https://doi.org/10.26565/2220-637X-2023-40-02 ISSN 2220-637X
УДК: 544.169+519.237.5

INTERNAL VALIDATION PARAMETERS OF LINEAR REGRESSION EQUATIONS IN QSAR PROBLEM

I. V. Khristenko[a], V. V. Ivanov[b]

V. N. Karazin Kharkiv National University, School of Chemistry, 4 Svobody sqr., Kharkiv, 61022 Ukraine

a) ✉ khristenko@karazin.ua https://orcid.org/0000-0001-7227-8333
b) ✉ vivanov@karazin.ua https://orcid.org/0000-0003-2297-9048

The article discusses a set of internal validation parameters that are (or can be) used to describe the quality of regression models in quantitative structure-activity relationship problems. Among these parameters there are well known determination coefficient, root mean square deviation, mean absolute error, etc. Also the indices based at Kullback-Leibler divergence as a measure of distance between two sets have been investigated. All the parameters (indices) were calculated for several regression models which describe boiling point of saturated hydrocarbons (alkanes). Regression models include a four-component additive scheme and equations describing the property as a function of topological indices. The two types of regressions based on these indices are linear dependencies on only one topological index and linear dependencies on topological index and the number of carbon atoms in the hydrocarbon. Various linear regression equations have been described with internal validation parameters that evaluate the quality of the equations from different perspectives. It is shown that a wide set of test parameters is not only an additional yet alternative description of regression models, but also provides the most complete description of the predictive characteristics and quality of the obtained regression model.

Keywords: *Quantitative Structure-Activity Relationships (QSAR), regression models, internal validation, topological descriptors*

Introduction

It is easier to calculate a regression equation than to prove its predictive ability. This sentence is especially true for QSAR (*Quantitative Structure-Activity Relationships*) linear regression models. Necessity of proper investigation of obtained equations has been recognized during the last years. It has been demonstrated that poorly validated regression equations can be misleading when evaluating molecular activity/property. Several important articles discuss typical situations and difficulties in description of the predictive ability of regression models. The provocative titles of the articles – "The importance of being earnest[1]…" [1] "Beware of q2!" [2], "Beware of R^2…." [3] call for attention to this problem. In the presented paper, we consider the problem of validating of QSAR regression equations from a somewhat specific point of view.

First of all, we note that, for common practice, QSAR studies involve dividing the primary data into two data sets. These sets are the *training set* that is used to generate the corresponding QSAR model, and the *test set* is the data for validation of the resulting models (equations). The parameters characterizing the description of the training set by the obtained equations are considered as internal validation, while the parameters characterizing the quality of the description of the test set are external validation. In recent years, significant attention has been paid to external validation, which can be considered as a model for the practical use of the obtained equations. Regarding the content of external validation, several important issues should be noted. The primary set must be divided in a certain ratio between the training and test sets. What is this ratio? How to specify the separation of systems (points) between two sets? How to prove the correctness of the division? And, in the end, will such a division lead to a decrease in the predictive ability of equations due to a decrease in the size of the training sample? So we see that external validation leads to additional questions for which there are no general answers yet (see, however, the article and references therein discussing this problem [4-6]). Hence

[1] *The quote from the famous play by Oscar Wilde emphasizes the main idea of the authors of the article – "first, validate, and then explore".*

such a procedure is not computationally well defined. In contrast to external validation, the internal validation does not need dividing the input data into two subsets.

Without denying the necessity for an external validation procedure, in the present paper we propose to take a closer look at the internal validation (goodness-of-fit) of QSAR regression. Usually the restricted set of parameters used for the internal validation. Among these, the most important are the determination coefficient (R^2) and the standard deviation (root mean square deviation, *RMSD*). Such parameters cannot be considered as those that give a complete description of the training sample and the corresponding regression equation. Also these parameters usually demonstrate a low sensitivity to variation of the model. An extremal example is the classic Anscombe paper, where the very different data unexpectedly fit the same equation, with the same R^2 and the same *RMSD* [7,8].

In this paper, we analyze a wide set of known internal validation parameters and a few new parameters that we have proposed. As an example we describe boiling points (BP, C°) of saturated hydrocarbons (alkanes). QSAR-models of these properties include additive scheme and graph theory approaches based on known topological indices. It should be noted that the interest to topological indices has been stable for a long time up to the present day. For instance, in the paper [9], new graph theory model for description of boiling points of alkanes is discussed. Also graph theory approaches currently used for description of anti-cancer activity [10] and even for description of potential anti-COVID-19 substances [11,12].

All the calculations were performed by using Python3 script language. RDKit package was used for manipulations with chemical structures and calculations of molecular descriptors. [13] The experimental data for BP of the alkanes were obtained from [14]. When information about the physicochemical properties of alkanes includes several values, we used the average values. In total, the training set contains information on 39 different saturated hydrocarbons with 1 to 9 carbon atoms.

