Environ Chem Lett (2008) 6:175–181
DOI 10.1007/s10311-007-0119-9

ORIGINAL PAPER

Modelling the property of compounds from structure: statistical methods for models validation

Sorana-Daniela Bolboacă · Lorentz Jäntschi

Received: 21 June 2007 / Accepted: 12 September 2007 / Published online: 16 October 2007

Abstract This study discusses some statistical methods used in quantitative structure–property relationships model's validation and comparison. The paper also introduces a series of online tools for model's validation. The implemented diagnostic tool that incorporated analytical approaches are exemplified on a series of models which estimate and predict the boiling points of a sample of 73 alkanes. The used parameters and methods are presented, and some reference values are provided.

Keywords Structure–property relationships (SPR) · Molecular descriptors family (MDF) · Model validation · Online application

Introduction

Structure–property relationships (SPR) and quantitative structure–property relationships (QSPRs), collectively referred to as (Q)SPRs, are theoretical models usually used to predict the physicochemical properties of chemicals from the knowledge of structure (Rogers and Hopfinger 1994). The method has been used since 1868, when Crum-Brown and Fraser (1868) studied the physiological action of the ammonium salts. Twenty-five years later, Richet studied the relationship between chemical structure and oil–water partition coefficient (Richet 1893). Since then, many properties were modelled using different approaches (Hemmateenejad 2006; Buchwald and Bodor 2002).

The processes by which the reliability (yielding the same or compatible results for a compounds or class of compounds by applying the same method) and the relevance (the ability to correctly predict the property of interest) of a QSPR model is tested are known as compulsory models validation methods (Worth and Cronin 2004). The main statistical methods used for models validation are (Tichy 2005): regression analysis, factor analysis, principal component analysis, cluster analysis, pattern recognition analysis, discriminant analysis, artificial neuron networks, and cybernetic techniques (self-learning machine, genetic algorithms, linear surface decision). The cross-validation analysis techniques, as leave-one-out (LOO) analysis, leave-many-out analysis (Tropsha et al. 2000) and bootstrapping (Wehrens et al. 2000) are also used for models validation.

The main goal of the paper was to highlight and to exemplify the analytical methods transposed into online programs useful in (Q)SPR models assessment. In so doing, a data set of 73 alkanes were investigated by analyzing one previously reported SPR models (Toropov et al. 1998) and two models obtained on the same data set by applying of the molecular descriptor family on the structure–property relationships approach (Jäntschi 2005).

Experimental

Models for alkanes boiling points

Toropov et al. reported a series of nine QSPRs obtained by using the 3D weighting of molecular descriptors (Toropov et al. 1998) which analyzed the experimental boiling points

S.-D. Bolboacă (✉)
"Iuliu Hatieganu" University of Medicine and Pharmacy,
6 Louis Pasteur, 400349 Cluj-Napoca, Romania
e-mail: sbolboaca@umfcluj.ro

L. Jäntschi
Technical University of Cluj-Napoca, 103–105 Muncii Bdv,
400641 Cluj-Napoca, Romania

of a series of 73 alkanes (all alkanes isomers with three, four, five, six, seven, eight and nine carbons). The best performing model and associated statistics are presented in Eq. (1):

$$\begin{aligned} \mathrm{Bp}(^\circ\mathrm{C}) &= 727.26(\pm 20.76)\cdot 3\mathrm{D}0\chi - 19.46(\pm 0.9)\cdot 3\mathrm{DSRW2} \\ &\quad + 7.99(\pm 0.39)\cdot \mathrm{M2} - 779.42(\pm 20.08) \\ & n = 73; r = 0.9986; s = 2.17; F = 8340 \end{aligned} \tag{1}$$

where Bp(°C) is the boiling point, $3\mathrm{D}0\chi$ and 3DSRW2 are method of ideal symmetry (MIS) indices, and M2 is a 3D modification of the Zagreb index; *n* the volume of the sample, *r* the correlation coefficient, *s* standard error of the estimate, and *F* Fisher parameter.

