Journal of the Serbian Chemical Society

JSCS-info@shd.org.rs • www.shd.org.rs/JSCS

J. Serb. Chem. Soc. *84* (4) 405–416 (2019)
JSCS–5193

UDC 635.71:539.2:539.16:665.5
Short communication

SHORT COMMUNICATION

Quantitative structure–retention relationship model for predicting retention indices of constituents of essential oils of *Thymus vulgaris* (Lamiaceae)

YOUSSOUF DRIOUCHE and DJELLOUL MESSADI*

Environmental and Food Safety Laboratory, Badji Mokhtar-Annaba University, BP.12, 23000 Annaba, Algeria

(Received 17 August, revised 6 December 2018, accepted 18 January 2019)

Abstract: In this paper, a quantitative structure–retention relationship (QSRR) model was developed for predicting the retention indices (log *RI*) of 36 constituents of essential oils. First, the chemical structure of each compound was sketched using HyperChem software. Then, molecular descriptors covering different information of molecular structures were calculated by Dragon software. The results illustrated that linear techniques, such as multiple linear regression (MLR), combined with a successful variable selection procedure are capable of generating an efficient QSRR model for predicting the retention indices of different compounds. This model, with high statistical significance ($R^2 = 0.9781$, $Q^2_{\rm LOO} = 0.9691$, $Q^2_{\rm ext} = 0.9546$, $Q^2_{\rm L(5)O} = 0.9667$, $F = 245.27$), could be used adequately for the prediction and description of the retention indices of other essential oil compounds. The reliability of the proposed model was further illustrated using various evaluation techniques: leave-5-out cross-validation, bootstrap, randomization test and validation through the test set.

Keywords: essential oils; retention indices; QSRR; multiple linear regression; *Thymus vulgaris* (Lamiaceae).

INTRODUCTION

Essential oils, a new approach to prevent the proliferation of microorganism or protection of food from oxidation, are ubiquitously used as antibacterial,[1–3] antifungal[3,4] and antioxidant agents.[5] They are also used to control human diseases of microbial origin and to cure diseases such as atherosclerosis and cancer[6] and are widely used in the food industry, medicine, and the fragrance industry. However, essential oils may exert toxic effects, such as human carcinogenicity, reproductive and developmental toxicity, neurotoxicity, and acute toxicity. The constituents present in essential oils have the potential to

*Corresponding author. E-mail: d_messadi@yahoo.fr

https://doi.org/10.2298/JSC180817010D

create serious and even fatal toxic effects if ingested in overly large quantities or used incorrectly.[7]

Essential oils have been used in folk medicine for thousands of years as antimicrobial agents. Therefore, the assessment of the gas chromatographic (GC) retention index (*RI*) of essential oils ingredients is a matter of great importance in the health of human beings.[8]

Seeking a quantitative relationship between molecular structure and the gas chromatographic retention indices has been a basic task in chemistry. Correlations between the GC retention indices and the molecular structures can provide more profound insights into the interactions between the eluents and the stationary phases from a theoretical viewpoint. In addition, they can provide very important information about the effect of the chemical structures on the retention behavior and the possible mechanism of absorption and elution.[8]

The study of quantitative structure–retention relationships (QSRRs) has been extensively used for predicting the GC retention. QSRR based on the experimental value of the retention index provides a promising faster way for predicting the retention index of essential oils using descriptors derived solely from the molecular structure. The advantage of a QSRR study over experimental methods lies in the fact that it requires only knowledge of the chemical structure. In addition, a QSRR model can be applied to predict the retention index of a new compound that belongs to the application domain of the model.[7]

Recently, several works that reported QSRR studies on the retention indices of essential oils have been published. Multiple linear regression (MLR) and partial least squares (PLS) models were built for the retention indices of 80 essential oils using the genetic algorithm (GA) to select the variables.[9] A nonlinear model based on the support vector machine (SVM) was also developed for 80 essential oils. Both linear and nonlinear models were used to predict 20 constituents in the test set. Liao *et al.*[10] reported an MLR model for predicting retention indices of 106 oxygen-containing organic compounds using the hydrogen-association classified molecular electronegativity-distance vector (H-MEDV) descriptors. Noorizadeh and Fermany,[11] and Noorizadeh *et al.*[12,13] built several QSRR models: (GA-MLR, GA-PLS, Kernel PLS and the Levenberg–Marquardt artificial neural network (L–M ANN)) for the retention indices of essential oils. The GA-MLR model for the prediction of the retention indices of 32 compounds was investigated by Azar *et al.*[14] Qin *et al.*[7] developed a QSRR model (based on the MLR method) for the prediction of the retention indices of 169 compounds in essential oils. Variable selections were performed by the ordered predictors selection (OPS) algorithm.[15] The reliability of the model was reviewed against the principles of the Organization for Economic Co-operation and Development (OECD).[16]

