Is the Cook's Distance useful in Refining the Quantitative Structure-Property Linear Models? – The Case of Partition Coefficient

Sorana BOLBOACĂ

"Iuliu Hatieganu" University of Medicine and Pharmacy Cluj-Napoca

Outline

- Aim
- Material and Methods
- Results
- Conclusion

Aim

- QSPR = Quantitative Structure-Property Relationship
- Partition coefficient (logP): lipophilicity governs both pharmacokinetics and pharmacodynamics of drugs.
- *In silico* molecular modeling

- Research aim:
  - to identify if and how withdrawing of influential compound(s) identified using the Cook's distance affect the characteristics of QSPRs models in case of partition coefficient

MATERIAL AND METHODS

Data set	Characteristics
organohalogen compounds	Set abbreviation: OC n (sample size) = 207 Molecular descriptors that relate to electronic structure: E_{HOMO} (Highest Occupied Molecular Orbital energy), E_{LUMO} (Lowest Unoccupied Molecular Orbital energy)
aliphatic organic compounds	Set abbreviation: AC n (sample size) = 125 I_{SET}: semi-empirical electrotopological index (takes into account the charges of the heteroatom and the carbon atoms attached to them through the definition of an equivalent local dipole moment)

Material and Methods

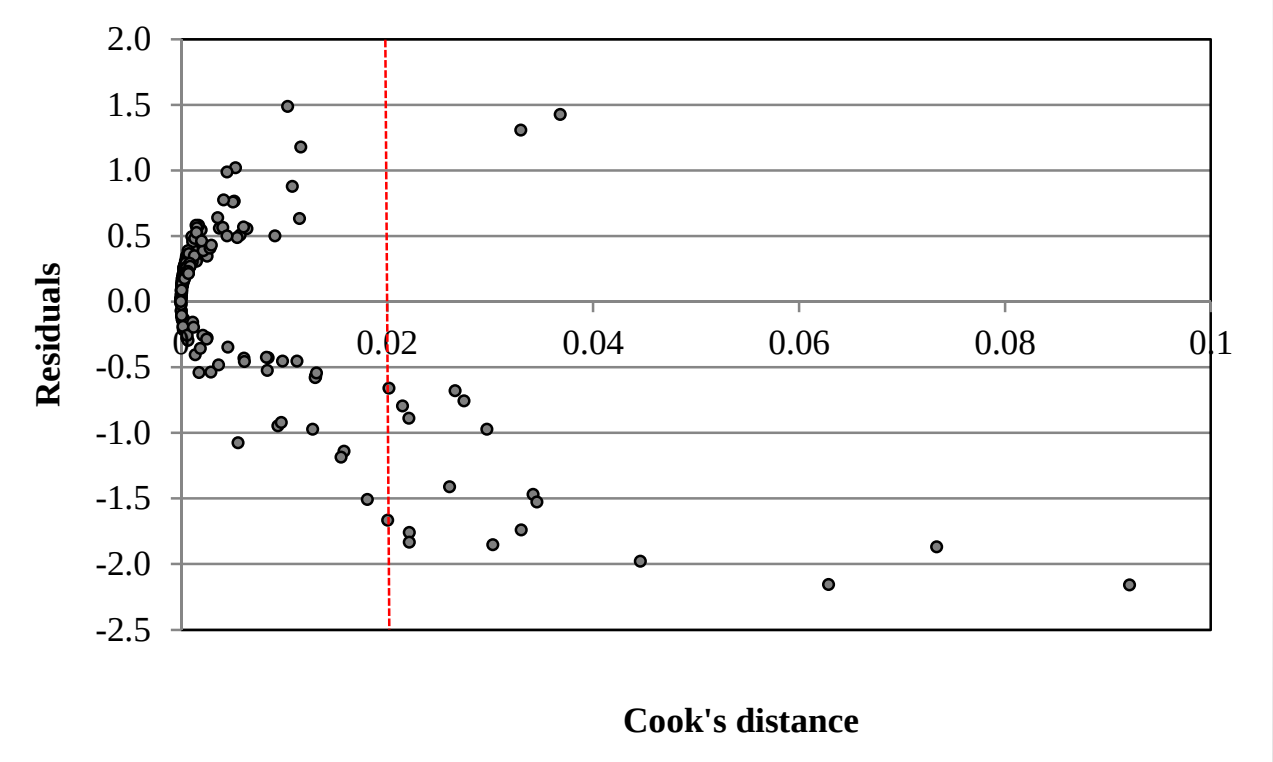
- Test normality - Chi-Squared test
- Identify outlier(s) - Grubb's test
- Identify influential(s) - Cook's distance - $D_i > 4/n$
- Construct and evaluate QSAR model – R^2 & R_{adj}^2 &s & F-value (p-value)
- Compare the full model with D_i-model in terms of correlation coefficient – Steiger's test – 5% significance level

Material and Methods

- Validate the models:
  - 8 statistical parameters
  - Training (~2/3·n compounds, n = sample size) vs. Test experiment
- Test the overall performances of Cook's distance method using Fisher's Chi-Squared test

RESULTS

- Cook's distance as method to identify influential (threshold equal to 0.02): organohalogen compounds.



