Externally Predictive QSAR Models: Thresholds of Acceptance by Various External Validation Criteria and Critical Inspection of Scatter Plots

Nicola Chirico, Ester Papa and Paola Gramatica

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Theoretical and Applied Sciences, University of Insubria, Varese, Italy.

E-mails: nicola.chirico@uninsubria.it; paola.gramatica@uninsubria.it - web: http://www.qsar.it

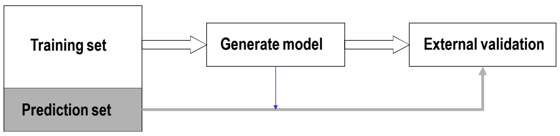
ABSTRACT

The evaluation of linear regression QSAR models performances, both in fitting and external prediction, is of pivotal importance [1][2]. In the last decade different external validation parameters have been proposed: Q^2_{F1} (Shi) [3], Q^2_{F2} (Schuurmann) [4], Q^2_{F3} (Todeschini) [5], average r^2_m (Roy) [6] and the Golbraikh –Tropsha (GT) method [7]. Recently, the concordance correlation coefficient (CCC, Lin) [8] has been proposed by our group as an external validation parameter to be used in QSAR studies. In our recent work, published in 2011 on JCIM [9], we have shown that, comparing with the commonly used acceptance thresholds (Q^2_{Fn}=0.6, average r^2_m=0.5), the concordance correlation coefficient threshold value (=0.85) is usually the most restrictive in the acceptance of QSAR models as externally predictive. This fact suggested that the CCC could be used as the preferred validation parameter in a precautionary approach, if the aim of QSAR developers is to have the smallest differences, within a certain range, among the experimental data and the predictions of the external data set.

In this new work [10], we have studied and compared the general trends of the various criteria in dependence of different possible bias in the external data distributions (scale, location, and location plus scale shifts), by means of a wide range of different simulated scenarios. This study highlighted, also by visual inspections of the experimental vs. predicted plots, some problems related to a few criteria; in particular, average r^2_m, if based on the proposed cut-off, could be prone to accept also not predictive models. This analysis allowed also to propose recalibrated, and inter-comparable, new thresholds for each criteria in the definition of a QSAR model as externally predictive. Two additional relevant topics emerged from the analysis of the results: 1) the scatter plot of the external predictions must always be evaluated and 2) the root mean squared error (RMSE) must also be calculated, as it is usually done in the good QSAR practice. In fact, we have verified that the sensitivity of the various validation criteria to RMSE often differs.

An additional important topic, here considered and applicable only to CCC, was to check by hypothesis test if the value of the calculated CCC is statistically significant [11]. This procedure allowed, consequently, to determine the minimum acceptable size of the external data set, an important point in QSAR studies, where the data set sizes are often small.

EXTERNAL VALIDATION PARAMETERS



External validation is basically based on two techniques:

- Q^2 formulas
- Other metrics

Q^2 formulas

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2} \quad [3]$$

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2} \quad [4]$$

$$Q_{F3}^2 = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2\right] / n_{EXT}}{\left[\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2\right] / n_{TR}} \quad [5]$$

Other metrics

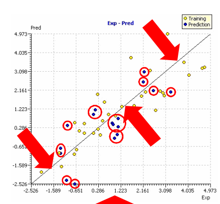
GOLBRAIKH AND TROPSHA METHOD [7]

- R^2 and R^2_0 (origin forced)
- Angular coefficients
- Closeness: $(R^2 - R^2_0) / R^2$

ROY METHOD [6]

$$\overline{r_m^2} = \frac{r_m^2 + r_m'^2}{2} \qquad \Delta_m^2 = \left| r_m^2 - r_m'^2 \right|$$

where: $r_m^2 = R^2\left(1 - \sqrt{R^2 - R_0^2}\right)$



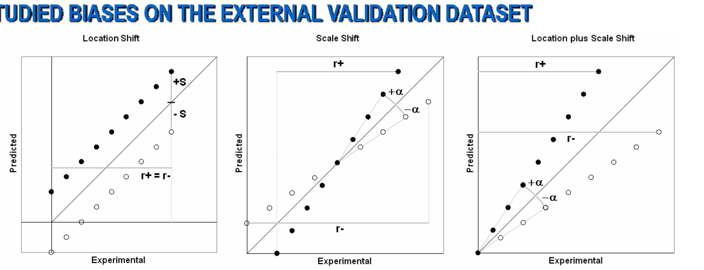
WE PROPOSE [9, 10]