Quantitative and qualitative characterizations of the essential oils and volatiles from different parts of *Artemisia tschernieviana* plant were the main objectives of the works reported by Zanousi *et al.*[17] Additionally, a straightforward model was developed to model the *RI*s of diverse natural compounds present in the obtained essential oils and volatile fractions based on a simple stepwise multiple linear regression (SW-MLR) approach.

A robust screening approach and a sparse QSRR model for predicting the *RIs* of 169 constituents of essential oils, obtained from Conforti *et al.*,[18] was reported by Al-Fakih *et al.*[19] The proposed method consists of two basic steps. The first step is dimension reduction, in which data are reduced from high dimensional space to a lower d-dimensional descriptor space. The second step is prediction, in which the response variable is predicted using a sparse method on the screened descriptors.

In the particular case of essential oils, despite the vast literature on the identification and characterization of novel constituent compounds, and apart from the study presented by Marrero-Ponce *et al.*[20] involving a database of 791 essential oil components with corresponding gas chromatographic retention properties, which seems to be an exception, the majority of QSRR models have generally been built on data sets of sizes ranging from 25 to 169 compounds, with most of these belonging to a single chemical series.

The aim of the present study was to develop a QSRR model for RI prediction of 36 essential oil components, representing chemical variation of leaf essential oil, at different stages of *Thymus vulgaris* (Lamiaceae) growth of Iranian plants.

EXPERIMENTAL

Experimental data

The essential oil composition was analyzed using a Shimadzu QP 5050 GC/MS with DB-5 capillary column.[21] *n*-Alkanes mixtures were analyzed under the GC/MS temperature condition program to calculate the Kovats *RI*s of the thirty-six identified components in the oil and are listed in Table I. The data are presented as the logarithm of *RI* to reduce the range of variation.

Descriptors generation

The chemical structure of each compound was sketched on a PC using the HyperChem program[22] and pre-optimized using the MM$^+$ molecular mechanics method (Polack–Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi empirical PM3 method at the restricted Hartree–Fock level with no configuration interaction applying a gradient norm limit of 0.01 kcal* Å^{-1} mol^{-1} as a stopping criterion.

The resulting geometries were used as the input for the generation of 74 three-dimensional-geometrical descriptors using the Dragon software (version 6).[23] These are the molecular descriptors defined in several different ways but always derived from the three--dimension structure of the molecules.

* 1 kcal = 4184 J

TABLE I. The data set and the corresponding observed and predicted values of log *RI* by MLR for the training and test sets