RESULTS

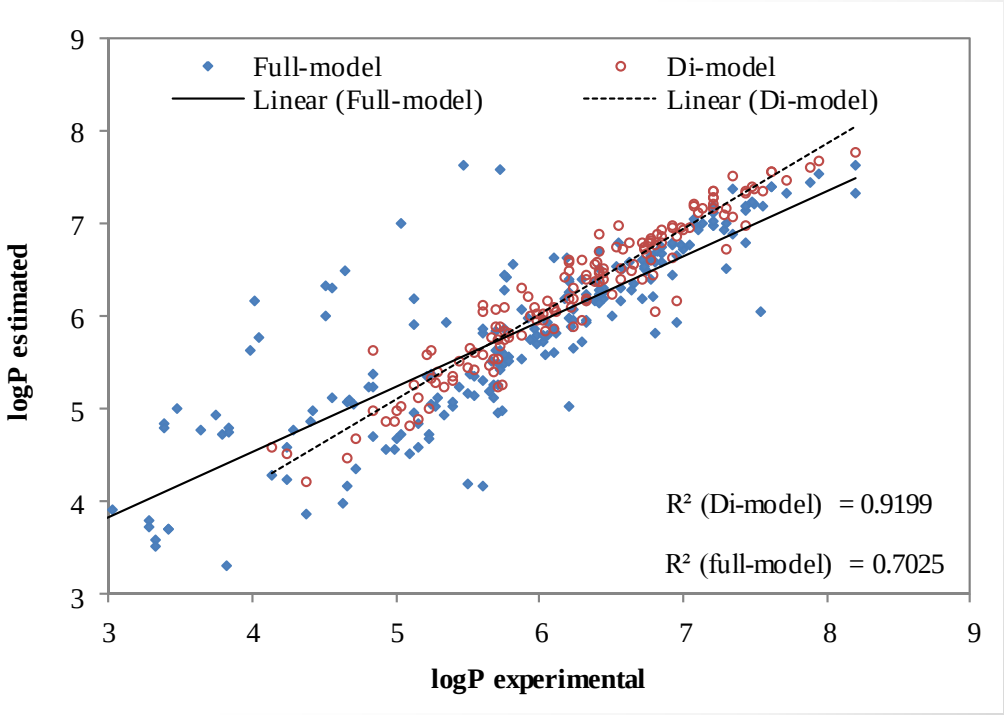
- Influential
  - Organohalogen set: 53 compounds (26%, 95%CI [20%-32%])
  - Aliphatic organic set: 33 compounds (27%, 95%CI [19%-35%])

Parameter	OC-full (n=207)	OC-D_i (n=154)	AC-full (n=124)	AC-D_i (n=91)
R^2	0.7025	0.9199	0.4259	0.9150
R^2_{adj}	0.6996	0.9188	0.4212	0.9140
s	0.6294	0.2302	1.2925	0.3211
F-value	241*	867*	90*	958*
R^2_{loo}	0.6928	0.9162	0.4077	0.9118
s_{loo}	0.6396	0.2355	1.3139	0.3270
F_{loo}	230*	825*	84*	920*

R^2 = determination coefficient; R^2_{adj} = adjusted determination coefficient; s = standard error of estimated; F=value = statistics of Fisher's test; R^2_{loo} = determination coefficient in leave-one-out analysis; s = standard error of predicted; F_{loo} = Fisher's statistics in leave-one-out analysis; * = p-value < 0.0001

RESULTS

- The correlation coefficient proved significantly higher ($p < 0.0001$) in D_i-model compared to full-model for both investigated sets, organohalogen compounds (Z-statistics = 6.704) and aliphatic organic compounds (Z-statistics = 8.022).



RESULTS: MODELS VALIDATION

Parameter	Organohalogen compounds		Aliphatic organic compounds	
	Full-model	D_i-model	Full-model	D_i-model
MAE	0.7082	0.7124	0.7622	0.8190
MAPE	0.0874	0.0267	0.9086	0.3074
SEP	1	1	1	1
REP	17.05	15.89	41.41	52.90
RMSE	1.0049	1.0066	1.0082	1.0112
APV	1.0146	1.0197	1.0245	1.0335
APMSE	0.0049	0.0067	0.0083	0.0115
%PredErr	17.88	18.66	15.38	14.73

MAE = mean absolute error; MAPE = mean absolute percentage error;
SEP = standard error of prediction; REP = relative error of prediction;
RMSE = root-mean-square error; APV = average prediction variance;
APMSE = average prediction mean squared error;
%PredErr = percentage prediction error

