptors is used, prediction of retention time in test step, with a small error is possible; this is improved in comparison with other method (PLS and OSC-PLS). This shows that by using all structural descriptors and also LS-SVM method, the retention time of complex petroleum compounds is predicted with satisfactory results.

Conclusion

A least squares-support vector machine (LS-SVM) model was established to predict the GC retention time of some organic compounds in petroleum sample. A proper model with high statistical quality and low prediction errors was obtained. The model could predict the GC retention time of organic compounds not existed in the modeling procedure accurately. The structural and topological descriptors concerning to the whole molecular properties and those of individual atoms in the molecule were found to be important factors controlling the retention time behavior.

Acknowledgment

The author gratefully acknowledges the support to this work from Young Researcher Club, Islamic Azad University, Arak Branch, research council.

References:

1. Moustafa, N.E., *Chromatographia*, **67**, 85 (2008).
2. Golfinopoulos, S.K., Lekkas, T.D. and Nikolaou, A.D., *Chemosphere*, **45**, 275 (2001).
3. Yazdanipour, A., *J. S. I. A. U.*, **18**, 50 (2008).
4. Niazi, A., Jameh-Bozorgi, S., and Nori-Shargh, D., *J. Hazard. Mat.*, **151**, 603 (2008).
5. Niazi, A., Jameh-Bozorgi, S., and Nori-Shargh, D., *Turk. J. Chem.*, **30**, 619 (2006).
6. Buyukbingol, E., Sisman, A., Akyildiz, M., Alparlan, D.N., Adejare, A., *Bioorg. Med. Chem.*, **15**, 4265 (2007).
7. Vapnik, V., *Statistical Learning Theory*, John Wiley, New York (1998).
8. Cortes, C., and Vapnik, V., *Mach. Learn.*, **20**, 273 (1995).
9. Suykens, J.A.K., and Vandewalle, J., *Neural Process. Lett.*, **9**, 293 (1999).
10. Suykens, J.A.K., Gestel, T., Brabanter, J., Moora, B., and Vandewalle, J., *Least-Squares Support Vector Machines*, World Scientifics, Singapore (2002).
11. Suykens, J.A.K., *Eur. J. Control*, **7**, 311 (2001).
12. Niazi, A., and Azizi, A., *Turk. J. Chem.*, **32**, 217 (2008).
13. Niazi, A., Ghasemi, J., and Yazdanipour, A., *Anal. Lett.*, **38**, 2377 (2005).
14. Niazi, A., Soufi, A., and Mobarakabadi, M., *Anal. Lett.*, **39**, 2359 (2006).
15. Niazi, A., *J. Braz. Chem. Soc.*, **17**, 1020 (2006).
16. Niazi, A., *Croa. Chem. Acta*, **79**, 573 (2006).
17. Niazi, A., Yazdanipour, A., *J. Hazard. Mat.*, **146**, 421 (2007).
18. Niazi, A., Ghasemi, J., and Zendehtdel, M., *Talanta*, **74**, 247 (2007).
19. Niazi, A., Ghasemi, J., and Yazdanipour, A., *Spectrochim. Acta Part A*, **68**, 523 (2007).
20. Niazi, A., Sharifi, S., and Amjadi, E., *J. Electroanal. Chem.*, **623**, 86 (2008).

21. Todeschini, R., and Consonni, V., *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim (2000).
22. Haaland, D.M., and Thomas, E.V., *Anal. Chem.*, **60**, 1193 (1988).

Archive of SID