ID	Object	log *RI* (Exp.)	log *RI* (Calcd.)	log *RI* (Pred.)	*HAT* (h_{ii})	$e_{i\mathrm{std}}$
1	*α*-Thujene	2.98	2.9944	2.9988	0.237	1.7166
2	β-Pinene	3.00	2.9896	2.9864	0.232	–1.2321
3	*β*-Myrcene	3.00	2.9958	2.9948	0.203	–0.4673
4	*o*-Cymene	3.02	3.0409	3.0440	0.128	2.0480
5	1,8-cineole	3.03	3.0355	3.0363	0.123	0.5323
6	*cis*-*β*-Terpineol	3.04	3.0539	3.0560	0.134	1.3702
7	Terpinolene	3.04	3.0247	3.0211	0.189	–1.6718
8	Linalool	3.05	3.0578	3.0598	0.213	0.8836
9	Isopulegol	3.05	3.0610	3.0639	0.212	1.2472
10	Camphor	3.07	3.0746	3.0761	0.250	0.5629
11	Isoborneol	3.08	3.0490	3.0450	0.113	–2.9546
12	Thymol methyl ether	3.10	3.1038	3.1062	0.386	0.6269
13	Verbenone	3.10	3.1069	3.1099	0.304	0.9443
14	Dihydrocarvone	3.11	3.0935	3.0891	0.212	–1.8779
15	Thymol	3.14	3.1350	3.1347	0.058	–0.4323
16	Eugenol	3.15	3.1500	3.1500	0.507	–0.0008
17	*β*-Caryophyllene	3.17	3.1623	3.1617	0.075	–0.6879
18	*α*-Bergamotene	3.16	3.1701	3.1712	0.101	0.9418
19	Germacrene D	3.18	3.1785	3.1782	0.149	–0.1563
20	*γ*-Elemene	3.18	3.1769	3.1766	0.071	–0.2801
21	*β*-Bisabolene	3.19	3.1735	3.1713	0.117	–1.5873
22	*δ*-Cadinene	3.19	3.1908	3.1910	0.157	0.0863
23	Caryophyllene oxide	3.20	3.2100	3.2116	0.138	0.9970
24	Spathulenol	3.21	3.2103	3.2104	0.202	0.0323
25	Aromadendrene oxide	3.21	3.2001	3.1985	0.139	–0.9860
26	Muurolol	3.22	3.2280	3.2298	0.180	0.8614
27	Bisabolol	3.22	3.2233	3.2240	0.168	0.3452
28[a]	α-Pinene	2.98	2.9964	2.9980	0.236	1.6432
29[a]	Camphene	2.99	2.9838	2.9858	0.213	–0.3784
30[a]	*α*-Phellandrene	3.01	3.0059	3.0066	0.194	–0.3043
31[a]	*α*-Terpinene	3.01	3.0239	3.0245	0.203	1.2934
32[a]	*γ*-Terpinene	3.04	3.0190	3.0198	0.189	–1.7839
33[a]	Borneol	3.07	3.0429	3.0456	0.121	–2.0738
34[a]	4-Terpineol	3.08	3.0762	3.0772	0.116	–0.2341
35[a]	*α*-Terpineol	3.11	3.0843	3.0850	0.159	–2.1743
36[a]	Cadinol	3.22	3.2330	3.2338	0.216	1.2405

[a] Test compounds

Training set and test set

The present challenge in the process of the development of a QSAR model lies in the development of a model with the capability to accurately predict the activity of new chemicals.

The most effective method of validating a regression model with respect to its prediction performance is to collect fresh data and directly compare the model predictions against them. When this is not possible, a reasonable procedure is to split the available data into two parts: a training set from which the model is built and an external set on which to evaluate its predictive power, the later should contain between 15 and 40 % of the compounds in the full data set.[24] Several procedures can be adopted for the selection of the training and test sets. In this work, the Kennard and Stone algorithm (CADEX)[25] was used to arbitrarily split the whole data into 27 samples training set and 9 samples testing set (Table I).

MLR Modeling

The general purpose of multiple regressions is to quantify the relationship between several independent or predictor variables and a dependent variable. A set of coefficients defines the simple linear combination of independent variables (molecular descriptors) that best describes the retention indices. The value of the retention indices for each essential oil component would then be calculated as a composite of each molecular descriptor weighted by the respective coefficients. A multi-linear regression model with *k* regressors could be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \tag{1}$$

The term linear is used because Eq. (1) is a linear function of the unknown parameters β_j, $j = 0,1,\ldots,k$, called the regression coefficients. The parameter β_j represents the expected change in the response y per unit change in x_j when all of the remaining regressor variables x_i ($i \neq j$) are held constant.

It is more convenient to deal with multiple regression models if they are expressed in matrix notation. This allows a very compact display of the model data and results. In matrix notation, the model given by Eq. (1) is:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2}$$

In general, $\boldsymbol{y}$ is an $n\times1$ vector of the observations, $\boldsymbol{X}$ is an $n \times p$ matrix of the levels of the regressor variables, $\boldsymbol{\beta}$ is a $p\times1$ vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n\times1$ vector of random errors.

When $\boldsymbol{X}$ is of full rank, the least-squares solution is $\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, where $\boldsymbol{b}$ is the estimator for the regression coefficients in $\boldsymbol{b}$ and $\boldsymbol{X}^{\mathrm{T}}$ is a transpose of $\boldsymbol{X}$.

The vector of the fitted values $\hat{\boldsymbol{y}}_i$ corresponding to the observed values y_i is:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{H}\,\boldsymbol{y} \tag{3}$$

The $n\times n$ matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}\boldsymbol{X}^{T}$ is usually called the “hat” matrix because it maps the vector of the observed values into a vector of fitted values. The diagonal elements are often called the leverages, since examination of:

$$\hat{y} = h_{ii} y_i + \sum_{j\neq i} y_j \tag{4}$$

indicates *via* h_{ii} how heavily y_i contributes to $\hat{y}_i$.