Linear regression models and internal validation parameters

For the physical-chemistry property of alkanes we consider three types of linear regression models. The first one is correspond to simple additive scheme.

$$Y = n_1 x_1 + n_2 x_2 + n_3 x_3 + n_4 x_4 \tag{1}$$

Where Y is the dependent variable – the physicochemical property of alkanes is a function of four parameters (n_1, n_2, n_3, n_4) that describe the molecular structure. The partial values (increments) x_1, x_2, x_3, x_4 are contributions from elements of the molecular structure (Table 1).

Table 1. Parametrization of additive scheme for alkane molecules

Fragment	Number of Fragments in the molecule	Increments
H_3C—	n_1	x_1
H_2C<	n_2	x_2
HC<	n_3	x_3
>	n_4	x_4

Also we consider *two* regression models based on graph theory. The topological indices $X = \left\{ \chi^{(1)}, ZM_1, ZM_2, ZM_2, IC_1, InfD \right\}$ (see Table 2) were used in the calculations as the molecular descriptors. For a detailed descriptions of the indexes presented in the Table 2, see for example [15,16].

The first graph theory based model is single-parameter equations:

$$Y = a_0 + a_1 X \tag{2}$$

where X is the descriptor from Table 2. The second equation includes descriptor X and the number of carbon atoms (N_C) in the hydrocarbon:

$$Y = a_0 + a_1 N_C + a_2 X \tag{3}$$

The regression coefficients (a_0, a_1, a_2) as well as partial values for additive scheme (x_1, x_2, x_3, x_4) were obtained using the Ordinary Least Squares (OLS) method (see for example [17]).

Table 2. Topological indices used in the present article (v_i– order of vertex *i,* (i, j) are pairs of connected by edges carbon atoms)

№	Topological Index	Definition
1	First order Randich index	$\chi^{(1)} = \sum_{(i,j)} 1/\sqrt{v_i v_j}$
2	First Zagreb index	$ZM_1 = \sum_i v_i^2$
3	Second Zagreb index	$ZM_2 = \sum_{(i,j)} v_i v_j$
4	Third Zagreb index (so called “forgotten index”)	$ZM_3 = \sum_i v_i^3$
5	First order informational content. n_k- number of vertices with definite v, $N = \sum_k n_k$	$IC_1 = -\sum_k \frac{n_k}{N} \log_2 \frac{n_k}{N}$
6	Informational index of distances in graph. r_k– number of routs with topological distances equal to k, $N_D = \sum_k r_k$	$InfD = -\sum_k \frac{r_k}{N_D} \log_2 \frac{r_k}{N_D}$

The residuals between given (experimental) values Y and those obtained by using regression equations (1-3) for training set (Y_i^{calc}) are calculated as follow:

$$e_i = Y_i - Y_i^{calc} \tag{4}$$

Also, the correspondence between the calculated and given values of the variable Y is usually described as a linear form which can be presented by two equivalent, but not identical, equations:

$$Y^{calc} = \beta_0 + \beta_1 Y \tag{5}$$

$$Y = \gamma_0 + \gamma_1 Y^{calc} \tag{6}$$

Of course, for the absolute (or “ideal”) correspondence between Y and Y^{calc} values, one can write ($\beta_0 = \gamma_0 = 0$, $\beta_1 = \gamma_1 = 1$):

$$Y = Y^{calc} \text{ and } Y^{calc} = Y \tag{7}$$

However, for the typical (realistic) situation of QSAR investigations, for the equation (5) one can write

$$Y^{calc} = (1 - R^2)\overline{Y} + R^2 Y \tag{8}$$

Where R is Pearson correlation coefficient, $\overline{Y} = \sum_i Y_i / n$ is mean value, and n is size of sample.

Further, according to known expression $\beta_1 \gamma_1 = R^2$ [18] for the eq. (6) we ***always*** have absolute correspondence in the sense of least square method ($\gamma_1 = 0$, $\gamma_1 = 1$).

$$Y = Y^{calc} \tag{9}$$

However, note that both equations (8) and (9) must be interpreted in the spirit of OLS and, of course, have the same coefficient of determination, R^2. A discussion of these issues can be found in [18, 19].

Hence, in the general case, the deviation of Y^{calc} from Y can be expressed in the following most formal way

$$\eta = F(Y, Y^{calc}) . \tag{10}$$

Where the parameter η describes the quality of approximation for the selected regression model. Generally speaking, expression (10) implies the use of different metrics F.

The set of η – parameters calculated for the presented regression models based at (1-3) is presented in Table 3.