The same series of alkane's experimental boiling points has been investigated by using the molecular descriptors family on the structure–property relationships approach (Jäntschi 2005). The MDF-SPR approach proved its usefulness in estimation and prediction of a series of activities and properties (Jäntschi and Bolboaca 2007). The MDF-SPR models and associated statistics are presented in Eqs. (2) and (3):

$$\begin{aligned} &\hat{Y}_{1\mathrm{D}} = -507.95 + 188.40 \cdot lbMdsHg \\ &r^2_{\mathrm{MDF}} = 0.9913;\ F_{\mathrm{MDF}} = 8048\,(p < 0.0001); \\ &s_{\mathrm{MDF}} = 3.81,\ n = 73 \end{aligned} \tag{2}$$

$$\begin{aligned} &\hat{Y}_{2\mathrm{D}} = -323.02 - 105.92 \cdot liDmEHt + 17.76 \cdot IADmwHt \\ &r^2_{\mathrm{MDF}} = 0.9982;\ F_{\mathrm{MDF}} = 19344\,(p < 0.0001); \\ &s_{\mathrm{MDF}} = 1.75,\ n = 73 \end{aligned} \tag{3}$$

where $\hat{Y}_{1\mathrm{D}}$ is the estimated boiling point by the MDF-SPR model with one descriptor; $\hat{Y}_{2\mathrm{D}}$ is the estimated boiling point by the MDF-SPR model with two descriptors; lbMdsHg, liDmEHt and IADmwHt are molecular descriptors generated and calculated based on information extracted from alkanes chemical structure; r^2_{MDF} squared correlation coefficient of the MDF SPR model; F_{MDF} Fisher parameter at a significance level of 5%; s_{MDF} standard error of the estimate; and *n* sample size.

Models assessment methodology: evaluation, validation, comparison

The following approaches were proposed and implemented for models validation and comparison. Software is hosted by AcademicDirect domain and could be freely used from http://l.academicdirect.org.

1. Evaluation:
   - Correlation coefficient:

     It is a simple statistical measure of the relationship between dependent variable and one or more independent variables. According with the type of experimental data, one of the followings methods could be applied: Pearson, Spearman, semi-quantitative, Kendall's tau (*a*, *b*, or *c*), or Gamma (Bolboaca and Jäntschi 2006).

     It takes value between –1 and +1. The value of zero indicates no relationship while values of ±1 indicate a perfect fit. A value greater than 0.5 is generally consider as good, while a value greater than 0.9 is consider as excellent.

     It is use as measure of the statistical fit of a regression-based model, but the preferred form is its squared value (coefficient of determination –r^2). The closer the value of determination coefficient is to 1, the better the model is.

     It is a measure of collinearity: linear relationship exists among some or all independent variables in a regression model. Collinearity and correlation are not equivalent concepts; thus, collinearity implies correlation, but correllation does not necessary implies collinearity.

     Implemented software: see /Chemistry/SARs/MDF_SARs/rank/ and /Statistics/linear_dependence/ (from http://l.academicdirect.org).

   - Linear regression: The parameters used to analyze a linear regression model are:

     Correlation coefficient, determination coefficient (the proportion of variation in dependent variable that is explained by its linear relationship with independent variables)—see above

     Adjusted correlation coefficient

     Standard error of the estimate (the average error predicting the dependent variable by means of the regression equation)

     Fisher parameter and associated significance: give information about statistical significance of the regression model

     Student's *t* Stat (refers to a test of the hypothesis that the regression coefficients and slope are significantly different from zero)

     Implemented software: /Statistics/multi_regression/

2. Validation:
   - Leave-one-out analysis:

     An internal cross validation technique

     Employ *n* training sets and from each of these one compound is excluded

  For each training set a model is obtained and then it is used to predict the property/activity of excluded compound

  The cross validation LOO score ($r^2_{\text{cv-loo}}$) is obtained; for its interpretation see coefficient of determination

  The difference between coefficient of determination (r^2) and cross validation LOO score ($r^2_{\text{cv-loo}}$) ought not to exceed 0.3. A difference greater than or equal to 0.3 could indicate: an overfitted model, the presence of irrelevant independent variables, and/or the presence of outliers (an observation that lies an abnormal distance from other values in a sample)

  Implemented software: /Chemistry/SARs/MDF_SARs/loo/

- Training versus test analysis:

  A validation technique

  The compounds are randomly spitted in training and test sets

  For each training set a model is obtained and is used to predict the property/activity of compounds from test set