Chemometric methods

Once the molecular descriptors are generated, multiple linear analysis regression and variable subset selection were performed by the software MobyDigs[26] using the ordinary least square (OLS) regression method.

From a statistical viewpoint, the ratio of the number of samples (n) to the number of descriptors (m) should not be too low. Usually, it is recommended that $n/m \geq 5$.[27]

First, models with 1–2 variables were developed by the all-subset-method procedure in order to explore all the low dimension combinations. The number of descriptors was subsequently increased one by one, thereby forming new models. The best models were selected at each rank, and the final model must be chosen from among them.

Population of 100 regression models were ranked according to their decreasing internal predictive performance, verified by Q^2, and the optimal model was then selected. The goodness-of-fit of the calculated model was assessed by means of the multiple determination coefficient, R^2, and the standard deviation error in calculation (*SDEC*):

$$SDEC=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{5}$$

Cross-validation techniques allow an assessment of internal predictivity (Q^2_{LMO} cross--validation; bootstrap) in addition to the robustness of model (Q^2_{LOO} cross-validation).

Cross-validation methods consist in leaving out a given number of compounds from the training set and rebuilding the model, which is then used to predict the compounds left out. This procedure is repeated for all compounds of the training set, thereby obtaining a prediction for each one. If each compound is taken away one at a time, the cross-validation procedure is called the leave-one-out technique (LOO technique), otherwise it is a leave-many-out technique (LMO technique). An LOO or LMO correlation coefficient, generally indicated with Q^2, is computed by evaluating the accuracy of these “test” compounds prediction.

$$Q^2=10-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_{i/i})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=1-\frac{PRESS}{TSS} \tag{6}$$

The “hat” of the variable *y*, as is the usual statistical notation, indicates that it is a predicted value of the studied property and the sub index “*i/i*” indicates that the predicted values come from models built without the predicted compound. *TSS* is the total sum of squares.

The predictive residual sum of squares (*PRESS*) measures the dispersion of the predicted values. It is used to define Q^2 and the standard deviation error in prediction (*SDEP*):

$$SDEP\ =\ \sqrt{\frac{PRESS}{n}} \tag{7}$$

A value $Q^2 > 0.5$ is generally regarded as a good result and $Q^2 > 0.9$ as excellent.[28,29] However, studies[30,31] have indicated that while Q^2 is a necessary condition for high predictive power of a model, it is not sufficient. To avoid overestimating the predictive power of the model, the LMO procedure (repeated 5000 times, with 5 objects left out at each step) was also performed ($Q^2_{\text{L(5)O}}$).

In the bootstrap validation technique, *K* *n*-dimensional groups are generated by a randomly repeated selection of *n* objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 8000 times for each validated model.[32] By using the selected model, the values of the response for the test objects were calculated and the quality of these predictions is defined in terms of Q^2_{ext}, which is defined as:

$$Q^2_{\text{ext}} = 1 - \frac{\sum_{i=1}^{n_{\text{ext}}} (\hat{y}_{i/i} - y_i)^2 / n_{\text{ext}}}{\sum_{i=1}^{n_{\text{tr}}} (y_i - \bar{y}_{\text{tr}})^2 / n_{\text{tr}}} = 1 - \frac{PRESS / n_{\text{ext}}}{TSS / n_{\text{tr}}} \tag{8}$$

where n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively.

Another useful parameter is the external standard deviation error of prediction ($SDEP_{\text{ext}}$), defined as:

$$SDEP_{\text{ext}} = \sqrt{\frac{1}{n_{\text{ext}}} \sum_{i=1}^{n_{\text{ext}}} \left(y_i - \bar{y} \right)^2} \tag{9}$$

where the sum runs over the test set objects (n_{ext}).

Applicability domain analysis

The applicability domain (AD)[29,33] is a theoretical region in space defined by the descriptors of the model and the modeled response, for which a given QSRR should make reliable predictions. In this work, the structural AD was verified by the leverage approach. The leverage, $\boldsymbol{h}_{ii}$,[32] is defined as follows:

$$\boldsymbol{h}_{\boldsymbol{ii}} = \boldsymbol{x}_i^{\text{T}} \left(\boldsymbol{X}^{\text{T}} \boldsymbol{X} \right)^{-1} \boldsymbol{x}_i \quad \left(i = 1;...;n \right) \tag{10}$$

where $\boldsymbol{x}_i$ is the descriptor row vector of the i^{th} compound, $\boldsymbol{x}^{\text{T}}_i$ is the transpose of $\boldsymbol{x}_i$, $\boldsymbol{X}$ is the model matrix derived from the calibration set descriptor values.