Table 3. Internal Validation parameters

№	Parameter	Description	Best equation
1	Root Mean Square Deviation, *RMSD*	$RMSD = \sqrt{\sum_i \frac{e_i^2}{n-p}}$	$RMSD \to 0$
2	Determination Coefficient, R^2. For the *LOO* procedure designated as Q^2	$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i \left(Y_i - \bar{Y}\right)^2}$	$R^2 \to 1$
3	Mean Absolute Error, *MAE*	$MAE = \frac{1}{n}\sum_i \lvert e_i \rvert$	$MAE \to 0$
4	Asymmetry of residuals, *Asymm*	$Asymm = \frac{1}{n}\sum_i e_i$	$Asymm \to 0$
5	Relative error of worst point, *WPt*	$WPt = \max_i \left\{ \lvert e_i / Y_i \rvert \right\}$	$WPt \to 0$
6	Kullback-Leibler divergence between Y and Y^{calc} distributions, $D_{KL}\left(Y \,\|\, Y^{calc}\right)$	$D_{KL}\left(Y \,\|\, Y^{calc}\right) = \sum_i \lvert \tilde{Y}_i \rvert \log_2 \frac{\lvert \tilde{Y}_i \rvert}{\lvert \tilde{Y}^{calc} \rvert}$, $\sum_i \lvert \tilde{Y}_i \rvert = 1 \quad \sum_i \lvert \tilde{Y}_i^{calc} \rvert = 1$	$D_{KL}\left(Y \,\|\, Y^{calc}\right) \to 0$
7	Inhomogeneity of Residuals, *IhR*	$IhR = -\log_2 n - \frac{1}{n}\sum_i \log_2 \lvert e_i \rvert / \sum_k \lvert e_k \rvert$	$IhR \to 0$
8	Angle between ideal (7) and obtained lines (5, 8), Δφ	$\Delta\varphi = \arctan\left(\beta_1\right) - 0.7854$	$\Delta\varphi \to 0$

Here one can see several standard parameters. Among them the *RMSD* (p is number of regression coefficients) and determination coefficient – R^2. We have also included a parameter based at absolute values of error (*MAE)*. The advantages and disadvantages of parameters based at absolute values of error are discussed in details in [20-22]. Also we found a few simple parameters to be useful. *Asymm* is a measure of the under- or overestimation of the dependent variable Y. The parameters *WPt* is simple values that estimate the spread of the residual vector or can be treated as relative outlier of point.

We are using also parameters based at Kullback-Leibler informational theory [23-24]. The Kullback-Leibler divergence $D_{KL}\left(Y \,\|\, Z\right)$, also known as the relative entropy, in specific way describes how Z distribution differ from actual distribution Y. There are several interpretations of $D_{KL}\left(Y \,\|\, Z\right)$. One of them is designated as "informational lost" when Z used instead of Y (or Z approximates Y). The main properties of D_{KL} are: $D_{KL}\left(Y \,\|\, Y\right) = D_{KL}\left(Z \,\|\, Z\right) = 0$, and $D_{KL}\left(Y \,\|\, Z\right) \neq D_{KL}\left(Z \,\|\, Y\right)$. We use two indices based at Kullback-Leibler divergence. The first one is to describe divergence of Y from Y^{calc} $D_{KL}\left(Y \,\|\, Y^{calc}\right)$, and the second – to describe inhomogeneity of the residuals *IhR* (see Table 3). As an example of the use of the Kullback-Leibler informational theory in chemistry see [25].

And also we are calculating the angle ($\Delta\varphi$) between "ideal line" (7) and line that is result of the regression analysis (8). All the above mentioned parameters were calculated in two variants. The first one corresponds to the calculation of regression parameters for the full sample, and the second to the leave-one-out cross-validation procedure (*LOO*) [26-29].

Results of calculations and discussion

As a result of the OLS regression calculation for the additive scheme (Eq. 1, Table 1), the corresponding plot of the "theory-experiment" relationship is presented in Fig. 1. The red line is corresponding to "ideal" dependence (7). Here one can see, that dependence of Y^{calc} from experimental value $Y = Y^{experim}$, designated as green circle, too far from "ideal" dependence for additive scheme.

Among the results obtained with the simple one-parameteric equation (2), much better solutions can be found. Especially the equation with Randich index ($\chi^{(1)}$) demonstrated the best result.