  The model is consider valid and stable if the determination coefficient on training set (r2tr) is not statistically different by the determination coefficient on test set (r2ts) and the values of the correlation coefficients respect the 95% confidence intervals of the squared correlation coefficient of the model

  Implemented software: /Chemistry/SARs/MDF_SARs/qsar_qspr_s/

3. Comparison:
   - Correlated correlation analysis:

     A method of comparison two correlation coefficients taking into account the sample sizes on which those were obtained (Steiger 1980)

     Hotteling's *t*/Steiger's *Z* test: test the correlated correlation at a significance level of 5%

     Fisher's *Z*-test: test differences between correlation coefficients obtained on two different groups (at a significance level of 5%)

     The statistical approaches were applied on the models from Eqs. (2) and (3), and where appropriate from Eq. (1) too.

Results and discussion

Models evaluation (regression coefficients, residuals, other statistics)

Depending on the type of analytical techniques used (Q)SPR analysis had as results a set of parameters that provide information about the model(s). By using the linear regression techniques, regression equations consisting of coefficients are produced. The coefficients had a simple and appealing meaning. Note that, in accordance with the choice of regression method, there are other parameters that deserve attention when a (Q)SPR model is interpreted.

Let us note *Y* as the endpoint of interest (in our case the boiling point of alkanes) and as *X* the variables used in Eqs. (1)–(3). The main regression coefficients, residuals, and model statistics performance for models presented in are presented in second and third equations Table 1.

Analyzing the model from Eq. (3) it can be observed that the complete data set comprises 73 compounds, two *X* variable (two molecular descriptors, *liDmEHt* and *IADmwHt*, respectively) and one *Y* variable (the boiling point). The two *X* variable are not significantly correlated with each other, the squared correlation coefficient between them being of 0.2064. The multiple linear regression of the model from Eq. (3) is summarize in Fig. 1.

Figure 1a shows the relationship between observed and estimated by Eq. (3) endpoint. By analyzing the plot it could not be identified any obvious outliers, but it can be observed that in 3 cases out of 73 the values were negative. The

Table 1 Informative models parameters for Eqs. (2) and (3)

Equations	Parameter	Coef [$_{95\%}$CI]	SE$_{\text{coef}}$	t_{coef}	r [$_{95\%}$CI$_r$]	r^2_{adj}	RSS
(2)	Intercept	–507.96 [–521.70 to –494.22]	6.89	–73.70*	0.9956 [0.9929 to 0.9972]	0.9911	1033
	Slope	188.40 [184.22 to 192.60]	2.10	89.71*			
(3)	Intercept	–323.02 [–331.55 to –314.54]	4.27	–75.72*	0.9991 [0.9985 to 0.9994]	0.9981	213
	of *liDmEHt*	–105.92 [–107.39 to –104.48]	0.73	–145.38*			
	of *IADmwHt*	17.76 [17.08 to 18.44]	0.34	51.96*			

r, Correlation coefficient; [$_{95\%}$CI], 95% confidence interval; RSS, residual sum of squares; coef, coefficients of the regression models (see Eqs. (2)–(3)); SE$_{\text{coef}}$, standard error of coefficient; t_{coef}, Student *t* parameter for coefficient; [$_{95\%}$CI$_r$], 95% confidence interval for correlation coefficient