The warning leverage, h^* is, generally, fixed at $3(m+1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation.

RESULTS AND DISCUSSION

A major step in the construction of QSRR model is finding a set of molecular descriptors that represent the variation in the structural properties of the molecules. The modeling and prediction of the physicochemical properties of organic compounds is an important objective in many scientific fields[13]. MLR is one of the most modeling methods in QSRR. MLR method provides equation linking the structural features to the (log *RI*) of the compounds. The selected equation is:

$$\begin{aligned} \log RI \;=\; & 1.16 + 0.291 piPC02 + 1.94\, ChiA\_B(p) + \\ & + 0.0992\, SM2\_B(s) - 0.0342 Mor15u \end{aligned} \tag{11}$$

$R^2 = 0.9781$; $Q^2_{\text{LOO}} = 0.9691$; $Q^2_{\text{L(5)O}} = 0.9667$; $Q^2_{\text{ext}} = 0.9546$; $Q^2_{\text{Boot}} =$ $= 0.9592$; $SDEC = 0.011$; $SDEP = 0.013$; $SDEP_{\text{ext}} = 0.016$; $s = 0.013$ log unit; $F = 245.27$ ($p = 0.000$)

From the above equation, it could be concluded that the most significant descriptors (Table S-I of the Supplementary material) according to the MLR method are molecular multiple path count of order 02 (*piPC*02), average Randic-like index from the Burden matrix weighted by polarizability (*ChiA_B(p)*), spectral moment of order 2 from the Burden matrix weighted by I-state (*SM2_B(s)*) and 3D-MoRSE – signal 15/unweighted (*Mor15u*). A detailed description of the linear model based on compounds in the training set is summarized in Table II.

TABLE II. Selected descriptors of multiple linear regression

No.	Symbol	Descriptor description	Group descriptor	Coefficient	*VIF*
–	Constant	–	–	1.1622 (±0.1293)	–
1	*piPC*02	Molecular multiple path count of order 02	Walk and path counts	0.2906 4 (±0.0163)	1.950
2	*ChiA_B(p)*	Average Randic-like index from Burden matrix weighted by polarizability	2D matrix-based descriptors	1.9410 (±0.3718)	1.640
3	*SM2_B(s)*	Spectral moment of order 2 from Burden matrix weighted by I-state	2D matrix-based descriptors	0.0992 (±0.0121)	1.334
4	*Mor15u*	3D-MoRSE – signal 15/unweighted	3D-MoRSE descriptors	–0.0342 (±0.0049)	1.216

The multi-collinearities between the above four descriptors were detected by their variation inflation factors (*VIF*), which can be calculated as follows:

$$VIF = \frac{1}{1-r^2} \tag{11}$$

where r is the correlation coefficient of the multiple regression between the variables in the model. If *VIF* equals 1, then no inter-correlation exists for each variable; if *VIF* falls into the range of 1–5, the related model is acceptable and if *VIF* is larger than 10, the related model is unstable and a recheck is necessary. The corresponding *VIF* values of the four descriptors are given in Table II. As can be seen from this table, the variables had *VIF* values of less than 5, indicating that the obtained model has statistical significance.[8]

The results for the randomized models can be compared with the real starting one only by representation in a plot of the statistical coefficients R^2 and Q^2. This is depicted in Fig. 1. The statistics for the modified log *RI* vectors are clearly lower than the real QSRR model. This ensures that a real structure–property relationship has been found.

The value of $R^2 = 0.9781$ attests the good fitting performances of the model. In general, the larger the magnitude of the F ratio, the better the model predicts the property values in the training set. The large F ratio of 245.27 indicates that the model performs excellently in predicting the log *RI* values. The model is robust, the difference between R^2 and Q^2 is small (< 0.5 %).

Plots of cross-validation of log *RI* values *vs*. experimental log *RI* values are shown in Fig. 2. Obviously, there is a close agreement between the experimental and predicted log *RI* values and the data present very low scattering around a straight line with respective slope and intercept values close to one and zero, respectively.