$$BP(C°) = -141.6 + 71.026\chi^{(1)}. \quad (11)$$

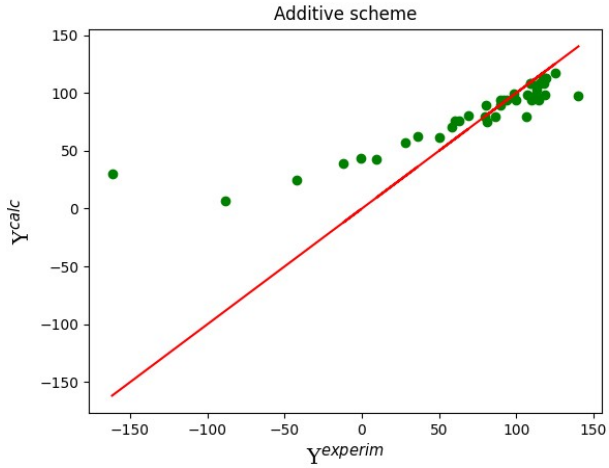
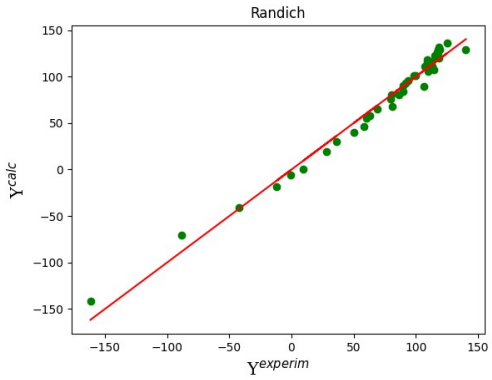


Figure 1. Dependence "theory-experiment" according to the additive scheme.

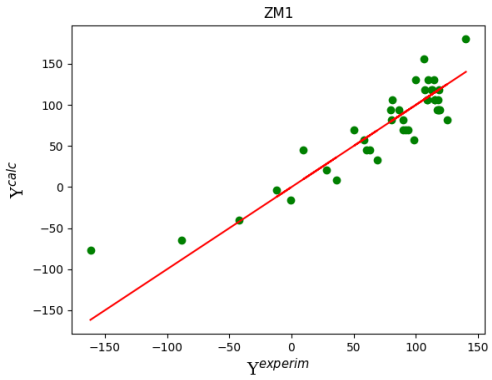
The internal validation parameters of the linear regressions obtained using the additive approach and those that follow eq. 2 are collected in Table 4. According to the obtained data, the equation based at $\chi^{(1)}$ is the best equation for all the parameters presented. Nevertheless, the choice of the following equations (in quality) depends on the chosen validation parameter. According to R^2 (and Q^2, MAE, *Δφ*) the next best equation is function from *ZM1*. The worst result was demonstrated by the *ZM3* index ($R^2 = 0.5265$ and a poor correspondence to the distribution of experimental data $D_{KL}\left(Y \,||\, Y^{calc}\right) = 0.5624$) with abnormal sensitivity to selected groups of molecules even compared to *InfD*. For the *InfD* the value $D_{KL}\left(Y \,||\, Y^{calc}\right) = 0.1445$ is significantly better than for the *ZM3* $D_{KL}\left(Y \,||\, Y^{calc}\right) = 0.3622$. It is also interesting that the average inhomogeneity of the residuals (*IhR*) is quite small for all indices (equations), even though there are large differences in *WPt* (relative outliers). However, *IhR* and *LOO IhR* for the additive scheme is noticeably greater than for all other regressions. Significant values of *Δφ* and *LOO Δφ* for additive approach are indicators of large difference between "ideal line" and actual.

For the two-parametric equations (3), which also include the number of carbon atoms in the molecule, the pictures are more optimistic. The quality of the equations in terms of validation parameters is much better (Fig. 3 and Table 5). Formally the best equation is:

