Predictive uncertainty may be improved by efficient use of experimental information for QSARs

- Weighting versus averaging in linear regression

Marit Strömberg and Ullrika Sahlin*

Linnaeus University, School of Natural Sciences, Kalmar/Växjö, Sweden.

*corresponding author

Introduction and Aim

The use of QSARs in chemical regulation or other applications of decision making is possible if they provide predictions with acceptable confidence (Comber et al. 2003, Cronin et al. 2003). Confidence is evaluated in terms of a model's predictive ability, which includes a precise assessment of uncertainty in predictions. Uncertainty in predictions from QSAR models arises not only from the strength of the analogy assumption, saying that molecules with similar structure should have similar physiochemical properties, but also from the application of a statistical model/learning algorithm and from the quality of the experimental data (Schultz et al. 2003, Trophsa 2010). Experimental data show variation e.g. from having experiments done at different labs. Even though there may be more than one experimental value for a given compound, QSARs are today mostly developed by using only one experimental value for each compound, selected by expert judgment or as averages (Papa et al. 2009). The question that arises is if consideration of more information in empirical data in QSAR-development may improve the predictive ability of the model. There are examples where differences in quality of measurement methods have been considered by weighting motivated as prior information based on expert judgment (Willighagen et al. 2011).

The **aim** was to compare predictive ability of QSARs developed on several experimental values per compound to QSARs developed on averaged experimental values.

Table 1. QSARs in Papa et al. (2009)

	Model ID			
Characteristic	2	3	5	6
Endpoint	T_M	Log(1/P_i)	LogK_{oa}	LogK_{ow}
Descriptor	X2A	T(O..Br)	T(O..Br)	T(O..Br)
Number of training data	20	28	24	14
Number of multiple measurements	7	6	5	6
Number of test data	5	6	6	6

Table 3. Comparison of LRW and LRAV based on Papa et al.'s (2009) models.

	Model ID			
Statistic	2	3	5	6
difference in log likelihood score	0.08	0.02	0.07	0.80
Kolomogorov Smirnov Stat. (LRAV)	0.51	0.37	0.31	0.37
Kolomogorov Smirnov Stat. (LRW)	0.40	0.37	0.26	0.23

Fig 1

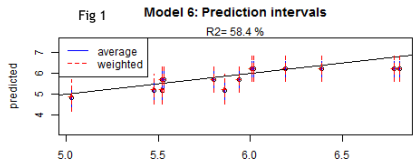
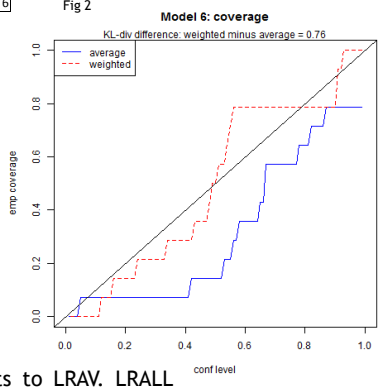


Fig 2



Models and Analysis

Multiple point estimates was considered by building weighted linear regressions with weights assigned such that each chemical had equal contribution to the loss function in the least squares regression. The weighted linear regression (LRW) and the linear regression based on all experimental data (LRALL) were each compared to the linear regression based on averages (LRAV). The modeling approaches ability to predict (including to assess predictive uncertainty) were evaluated by

1) The correlation between predicted and observed values in an external test data set,
2) Empirical coverage to theoretical confidence levels, and
3) Log likelihood scores derived for a common external data set under the corresponding predictive distributions.

Predictive uncertainty was here assessed as a non-parametric distribution by model-based bootstrap.

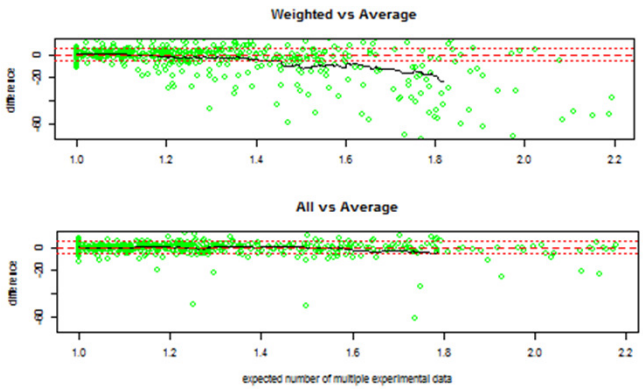
First, the effect of considering more experimental information was evaluated on four QSAR data sets from models developed by Papa et al. (2009) (Table 1). Second, in order to seek generality artificial datasets were constructed (Table 2). The comparisons were done on models judged as having good predictivity on average, which were those with R^2>0.6 for the training data, and where at least one of the approaches succeeded reasonably well in assessing the predictive uncertainty[1]. Differences in performance between modeling approaches were evaluated by the difference in logged likelihood scores, where a difference within 5 is "barely worth mentioning"[2].

[1] Judged as those with a Kolmogorov Smirnov statistic < 0.2, i.e. a significance level of 0.05
[2] According to the decibans scale for Bayes' factor.

Table 2. Characteristics of the artificial QSAR data sets

Model	Formula/Distribution
Endpoint values, Y =	X·B + random error + model error
Descriptor values, X	uniform(0,1)
Regression coefficients, B	uniform(0,1)
Random errors	normal(0,1) ·chisquare(1,s)
Model errors	normal(0,e)
Number of measurements per chemical	bionomial(m,p)
Characteristic	**Range**
Probability of multiple measurements, p	0-0.3
Size of multiple measurements, m	1-5
Size of training data, n	10-50
Random error, s	0.01-0.3
Model error, e	0.01-0.3
Number of descriptors, k	1-4

Fig 3. Differences in log likelihood score with a trend shown by moving average. The dotted lines indicate the zone where the differences between modeling approaches are "barely worth to mention". A negative difference favors using averages of experimental values.



Results

LRW rendered identical regression coefficients to LRAV. LRALL gave slightly different regression parameters (Fig 1). The estimates of model error and thereby uncertainty in predictions for the three models were all different (Fig 2). All of the four models by Papa et al. (2009) LRW showed an improved predictivity as compared to LRAV (Table 3) indicating that uncertainty in QSAR predictions may be improved by using weighting instead of averaging.

Regarding the QSAR models based on the artificial data sets none of the three modeling approaches had always better predictive performance than the others, and most differences in models' prediction ability were within the "barely worth mentioning" zone (Fig 3). LRW performed on average worse than LRAV, and the performance got worse with increasing expected number of experimental values per compound (p-value less than 0.001). LRALL had a slightly lowered performance, compared to LRAV, with increasing expected experimental values as well (p-value less than 0.01). Neither the number of compounds per descriptor nor expected total variance influenced the relative performances of the models.

Conclusion

The general conclusion is that of the three investigated model types there is no specific model type that always is in favor in terms of model predictivity, and which approach that is best depends on the specific data set. Therefore it could be worthwhile to consider all three types when developing a QSAR by linear regression.

Acknowledgement
This study was funded by the FP7 project CADASTER (grant agreement 212668). For more information, visit www.cadaster.eu

References
Comber et al (2003). Environ. Tox. and Chemistry, 22(8): 1822-1828; Cronin et al (2003). Environ. Health Persp. 111(10): 1376-1390; Papa et al (2009). QSAR & Comb Science. 28(8): 790-796; Schultz and Cronin (2003). Environ. Tox. and Chemistry. 22(3): 599-607; Trophsa (2010). Mol. Inf, 29: 476-488; Willighagen et al (2011). J. Biomedical Semantics. 2(Suppl 1):S6