* $p < 0.0001$

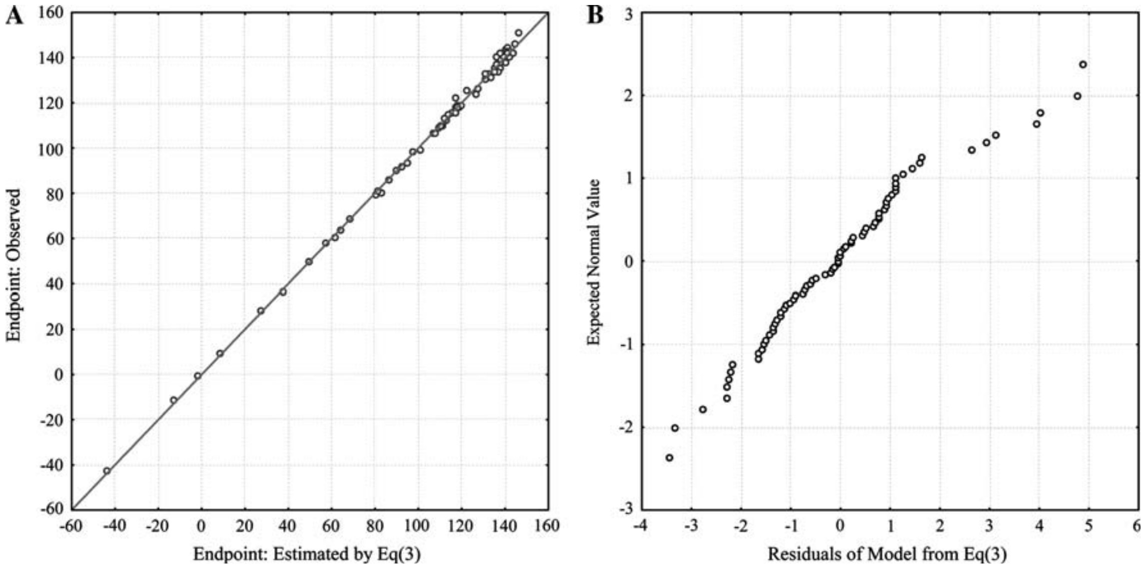

Fig. 1 **a** Relationship between observed and estimated endpoint by Eq. (3). **b** Normal probability plot of residuals

normal probability plot of residuals (Fig. 1b) was drawn in order to identify the pinpoint outliers. In the literature, it is considered that all observation points that lie on imagined straight line that goes through zero residual and 0.5 probability had approximately normal distributed residual. Any point that falls off the imagined straight line has a residual, which informs that the difference between measured and estimated endpoint is much larger or smaller then expected based on assumption of normally distributed residuals. Note that this is just an empirical interpretation of normality. Analyzing Fig. 1b it could be observed that there are seven compounds in the right-top corner (from left to right: 2,3,3,4-MMMMC5, 2,2,3,3-MMMMC4, 2,2,3-MMEC5, 3,5-MMC7, 2,3,4-MMMC6, 3,3,4-MMMC6, and 2,2,3,4-MMMMC5, respectively) of the graphic and three in the left-bottom corner (from left to right: 2,4-MEC6, 2,4,4-MMMC6, and 3M-C6, respectively). It could be considered that these ten compounds are outliers. By removing them from the dataset ($\sim$14% compounds removed) a squared correlation coefficient of 0.9994 and a standard error of the estimate of 1.07 are obtained for this model (Fisher parameter 46921, $p < 0.0001$). As a result of removing the ten compounds considered “outliers”, the squared correlation coefficient it is improved with 0.0012, which could not be considered a significant improvement. As a conclusion of the model evaluation, it can be said that the model presented in Eq. (3) is a good and accurate model.

Models validation

It is well known that in modelling, in some cases, the removal of compound(s) and/or variable(s) that “do not fit” according to some subjective criterion could significantly improve a model. The model obtained by removing the compounds and/or variables that “do not fit” had as main disadvantage the absence of representatively for additional compounds. A measure of the improvement is the goodness-fit of the model, usually expressed as the correlation coefficient and associated squared value. The main problem of the goodness-fit of a model is that with sufficiently many independent variables could be easily obtain a squared correlation coefficient close to the optimal value of 1. A model is proper to be investigated for its goodness-of-fit when the sample size of investigated compounds is four or five times greater than the number of independent variables, $n = 5v$, where n is the volume of the sample size and v the number of variables used by the model (Hawkins 2004). All investigated models presented in Eqs. (1)–(3) were valid from the point of view of ration between sample size and number of independent variables. The assessment of the predictive power of the models from Eqs. (2) and (3) was done by using two techniques: cross validation LOO approach and training versus test analysis.

Leave-one-out analysis

The LOO analysis software has been used in order to obtained the cross validation LOO score for the models from Eqs. (2) and (3). The results are presented in Table 2.

The value of the obtained $r^2_{\text{cv-loo}}$ (see Table 2) shown that the models from Eqs. (2) and (3) are excellent models in terms of estimation capabilities. The value of the difference between r^2 and $r^2_{\text{cv-loo}}$ indicated the absence of any irrelevant variable in the investigated models or any outlying data points. This observation sustained also the absence of the “possible” outlier’s identified in Fig. 1b.