$$BP(C°) = -168.6 + 27.8Nc + 44.3InfD. \quad (12)$$



A



B

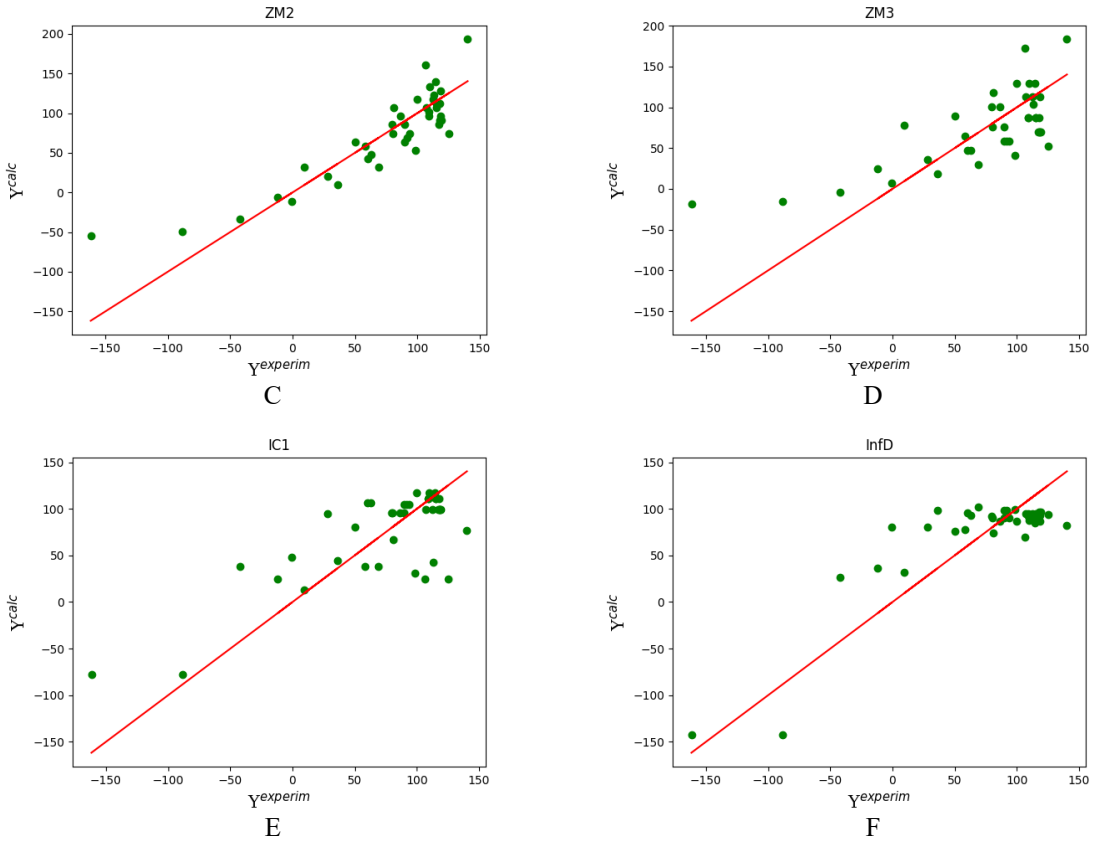


Figure 2. The QSAR estimations according to eq. 2 versus experimental data.

The equation is characterized by a large value of R^2, a smaller value of $D_{KL}\left(Y \,||\, Y^{calc}\right)$ and values of the other indices except inhomogeneity parameters *IhR* and *LOO IhR*. The *IhR* and *LOO IhR* values for *InfD* are only slightly worse than the corresponding values for other equations based on topological indices. The parameter *WPt*, characterizing the maximal relative outlier for (12), is much smaller than for the other equations. However, it can be noted visually that the best equation (12) is characterized by typical "steps" on the graph of the "calculation-experiment" dependence (Fig. 3, F). This indicates poor recognizability of some groups of molecules by regression model. This is in spite of high values of R^2, and low values of $D_{KL}\left(Y \,||\, Y^{calc}\right)$.

Table 4. The internal validation parameters for additive scheme (1) and regression equations (2).

	additive	$\chi^{(1)}$	ZM_1	ZM_2	ZM_3	IC_1	*InfD*
RMSD	41.2	8.7	26.2	29.7	43.1	40.0	33.0
LOO RMSD	61.1	9.8	29.2	33.1	47.0	43.5	36.7
R^2	0.5885	0.9808	0.8245	0.7740	0.5265	0.5914	0.7216
$Q^2 = LOO\ R^2$	0.0946	0.9754	0.7827	0.7201	0.4369	0.5169	0.6562
MAE	20.8	7.0	19.8	21.2	32.9	28.2	26.2
LOO MAE	26.1	7.6	21.3	22.9	35.1	30.1	28.4
Asymm	-0.93	10^{-14}	10^{-14}	10^{-13}	10^{-15}	10^{-14}	10^{-14}
LOO Asymm	-12.9	-0.03	-0.7	-0.9	-1.20	-0.7	0.6
WPt	80.2	9.3	27.9	20.1	14.5	88.7	147.4
LOO WPt	87.3	10.0	30.6	22.0	15.7	91.6	151.1
$D_{KL}(Y \,\|\, Y^{calc})$	0.5074	0.0454	0.1912	0.2539	0.5624	0.3622	0.1445
$LOO\ D_{KL}(Y \,\|\, Y^{pred})$	0.4042	0.0650	0.2445	0.3244	0.8227	0.4575	0.1314
IhR	1.17	0.48	0.64	0.79	0.55	0.70	0.55
LOO IhR	1.34	0.52	0.67	0.81	0.57	0.72	0.59
$\Delta\varphi$	-23.1	-0.6	-5.5	-7.3	-17.2	-14.4	-9.2
$LOO\ \Delta\varphi$	-36.2	-1.1	-6.8	-8.6	-19.3	-17.3	-8.7