Concordance correlation coefficient [8]

$$CCC = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 + n(\bar{y} - \bar{\hat{y}})^2}$$

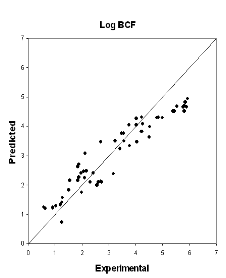
It is similar to the correlation coefficient (linear alignement), but, in addition, it takes into account the closness to the diagonal (perfect match)

STUDIED BIASES ON THE EXTERNAL VALIDATION DATASET



External validation data can be biased in different ways. The performances of the validation criteria are here studied using the three biases studied by Lin [9]: location shift, scale shift and location plus scale shift.

REQUESTED NUMBER OF EXTERNAL VALIDATION ELEMENTS



Using the method proposed by Lin [11] it is possible to calculate the minimum number of external elements requested to perform an hypothesis test (i.e. in rejecting the computed CCC if smaller or equal to the least acceptable one, which is calculated by the Lin's method).

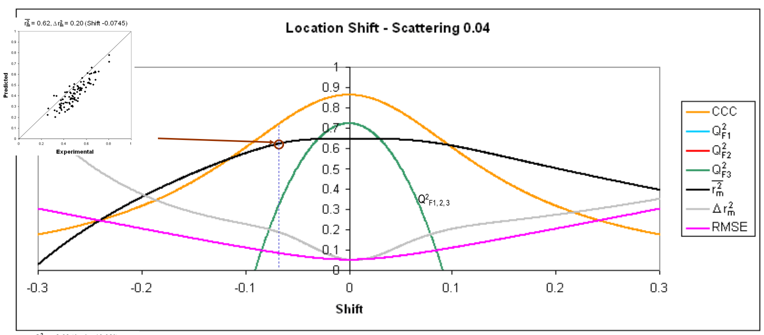
We thus calculated the minimum number of elements requested in different simulated data sets. Here we present an example on a real dataset [12].

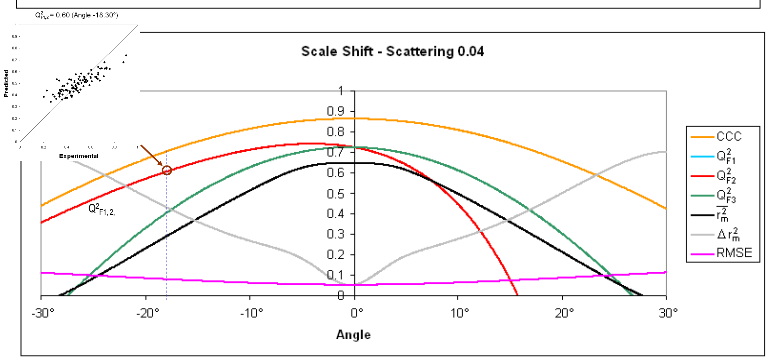
The minimum number of elements resulted to be from 52 to 66, with a confidence interval of 0.95. The number of elements in the studied dataset is 59, thus within the reported interval.

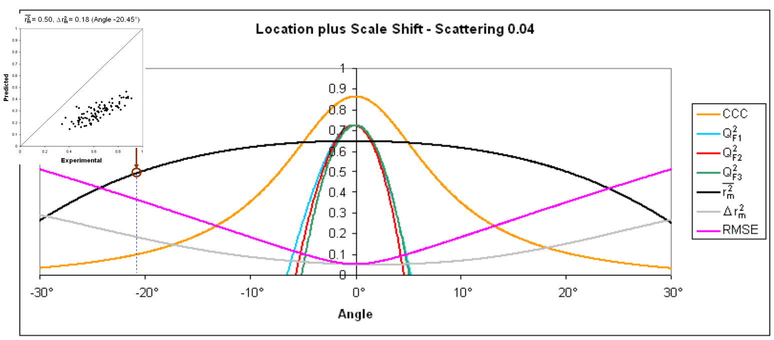
MATERIAL AND METHODS

Datasets are generated at random, following a gaussian distribution, by means of a custom simulation software. For every dataset, different level of biases (location, scale and scale plus location shift) have been applied, for different levels of data scattering (ranging from 0 to 0.06), resulting on a total of 9×10^6 of datasets. Every new inter-comparable threshold is calculated averaging 100 datasets.