Table 2 Results of leave-one-out analysis

Equations	r^2	$r^2_{\text{cv-loo}}$	$F_{\text{cv-loo}}$	$r^2 - r^2_{\text{cv-loo}}$
(2)	0.9913	0.9908	7654^*	0.0004
(3)	0.9982	0.9980	17562^*	0.0002

* $p < 0.0001$

Training versus test analysis

Two different training versus test analysis approaches were applied for validation of the model presented in Eq. (3):

(1) The sample of compounds was randomly spitting into training and test set by using the training vs. test experiment software. A number of 24 experiments were conducted and the results are presented in Table 3.

(2) Ten compounds that were suspected to be outliers by analyzing the graphical representation presented in Fig. 1b (2,3,3,4-MMMMC5, 2,2,3,3-MMMMC4, 2,2,3-MMEC5, 3,5-MMC7, 2,3,4-MMMC6, 3,3,4-MMMC6, 2,2,3,4-MMMMC5, 2,4-MEC6, 2,4,4-MMMC6, and 3M-C6, respectively) were included into test set while the other compounds formed the training test.

As it can be observed from Table 3, all investigated models were statistical significant ($p < 0.05$). A deepen analysis revealed that all correlation coefficients in training sets respected the 95% confidence intervals of Eq. (3). In 6 cases out of 24 the correlation coefficient obtained in test set was greater than the upper boundary of the 95% confidence intervals obtained by Eq. (3). Moreover, in 95 percent of cases, the coefficients of the models (Table 3) respected the coefficient 95% confidence intervals of the model from Eq. (3) (Table 1). In 5 cases out of 24, the correlation coefficients obtained in training and test sets were statistically different (note that just in one case the

Table 3 Training versus test analysis on model from Eq. (3): results on 24 experiments

$\hat{Y} = a_0 + a_1 \cdot liDmEHt + a_2 \cdot IADmwHt$			Training set			Test set			$Z_{\text{rtr–rts}}$
a_0	a_1	a_2	n_{tr}	r^2_{tr}	F_{tr}	n_{ts}	r^2_{ts}	F_{ts}	
–327.86	–104.37	18.30	40	0.9989	$16370^{\dagger}$	33	0.9971	$4664^{\dagger}$	$1.87^{\ddagger}$
–325.06	–106.50	17.83	41	0.9966	$5533^{\dagger}$	32	0.9991	$13997^{\dagger}$	$2.48^{\ddagger}$
–322.09	–105.41	17.73	42	0.9985	$13212^{\dagger}$	31	0.9982	$6045^{\dagger}$	0.24^*
–320.57	–107.00	17.46	43	0.9978	$9243^{\dagger}$	30	0.9988	$8219^{\dagger}$	1.22^*
–321.72	–106.30	17.64	44	0.9984	$12938^{\dagger}$	29	0.9980	$4464^{\dagger}$	0.45^*
–317.82	–107.05	17.23	45	0.9981	$10952^{\dagger}$	28	0.9981	$6383^{\dagger}$	0.00^*
–324.34	–105.55	17.91	46	0.9980	$10810^{\dagger}$	27	0.9986	$8356^{\dagger}$	0.70^*
–331.61	–105.53	18.44	47	0.9985	$14383^{\dagger}$	26	0.9975	$4169^{\dagger}$	0.94^*
–335.03	–105.28	18.72	48	0.9986	$16162^{\dagger}$	25	0.9972	$3605^{\dagger}$	1.33^*
–324.32	–106.20	17.81	49	0.9982	$12584^{\dagger}$	24	0.9983	$5788^{\dagger}$	0.00^*
–324.49	–106.04	17.83	50	0.9967	$7024^{\dagger}$	23	0.9996	$15570^{\dagger}$	$4.01^{\ddagger}$
–329.76	–106.00	18.25	51	0.9979	$11593^{\dagger}$	22	0.9986	$5436^{\dagger}$	0.83^*
–320.98	–106.07	17.60	52	0.9981	$12880^{\dagger}$	21	0.9985	$5510^{\dagger}$	0.41^*
–322.94	–106.33	17.70	53	0.9971	$8506^{\dagger}$	20	0.9993	$12350^{\dagger}$	$2.36^{\ddagger}$
–324.79	–106.55	17.83	54	0.9980	$12533^{\dagger}$	19	0.9994	$5287^{\dagger}$	$2.10^{\ddagger}$
–322.49	–105.94	17.72	55	0.9982	$14567^{\dagger}$	18	0.9981	$3997^{\dagger}$	0.18^*
–324.48	–105.75	17.88	56	0.9978	$11799^{\dagger}$	17	0.9987	$5353^{\dagger}$	0.75^*
–320.21	–106.17	17.51	57	0.9976	$11117^{\dagger}$	16	0.9989	$5082^{\dagger}$	1.12^*
–325.53	–106.16	17.91	58	0.9978	$12430^{\dagger}$	15	0.9991	$5682^{\dagger}$	1.24^*
–323.75	–105.80	17.83	59	0.9983	$16044^{\dagger}$	14	0.9977	$2387^{\dagger}$	0.44^*
–324.08	–106.40	17.79	60	0.9981	$15104^{\dagger}$	13	0.9993	$3092^{\dagger}$	1.34^*
–327.75	–105.81	18.12	61	0.9980	$14601^{\dagger}$	12	0.9989	$3282^{\dagger}$	0.71^*
–321.58	–105.75	17.67	62	0.9984	$18507^{\dagger}$	11	0.9974	$733^{\dagger}$	0.65^*
–323.89	–105.80	17.84	63	0.9980	$15306^{\dagger}$	10	0.9985	$2164^{\dagger}$	0.28^*