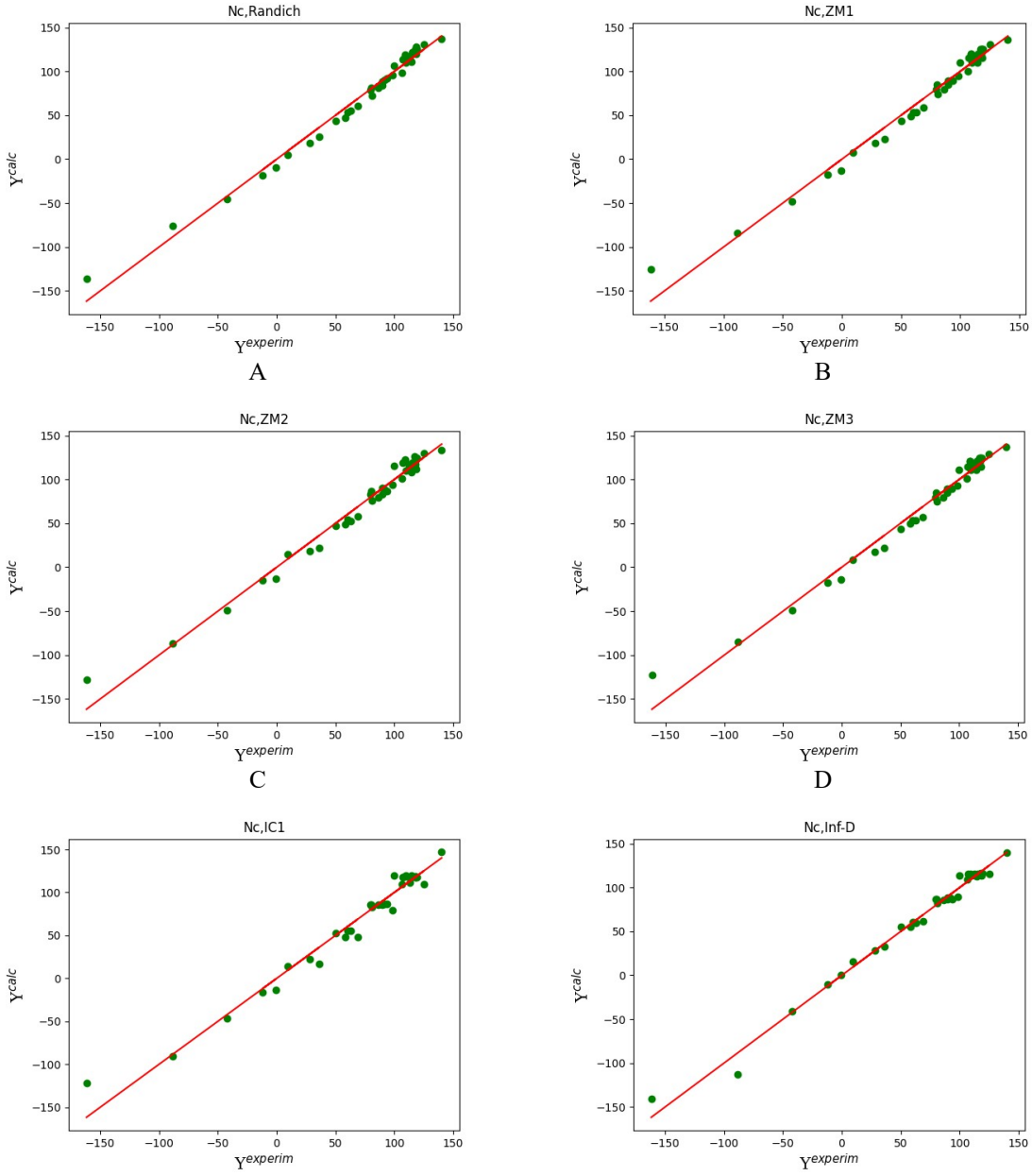


Figure 3. Data from QSAR regression (eq. 3) *vs* experimental data.

A similar by quality equation

$$BP(C°) = -146.51 + 10.75Nc + 48.91\chi^{(1)}. \quad (13)$$

is visually free of this drawback (Fig. 3, A), but appears to have noticeable differences between estimated and theoretical distributions of the dependent variable ($D_{KL}\left(Y \,||\, Y^{calc}\right) = 0.03$ and $LOO\, D_{KL}\left(Y \,||\, Y^{pred}\right) = 0.05$).

Table 5. Internal validation parameters for regression equations (3)