BEHAVIOR OF THE VALIDATION CRITERIA AT DIFFERENT DATA BIASES







Some of the studied validation criteria tend to accept not predictive external data sets, in some of the applied biases: in particular the averaged r^2_m in the location and location plus scale shift scenario, and $Q^2_{F1,2}$ in the scale shift one for negative values of the shift.. In addition, some of the studied criteria showed to be unbalanced with respect to the RMSE values: the averaged r^2_m for the location shift scenario and $Q^2_{F1,2}$ for the location plus scale shift and, to a much higher level, for the scale shift scenario.

NEW INTER-COMPARABLE THRESHOLDS

Shift type	Shift Angle	CCC	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	$\overline{r_m^2}$	Δr_m^2
Location Shift	0	0.86 ± 0.03	0.72 ± 0.05	0.72 ± 0.05	0.72 ± 0.05	0.65 ± 0.06	0.05 ± 0.04
Location Shift	-0.0375	0.81 ± 0.03	0.60 ± 0.07	0.60 ± 0.07	0.60 ± 0.07	0.65 ± 0.06	0.12 ± 0.06
Location Shift	0.0375	0.81 ± 0.03	0.60 ± 0.07	0.60 ± 0.07	0.60 ± 0.07	0.64 ± 0.06	0.10 ± 0.06
Location Shift	-0.0745	0.70 ± 0.04	0.24 ± 0.12	0.24 ± 0.12	0.24 ± 0.12	0.62 ± 0.05	0.20 ± 0.04
Location Shift	0.0935	0.63 ± 0.04	-0.04 ± 0.16	-0.04 ± 0.16	-0.04 ± 0.16	0.62 ± 0.05	0.20 ± 0.04
Scale Shift	0°	0.86 ± 0.03	0.72 ± 0.05	0.72 ± 0.05	0.72 ± 0.05	0.65 ± 0.06	0.05 ± 0.04
Scale Shift	-18.30°	0.70 ± 0.04	0.60 ± 0.04	0.60 ± 0.04	0.39 ± 0.05	0.28 ± 0.05	0.44 ± 0.04
Scale Shift	6.35°	0.84 ± 0.03	0.60 ± 0.07	0.60 ± 0.07	0.68 ± 0.06	0.59 ± 0.05	0.22 ± 0.04
Scale Shift	-11.20°	0.80 ± 0.03	0.70 ± 0.04	0.70 ± 0.04	0.60 ± 0.05	0.48 ± 0.05	0.29 ± 0.03
Scale Shift	10.55°	0.80 ± 0.03	0.42 ± 0.10	0.42 ± 0.10	0.60 ± 0.07	0.48 ± 0.05	0.29 ± 0.03
Scale Shift	-5.65°	0.85 ± 0.03	0.74 ± 0.05	0.74 ± 0.05	0.69 ± 0.05	0.61 ± 0.05	0.20 ± 0.04
Scale Shift	5.20°	0.85 ± 0.03	0.63 ± 0.07	0.63 ± 0.07	0.69 ± 0.06	0.61 ± 0.05	0.20 ± 0.04
Location + Scale Shift	0°	0.86 ± 0.03	0.72 ± 0.05	0.72 ± 0.05	0.72 ± 0.05	0.65 ± 0.06	0.06 ± 0.04
Location + Scale Shift	-2.50°	0.80 ± 0.03	0.60 ± 0.06	0.58 ± 0.07	0.55 ± 0.07	0.65 ± 0.06	0.06 ± 0.04
Location + Scale Shift	2.00°	0.82 ± 0.03	0.60 ± 0.07	0.59 ± 0.07	0.61 ± 0.07	0.65 ± 0.06	0.05 ± 0.04
Location + Scale Shift	-2.35°	0.80 ± 0.03	0.61 ± 0.06	0.60 ± 0.07	0.57 ± 0.07	0.65 ± 0.06	0.06 ± 0.04
Location + Scale Shift	1.90°	0.82 ± 0.03	0.61 ± 0.07	0.60 ± 0.07	0.62 ± 0.07	0.65 ± 0.06	0.05 ± 0.04
Location + Scale Shift	-2.15°	0.81 ± 0.03	0.63 ± 0.06	0.62 ± 0.06	0.60 ± 0.07	0.65 ± 0.06	0.06 ± 0.04
Location + Scale Shift	2.10°	0.81 ± 0.03	0.59 ± 0.07	0.57 ± 0.07	0.60 ± 0.07	0.65 ± 0.06	0.05 ± 0.04
Location + Scale Shift	-20.45°	0.11 ± 0.02	-2.37 ± 0.22	-6.27 ± 1.07	-10.4 ± 1.6	0.50 ± 0.07	0.18 ± 0.06
Location + Scale Shift	20.10°	0.11 ± 0.02	-1.74 ± 0.04	-24.2 ± 3.4	-10.1 ± 1.6	0.50 ± 0.07	0.15 ± 0.07