a_0, a_1, and $a_{2,}$ regression coefficient; n_{tr}, number of compounds in training set; r^2_{tr}, squared correlation coefficient in training set; F_{tr} = Fisher parameter for regression model obtained in training set; n_{ts} = number of compounds in test set; r^2_{ts} = squared correlation coefficient in test set; F_{ts} = Fisher parameter for regression model obtained in test set

* $p \geq 0.05$; ‡ $p < 0.05$; † $p < 0.0001$

correlation coefficient obtained in test set was less than the correlation coefficient obtained in training set for $n_{\text{tr}} = 40$, see Table 3). The graphical representation of estimated versus observed endpoint when the sample size in training set was equal with $2/3{\cdot}n$ ($n = 73$) is presented in Fig. 2a.

The null hypothesis that the correlation coefficient obtained in training set is not significantly different by the correlation coefficient obtained in corresponding test set was tested by applying the Fisher Z test at a significance level of 5%. As it can be observed from Table 3, significant differences were identified in 5 cases out of 24. In all situations, the correlation coefficient obtained in test set was significant statistic greater compared with the correlation coefficient obtained in training set (see Table 3).

The second strategy on training versus test analysis was applied and the graphical representation from Fig. 2b was obtained. Starting with the observations obtained from Fig. 1b, the ten compounds considered as “possible” outliers were included in the test set while the others were included in the training set. The equation obtained in training set Eq. (4) and associated statistical parameters were:

$$\hat{Y}_{\text{2D-tr}} = -322.94 - 105.18 \cdot liDmEHt + 17.83 \cdot IADmwHt$$
$$r^2 = 0.9992;\ F = 39949\,(p < 0.0001);\ s = 1.17,\ n = 63 \tag{4}$$

The 95% confidence intervals of the correlation coefficient obtained in training set [0.9993–0.9997] did not overlap on the 95% confidence intervals of the correlation coefficient obtained in test set [0.9427–0.9969] suggesting a significant difference, but a good model (a correlation coefficient of 0.9867, and an associated squared value of 0.9736). The graphical representation of the models is presented in Fig. 2b.

Models comparison–correlated correlation analysis

The correlated correlation analysis was performed to identify differences between models. Note that, the correlations being compared were not independent, they sharing one variable (the endpoint). The analysis was able to revealed whether two correlations (obtained by two different models) hade different strengths (the Steiger’s Z test was applied at a significance level of 5%, see Table 4).