	$\chi^{(1)}$	ZM_1	ZM_2	ZM_3	IC_1	$InfD$
RMSD	7.9	9.2	9.4	9.6	11.0	7.2
LOO RMSD	10.1	11.4	11.6	11.8	13.3	11.0
R^2	0.9845	0.9791	0.9781	0.9772	0.9701	0.9871
Q^2=*LOO* R^2	0.9747	0.9679	0.9668	0.9656	0.9554	0.9698
MAE	6.1	6.7	7.0	6.9	7.3	4.7
LOO MAE	7.0	7.5	7.8	7.7	8.3	5.8
Asymm	10^{-12}	10^{-12}	10^{-13}	10^{-13}	10^{-13}	10^{-13}
LOO Asymm	-0.39	-0.29	-0.28	-0.28	-0.25	0.011
WPt	17	22.2	23.5	24	24.3	2.4
LOO WPt	18.8	24.4	25.8	26.2	27.0	2.9
$D_{KL}(Y \parallel Y^{calc})$	0.0301	0.0299	0.0263	0.0310	0.0343	0.004
LOO $D_{KL}(Y \parallel Y^{pred})$	0.0531	0.0491	0.0461	0.0505	0.0555	0.004
IhR	0.52	0.51	0.59	0.46	0.83	0.98
LOO IhR	0.59	0.56	0.64	0.51	0.88	1.17
$\Delta\varphi$	-0.45	-0.61	-0.63	-0.66	-0.87	-0.37
LOO $\Delta\varphi$	-1.26	-1.29	-1.38	-1.33	-1.63	-0.47

Conclusion

To date, work on QSAR regression equation testing problems has shifted significantly toward external validation. However, the multitude of internal validation parameters is a useful tool for multilateral analysis of the resulting regression equations. In this paper, we have examined several internal validation parameters that are different in nature. It has been shown that such parameters can complement each other. In particular, parameters based on the Kullback-Leibler informational theory (indices $D_{KL}\left(Y \parallel Y^{calc}\right)$ and *IhR*) describe the correspondence of theoretical, based on the regression model, and experimental data, from a different perspective than the determination coefficient and other known parameters. Assessing the results of calculations, also it should be noted that the standard set of parameters is still not sufficient to identify the presence of "steps" in the "calculation-experiment" graph. Their influence on the quality of approximation still awaits quantitative assessment. At the same time, visual analysis of the "theory-experiment" graph remains important.

Note also that there is obviously no universal solution in choosing the best (accurate) regression model to describe the properties of the system in a situation where there is a large scatter in the initial data. In this case, it is difficult to avoid a wide range of parameters (standard deviation, coefficient of determination, etc.) when estimating multiple models.

Acknowledgment

This research was partly supported by the Ukrainian Minister of Education and Science. Grant № 0122U001485 "Design and optimization of functional nanodisperse systems: lyophilic aggregates, biocompatible ashes, hybrid materials, photoelectric converters".

References

1. Tropsha A., Gramatica P. and Gombar V. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR Comb. Sci.* 2003, *22*, 69-77. https://doi.org/10.1002/qsar.200390007
2. Golbraikh A., Tropsha A. Beware of Q2! *Journal of Molecular Graphics and Modelling*. 2002, *20*, 269–276. https://doi.org/10.1016/S1093-3263(01)00123-1
3. Alexander D. L. J., Tropsha A., and Winkler D. A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. of Chem. Inform. and Model.* 2015, *55*(7), 1316-1322. https://doi.org/10.1021/acs.jcim.5b00206
4. Joseph V. R. and Vakayil A. SPlit: An Optimal Method for Data Splitting. *Technometrics*. 2021, *64(2)*, 166-176. https://doi.org/10.1080/00401706.2021.1921037