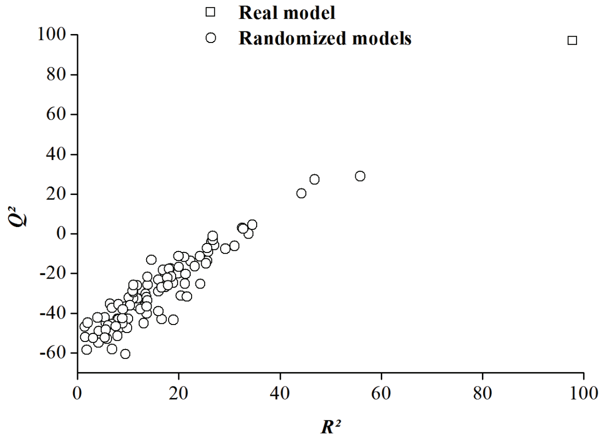

Fig. 1. Randomization test associated to previous QSRR models.

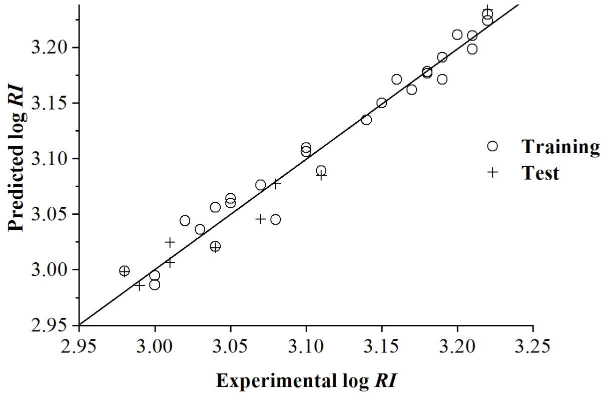

Fig. 2. Cross-validation *vs*. experimental log *RI*.

SDEP is similar to *SDEC* and hence, this model has internal predictivity not very dissimilar from the fitting power. The model demonstrates a very good stability in internal validation (the difference between Q^2 and $Q^2_{\rm L(5)O}$ is 0.0024 %), while bootstrapping confirms the internal predictivity and stability of the model.

The information obtained by $Q^2_{\rm ext}$ is somewhat optimistic. In fact, with small data sets (20–40 compounds), completely new compound external predictivity could only be verified *a posteriori*, case-by-case.

The AD of the linear model was analyzed in a Williams plot (Fig. 3), the plot of standardized residuals ($e_{\rm istd}$) *vs*. leverage values. Table I (column 7) shows samples with absolute standardized residual values smaller than 3 standard deviation units, and so, no *Y*-outlier was detected.

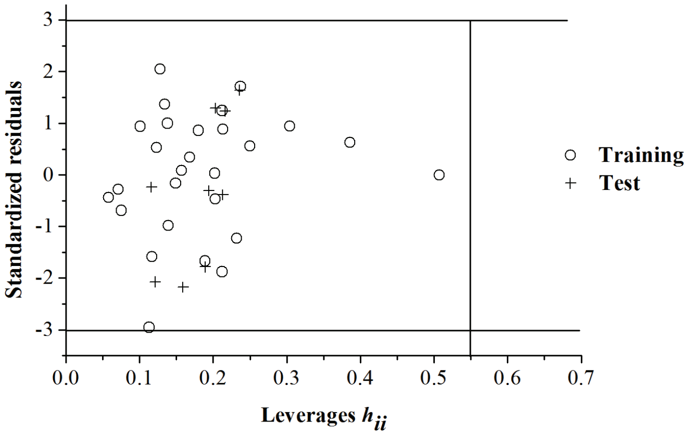

Fig. 3. The Williams plot of the data.

CONCLUSIONS

A QSRR study was performed for predicting the gas chromatographic *RI* of 36 essential oil compounds on a DB-5 stationary phase with low polarity, in which case the chromatographic retention is governed by the geometry and the size of the solute molecules.[34] Hence, only geometrical descriptors were considered as explicative variables.

The optimal model is a multivariate linear model, which has four variables, selected using the MLR technique. A detailed validation procedure demonstrated the accuracy and robustness of the proposed model not only by calculating its fitness on the training set, but also by testing its predictive ability. The QSRR model with simply calculated molecular descriptors could be employed to estimate the retention index for new compounds.