Analyzing the results presented in Table 4, it can be observed that the relationship between molecular descriptors and boiling points, Eq. (3), is strengthens comparing with the relationship between descriptors and boiling points Eq. (1). Note that, better performances in terms of strength are obtained by a model with two descriptors, Eq. (3), then by a model with three descriptors Eq. (1). The strength is also significantly greater when two molecular descriptors are used instead of one when models are obtained by using the molecular descriptors family on the structure–property relationships approach (see Table 4, $r_{\text{Eq (2)}} - r_{\text{Eq (3)}}$).

In (Q)SAR/(Q)SPR modelling it is easy to manipulate data such way that an apparently good model to be obtained. The most widely used statistical technique is multiple regression analysis, for its ability in establishing a correlation between independent variables and an endpoint. However, it is of crucial importance to realize the difference between models’s fit and prediction ability. The fit tells how well are able to reproduce the endpoint data of a

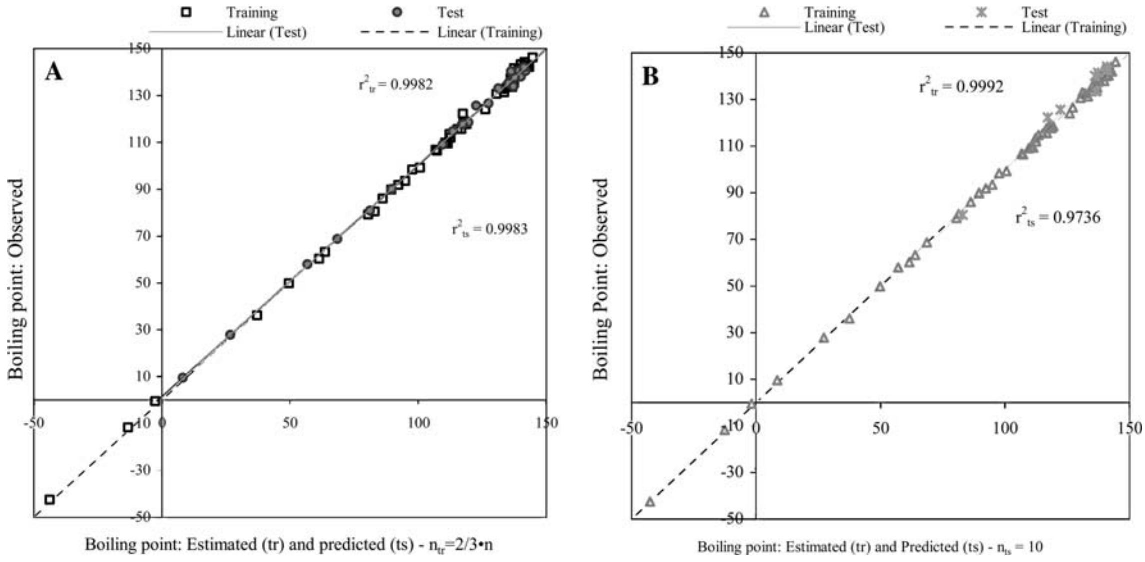

Fig. 2 **a** Relationship between observed and estimated/predicted endpoint ($n_{\text{ts}} = 1/3\ n$). **b** Relationship between observed and estimated/predicted endpoint ($n_{\text{ts}} = 10$)

Table 4 Steiger's *Z* test: results

Eq. (1)–Eq. (2)		Eq. (1)–Eq. (3)		Eq. (2)–Eq. (3)	
Y - $\hat{Y}_{\mathrm{Eq(2)}}$	0.9956	Y - $\hat{Y}_{\mathrm{Eq(3)}}$	0.9991	Y - $\hat{Y}_{\mathrm{Eq(3)}}$	0.9991
Y - $\hat{Y}_{\mathrm{Eq(1)}}$	0.9983	Y - $\hat{Y}_{\mathrm{Eq(1)}}$	0.9983	Y - $\hat{Y}_{\mathrm{Eq(2)}}$	0.9956
$\hat{Y}_{\mathrm{Eq(1)}}$ - $\hat{Y}_{\mathrm{Eq(2)}}$	0.9957	$\hat{Y}_{\mathrm{Eq(1)}}$ - $\hat{Y}_{\mathrm{Eq(3)}}$	0.9991	$\hat{Y}_{\mathrm{Eq(2)}}$ - $\hat{Y}_{\mathrm{Eq(3)}}$	0.9976
Steiger's *Z*	-4.24^{*}	Steiger's *Z*	$2.80^{‡}$	Steiger's *Z*	$7.02^{‡}$

‡ $p < 0.05$; * $p \geq 0.05$

set of compounds. This could be obtained easily with many variables or by removing the compounds and variables that do not fit according to some empirical criterion. Fortunately, the prediction ability (measures of how accurately the data of new compounds not previously used in the obtained model can be predict) is not as easy to manipulate.