5. Liu H., Cocea M. Semi-random partitioning of data into training and test sets in granular computing context. *Granul. Comput.* 2017, *2*, 357–386. https://doi.org/10.1007/s41066-017-0049-2
6. Joseph V. R. Optimal ratio for data splitting. *Stat. Anal. Data Min.: ASA Data Sci. J*. 2022, *15*, 531–538. https://doi.org/10.1002/sam.11583
7. Anscombe F. J. Graphs in Statistical Analysis. *Am. Stat.* 1973, *27*, 17-21. https://doi.org/10.2307/2682899
8. Besalu E., Julian-Ortiz J. V., Pogliani L. Trends and Plot Methods in MLR Studies. *J. Chem. Inf. Model.* 2007, *47,* 751-760. https://doi.org/10.1021/ci6004959
9. Mukwembi S., Nyabadza F. A new model for predicting boiling points of alkanes, *Scientific Reports*, 2021, *11*, 24261. https://doi.org/10.1038/s41598-021-03541-z
10. Mukwembi S, Nyabadza F. Predicting anti−cancer activity in flavonoids: a graph theoretic approach. *Scientific Reports*. 2023, *13*, 3309. https://doi.org/10.1038/s41598-023-30517-y
11. Zhen W., Khalid A., Ali P., Rehman H., Siddiqui M. K., Ullah H. Topological Study of Some Covid-19 Drugs by Using Temperature Indices. *Polycyclic Aromatic Compounds*. 2022. *43 (2)*, 1133-1144. 10.1080/10406638.2022.2025864
12. Zhang Y., Khalid A., Siddiqui M. K., Rehman H., Ishtiaq H., and Cancan M. On Analysis of Temperature Based Topological Indices of Some Covid-19 Drugs. *Polycyclic aromatic compounds*. 2023, *43*(4), 3810–3826. https://doi.org/10.1080/10406638.2022.2080238
13. https://www.rdkit.org/
14. https://www.chemeo.com/
15. Todeschini R., & Consonni V. (2000). *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH.
16. Devillers J., & Balaban A. T. (1999). *Topological Indices and Related Descriptors in QSAR and QSPR*. London: CRC Press
17. Roy K., Kar S., Das N.R. A Primer on QSAR/QSPR Modeling. Fundamental Concepts. Springer: 2015
18. M.R. Spiegel, John J. Schiller, R. A. Srinivasan Probability and Statistics. McGraw-Hill, New York, 2013, 424 p.
19. Besalú E., de Julián-Ortiz J. V., Iglesias M., Pogliani. L. An overlooked property of plot methods. *Journal of Mathematical Chemistry*. 2006, *39*, 475-484. https://doi.org/10.1007/s10910-005-9035-z
20. Hyndman R. J., Koehler A. B. Another look at measures of forecast accuracy, *International Journal of Forecasting.* 2006, *22*, 679– 688. https://doi.org/10.1016/j.ijforecast.2006.03.001
21. Hyndman R. J. Another Look at Forecast Accuracy Metrics for Intermittent Demand, *Foresight.* 2006, *4*, 43-46. https://robjhyndman.com/papers/foresight.pdf
22. Hewamalage H., Ackermann K., Bergmeir C. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Min Knowl Disc.* 2023, *37*, 788–832. https://doi.org/10.1007/s10618-022-00894-5
23. Kullback S. Information theory in statistics, Glouchester Mass, 1978, 399 p.
24. Kullback S. Leibler R.A. On information and sufficiency. *Ann. Math. Statist*. 1951, *22(1)*, 79 –86. https://doi.org/10.1214/aoms/1177729694
25. Hummer G., Garde S., Garcia A. E., Pohorille A., Prat L. R. An information theory model of hydrophobic interactions. *Proc. Natl. Acad. Sci. USA.* 1996, *93*, 8951-8955. https://doi.org/10.1073/pnas.93.17.895
26. Arlot S., Celisse A. A survey of cross-validation procedures for model selection. *Statistics surveys*, 2010, *4*, 40-79. https://doi.org/10.48550/arXiv.0907.4728
27. Quan N. T. The Prediction Sum of Squares as a General Measure for Regression Diagnostics. *J. Business & Economic Statistics*. 1988, *6(4)*, 501-504. https://doi.org/10.1080/07350015.1988.10509698
28. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society*. 1974, *Ser. B, 36*, 111-133. http://www.jstor.org/stable/2984809
29. Cawley G. C., Talbot N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research.* 2010*, 11*, 2079-2107. http://jmlr.org/papers/v11/cawley10a.html

Received 19.05.2023 *Accepted 11.09.2023*

INTERNAL VALIDATION PARAMETERS OF LINEAR REGRESSION EQUATIONS IN QSAR PROBLEM

І. В. Христенко, В. В. Іванов. Параметри внутрішньої валідації рівнянь лінійної регресії в проблемі QSAR.

Харківський національний університет імені В. Н. Каразіна, майдан Свободи 4, Харків, 61022, Україна

У статті обговорюється набір внутрішніх параметрів валідації, які використовуються (або можуть бути використані) для опису якості регресійних моделей у задачах QSAR. Серед цих параметрів добре відомі коефіцієнт детермінації, залишкове середнє квадратичне відхилення, середня абсолютна похибка тощо. Також були досліджені індекси, засновані на дивергенції Кульбака-Лейблера як міри відстані між двома множинами. Всі параметри (індекси) були розраховані для декількох регресійних моделей, які описують температуру кипіння насичених вуглеводнів (алканів). Регресійні моделі включають чотирьохкомпонентну адитивну схему та рівняння, що описують температуру кипіння як функцію топологічних індексів. Два типи регресій на основі цих індексів - лінійні залежності тільки від одного топологічного індексу та лінійні залежності від кількості атомів вуглецю у вуглеводневій речовині та топологічного індексу.

Описано різні лінійні рівняння регресії з внутрішніми валідаційними параметрами, які оцінюють якість рівнянь з різних точок зору. Показано, що широкий набір тестових параметрів є не тільки додатковим, чи альтернативним описом регресійних моделей, а й забезпечує більш повніший опис прогностичних характеристик та якості отриманої регресійної моделі.

***Ключові слова**: Кількісне співвідношення структура-властивість (QSAR), регресійні моделі, внутрішня валідація, топологічні дескриптори.*

Надіслано до редакції 19.05.2023 *Прийнято до друку 11.09.2023*

Kharkiv University Bulletin. Chemical Series. Issue 40 (63), 2023