ИЗВОД

МОДЕЛ КВАНТИТАТИВНЕ РЕЛАЦИЈЕ СТРУКТУРЕ И ИНДЕКСА РЕТЕНЦИЈЕ ЗА САСТОЈКЕ ЕТАРСКОГ УЉА ИЗ *Thymus vulgaris* (Lamiaceae)

YOUSSOUF DRIOUCHE и DJELLOUL MESSADI

Environmental and Food Safety Laboratory, Badji Mokhtar-Annaba University, BP.12, 23000 Annaba, Algeria

У овом чланку, развијен је модел квантитативне релације структуре и индекса ретенције (QSRR) за предвиђање ретенционих индекса (log *RI*) за 36 састојка етарских уља. Прво је хемијска структура сваког једињења скицирана коришћењем HyperChem софтвера. Онда су, коришћењем Dragon софтвера, израчунати молекулски дескриптори који покривају различне информације молекулских структура. Резултати су показали да су линеарне технике, попут вишеструке линеарне регресије (MLR) у комбинацији са успешном процедуром избора променљивих, у стању да генеришу успешан QSRR модел за предвиђање ретенционих индекса различних једињења. Овај модел са високом статистичком значајношћу (R^2 = 0,9781, Q^2_{LOO} = 0,9691, Q^2_{ext} = 0,9546, $Q^2_{\mathrm{L(5)O}}$ = 0,9667, F = 245,27), може се адекватно користити за предвиђање и описивање ретенционих индекса једињења и у другим етарским уљима. Поузданост предложеног модела је даље илустрована коришћењем разних техника процењивања: "leave-5-out" унакрсна валидација, "bootstrap", тест насумичности ("randomization test") и валидација преко скупа за проверу.

(Примљено 17.августа 2018, ревидирано 6. децембра 2018, прихваћено 18. јануара 2019)

REFERENCES

1. S. Burt, *Int. J. Food Microbiol*. **94** (2004) 223 (https://doi.org/10.1016/j.ijfoodmicro.2004.03.022)
2. S.-T. Chang, P.-F. Chen, S.-C. Chang, *J. Ethnopharmacol*. **77** (2001) 123 (https://doi.org/10.1016/S0378-8741(01)00273-2)
3. D. Kalemba, A. Kunicka, *Curr. Med. Chem*. **10** (2003) 813 (https://doi.org/10.2174/0929867033457719)
4. C. L. Wilson, J. M. Solar, A. El Ghaouth, M. E. Wisniewski, *Plant Dis*. **81** (1997) 204 (https://doi.org/10.1094/PDIS.1997.81.2.204)
5. M. Burits, F. Bucar, *Phytother. Res*. **14** (2000) 323 (https://doi.org/10.1002/1099-1573(200008)14:5<323::AID-PTR621>3.0.CO;2-Q)
6. P. H. Warnke, E. Sherry, P. A. J. Russo, Y. Acil, J. Wiltfang, S. Sivananthan, M. Sprengel, J. C. Roldàn, S. Schubert, J. P. Bredee, I. N. G. Springer, *Phytomedicine* **13** (2006) 463 (https://doi.org/10.1016/j.phymed.2005.09.012)
7. L.-T. Qin, S.-S. Liu, F. Chen, Q.-F. Xiao, Q.-S. Wu, *Chemosphere* **90** (2013) 300 (https://doi.org/10.1016/j.chemosphere.2012.07.010)
8. M. Rahimi, H. Farahbakhsh, N. Salehi, M. Nekoei, *Int. J. Adv. Appl. Sci.* **1** (2012) 91 (https://www.iaescore.com/journals/index.php/IJAAS/article/view/775)
9. S. Riahi, E. Pourbasheer, M. R. Ganjali, P. Norouzi, *J. Hazard. Mater.* **166** (2009) 853 (https://doi.org/10.1016/j.jhazmat.2008.11.097)
10. L. Liao, D. Qing, J. Li, G. Lei, *J. Mol. Struct.* **975** (2010) 389 (https://doi.org/10.1016/j.molstruc.2010.05.017)
11. H. Noorizadeh, A. Farmany, *Chromatographia* **72** (2010) 563 (https://doi.org/10.1365/s10337-010-1660-4)
12. H. Noorizadeh, A. Farmanya, A. Khosravi, *J. Chin. Chem. Soc*. **57** (2010) 982 (https://doi.org/10.1002/jccs.201000137)
13. H. Noorizadeh, A. Farmany, M. Noorizadeh, *Quim. Nova* **34** (2011) 242 (http://dx.doi.org/10.1590/S0100-40422011000200014)
14. P. A. Azar, M. Nekoei, R. Siavash, M. R. Ganjali, K. Zare, *J. Serb. Chem. Soc*. **76** (2011) 891 (https://doi.org/10.2298/JSC100219076A)
15. R. F. Teofilo, J. P. A. Martins, M. M. C. Ferreira. *J. Chemom.* **23** (2009) 32 (https://doi.org/10.1002/cem.1192)
16. OECD, *Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q)SAR] Models*, Paris, 2007 (https://doi.org/10.1787/9789264085442-en)
17. M. B. P. Zanousi, M. Nekoei, M. Mohammadhosseini. *J. Essent. Oil-Bear. Plants* **20** (2017) 672 (https://doi.org/10.1080/0972060X.2017.1329669)
18. F. Conforti, F. Menichini, C. Formisano, D. Rigano, F. Senatore, N. A. Arnold, F. Piozzi, *Food. Chem.* **116** (2009) 898 (https://doi.org/10.1016/j.foodchem.2009.03.044)
19. A. M. Al-Fakih, Z. Y. Algamal, M. H. Lee, M. Aziz, *SAR QSAR Environ. Res*. **28** (2017) 691 (http://dx.doi.org/10.1080/1062936X.2017.1375010)