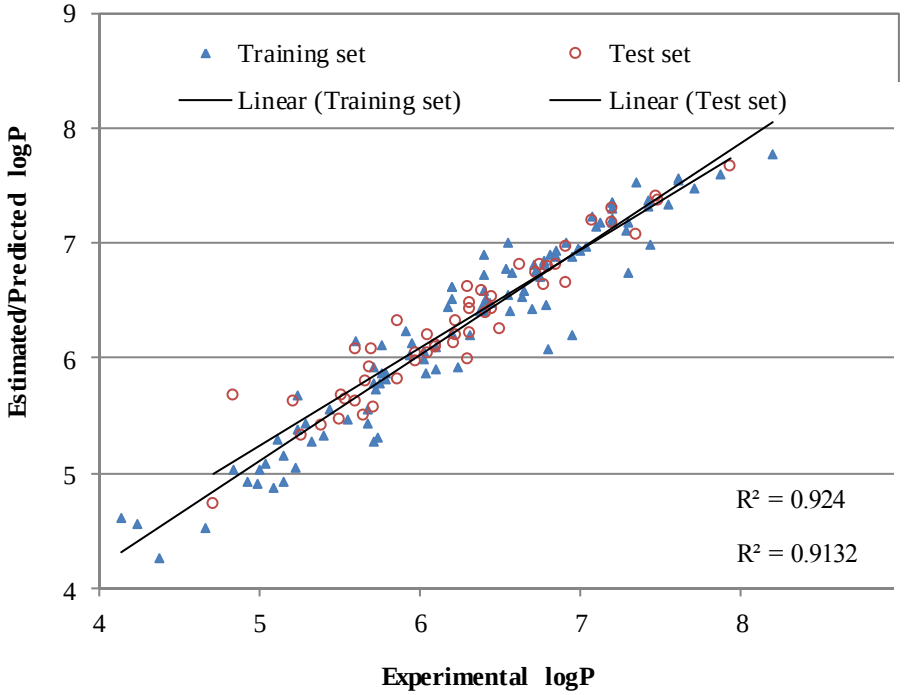
RESULTS: TRAINING VS TEST

Stat	Organohalogen compounds		Aliphatic organic compounds	
	Training	Test	Training	Test
n	103	51	61	30
R^2	0.9240	0.9132	0.9086	0.9249
F-value	608	226	586	309
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001

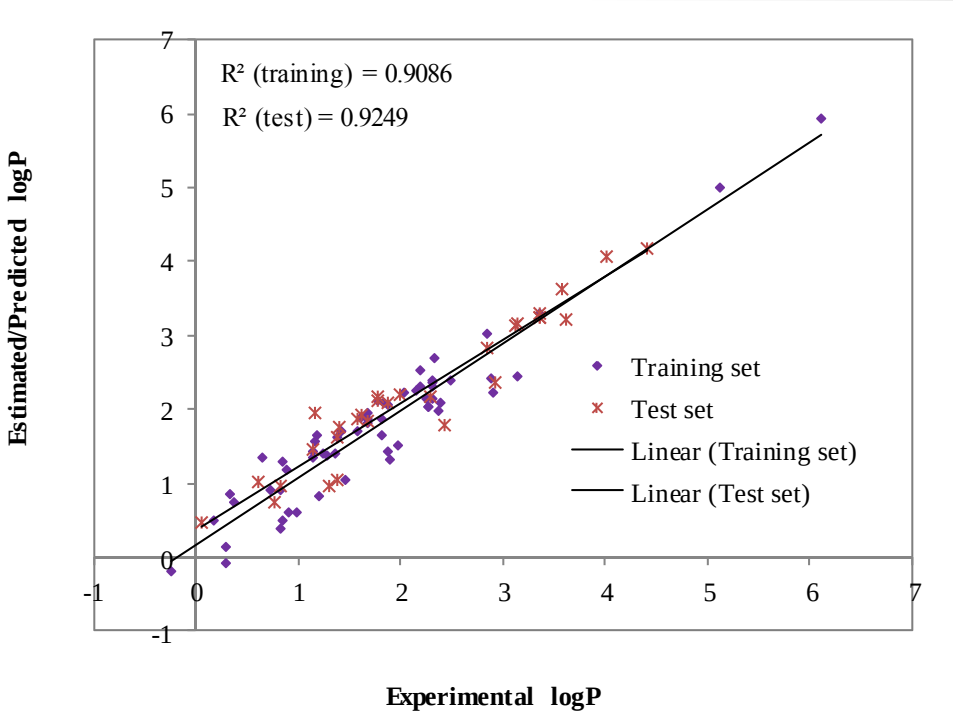
n = sample size; R^2 = determination coefficient;
F-value = Fisher's statistics; p-value = significance level

RESULTS

organohalogen compounds



aliphatic organic compounds



Conclusion

- Cook's distance approach proved able to identify those compounds with significant influence on the QSPR models in investigation of partition coefficient as function of descriptors on organohalogen and aliphatic organic compounds.

- Question(s) …
  - Is this behavior the same unconcerned the sample(s) when partition coefficient is of interest?
  - How the influential are different by the compounds in the sample?

THANK YOU FOR ATTENTION!