Due to the different behavior of the validation criteria with respect to the applied biases, especially the insensitiveness of some of them, new inter-comparable thresholds for the acceptance of QSAR models, in a precautionary aproach, are here proposed and summarized as:

$$Q^2_{Fn} = 0.70$$
$$\overline{r_m^2} = 0.65$$
$$CCC = 0.85$$

(It is important to note that CCC is more or less comparable to the square root of the other validation criteria: this is why its threshold is relatively high)

CONCLUSIONS

- ✓ $Q^2_{F1,2}$ and averaged r^2_m, in accepting models as predictive, are not very sensitive for some of the biased simulated scenario.
- ✓ Only CCC and Q^2_{F3} showed to be balanced respect to RMSE in all the simulated biased scenarios.
- ✓ New inter-comparable thresholds are here proposed for QSAR model validation.
- ✓ CCC allows to determine the minimum acceptable number of external elements for hypothesis test.
- ✓ For a better validation, a set of criteria and the scatter plots should be always verified [10] (as implemented in QSARINS [13])

REFERENCES

[1] Tropsha et al. The importance of Being Earnest: Validation in the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Comb. Sci.* 2003, 22, 69-76.

[2] Gramatica. Principles of QSAR models validation: internal and external. *QSAR & Comb. Sci.* 2007, 5, 694-701.

[3] Shi et al. QSAR Models Using a Large Diverse Set of Estrogens. *J. Chem. Inf. Comput. Sci.* 2001, 41, 186-195.

[4] Schüürmann et al. External Validation and Prediction Employing the Predictive Squared Correlation Coefficients Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf. Model.* 2008, 48, 2140–2145.

[5] Consonni et al. Comments on the Definition of the Q² Parameter for QSAR Validation. *J. Chem. Inf. Model.* 2009, 49, 1669-1678.

[6] Ojha, P.K.; Mitra, I.; Das, R.N.; Roy, K. Further exploring r2m metrics for validation of QSPR models. *Chemom. Intell. Lab. Syst.* 2011, *107*, 194-205.

[7] Golbraikh and Tropsha. Beware of q2. *J. Mol. Graph. Model.* 2002, 20, 269-276.

[8] Lin. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 1989, 45, 255-268.

[9] Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *J. Chem. Inf. Model.*, 2011, 51, 2320–2335.

[10] Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models. Part 2 New inter-comparable thresholds for different validation criteria and the need for scatter plot inspection . Under revision in *J. Chem. Inf. Model.*

[11] Lin, L. I. Assay Validation Using the Concordance Correlation Coefficient. *Biometrics* 1992, 48, 599-604.

[12] Gramatica, P.; Papa, E. An update of the BCF QSAR model based on theoretical molecular descriptors. *QSAR &Comb. Sci.* 2005, 24, 953-960.

[13] Chirico, N.; Papa, E.; Kovarich, S.; Cassani, S.; Gramatica P. *QSARINS, software for QSAR model development and validation*; University of Insubria, Varese, Italy, 2008-2012. http://www.qsar.it

This work was supported by the European Union through the project CADASTER FP7-ENV-2007-1-212668