20. Y. Marrero-Ponce, S. J. Barigye, M. E. Jorge-Rodriguez, T. Tran-Thi-Thu, *Chem. Pap.* **72** (2018) 57 (https://doi.org/10.1007/s11696-017-0257-x)
21. A. Nezhadali, M. Nabavi, M. Rajabian, M. Akbarpour, P. Pourali, F. Amini, *Beni-Seuf Univ. J. Appl. Sci*. **3** (2014) 87 (https://doi.org/10.1016/j.bjbas.2014.05.001)
22. HyperChemTM. Release 6.02 for Windows, *Molecular Modeling system*, 2000 (http://www.hyper.com/)
23. TALETE srl, Dragon (Software for Molecular Descriptors Calculation) version 6.0, 2011 (http://www.talete.mi.it/)
24. E. Benfenati, J. R. Chrétien, G. Gini, N. Piclin, M. Pintore, A. Roncaglioni, *Validation of the models, in Quantitative Structure–Activity Relationships (QSAR) for Pesticide Regulatory Purposes,* Elsevier, Amsterdam, 2007, pp. 185–199 (https://doi.org/10.1016/B978-044452710-3/50008-2)
25. R. W. Kennard, L. A. Stone, *Technometrics* **11** (1969) 137 (https://doi.org/10.1080/00401706.1969.10490666)
26. R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Paven, MobyDigs, version 1.1, Copyright TALETE srl, 2009 (http://www.talete.mi.it/)
27. J. Xu, H. Zhang, L. Wang, G. Liang, L. Wang, X. Shen, W. Xu, *Spectrochim. Acta, Part A* **76** (2010) 239 (https://doi.org/10.1016/j.saa.2010.03.027)
28. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, P. Gramatica, *Environ. Health Perspect*. **111** (2003) 1361 (https://www.ncbi.nlm.nih.gov/pubmed/12896860)
29. A. Tropsha, P. Gramatica, V. K. Grombar, *QSAR Comb. Sci*. **22** (2003) 69 (https://doi.org/10.1002/qsar.200390007)
30. H. Kubinyi, F. A. Hamprecht, T. Mietzner, *J. Med. Chem*. **41** (1998) 2553 (https://doi.org/10.1021/jm970732a)
31. A. Golbraikh, A. Tropsha, *J. Mol. Graphics Modell*. **20** (2002) 269 (https://doi.org/10.1016/S1093-3263(01)00123-1)
32. B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans,* Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982 (http://dx.doi.org/doi:10.1137/1.9781611970319)
33. M. Shen, C. Béguin, A. Golbraikh, J.P. Stables, H. Kohn, A. Tropsha, *J. Med. Chem*. **47** (2004) 2356 (https://pubs.acs.org/doi/abs/10.1021/jm030584q)
34. R. Kaliszan, *Quantitative* structure–*chromatographic retention relationships*, Wiley, New York, 1987 (https://www.osti.gov/biblio/6478095).