Thus, the (Q)SAR/(Q)SPR models that are fitted it is important to be validate in order to verify if had real predictions abilities. At least two standard ways of validation the model can be applied. The most exact way, considered by many authors, is by external validation, when a new test sample not used in the model development is investigating. The second method is by applying LOO cross-validation method and training versus test analysis, when the entire sample of available compounds is used both to fit the model and to assess its validity. The last two methods are most suitable for the MDF SPR models, because according with the methodology of descriptors generation and calculations the same family of descriptors is obtained on all compounds or on a part of the interest compounds.

Concluding remarks

A clear methodology for models evaluation that to comprise, as was illustrated in the paper, evaluation, validation and comparison methods is necessary any time when (Q)SPR models are reported.

Related to linear regression models, correlation coefficients (Pearson, Spearman, Semi-Quantitative, Kendall's tau (*a*, *b*, or *c*), or Gamma), coefficient of determination (squared correlation coefficients), Fisher (*F*), Student (Pearson, Spearman, and Semi-Quantitative) and *Z* statistical tests define the measure of collinearity. To these are adding adjusted correlation coefficient, standard error of the estimate, regression model significance and regression coefficients and slope analysis complete the model analysis. Regarding models validation, two instruments (LOO analysis and training versus test analysis) were applied on the investigated models. The comparison of the models is maybe the most difficult task. But, at least one method was proved its usefulness: correlated correlation analysis.

The (Q)SPR models will have an important place in future chemical management as priority settings, risk assessment, classification and labelling. Statistical methods had an important place in (Q)SPR/(Q)SAR models assessment and the implemented software help researchers all over thee world to validate the obtained models.

Acknowledgments The research was supported by UEFISCSU Romania through projects.

References

Bolboaca SD, Jäntschi L (2006) Pearson versus Spearman, Kendall's Tau correlation analysis on structure–activity relationships of biologic active compounds. Leon J Sci 9:179–200

Buchwald P, Bodor N (2002) Computer-aided drug design: the role of quantitative structure–property, structure–activity and structure–metabolism relationships (QSPR, QSAR, QSMR). Drugs Future 27(6):577–588

Crum-Brown A, Fraser TR (1868) On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia, Thebia, Codeia, Morphia, and Nicotia. Trans R Soc Edinbours 25:151–203

Hawkins DM (2004) The problem of overfitting. J Chem Inf Comput Sci 44:1–12

Hemmateenejad B (2006) Chemometrics in Iran. Chemometr Intell Lab 81(2):202–208

Jäntschi L (2005) Molecular descriptors family on structure activity relationships 1. review of the methodology. Leon Elect J Pract Techol 6:76–98

Jäntschi L, Bolboaca S (2007) Results from the use of molecular descriptors family on structure property/activity relationships. Int J Mol Sci 8(3):189–203

Richet MC (1893) Comptes rendus des séances de la Société de biologie et de ses filiales (C. R. Seances Soc. Biol. Fil.) 45:775–776

Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. J Chem Inf Comput Sci 34:854–866

Steiger JH (1980) Tests for comparing elements of a correlation matrix. Psychol Bull 87:245–251

Tichý M (2005) Experimental toxicology in silico. Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub 149(2):217–219

Toropov A, Toropova A, Ismailov T, Bonchev D (1998) 3D weighting of molecular descriptors for QSPR/QSAR by the method of ideal symmetry (MIS). 1. Application to boiling points of alkanes. J Mol Struct Theochem 424:237–247

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–76

Wehrens R, Putter H, Buydens LMC (2000) The bootstrapp: a tutorial. Chemom Int Lab Sys 54:35–52

Worth PA, Cronin MTD (2004) Report of the workshop on the validation of QSARs and other computational prediction models. ATLA 32(1B):703–706