# HOW THE METHOD OF STATISTICAL MODELING DOES DETERMINATE ESTIMATION OF CHLORONAPHTHALENE HALF-LIVES IN DIFFERENT ENVIRONMENTAL MEDIA?

Ziemecki P, Puzyn T, Falandysz J

Department of Environmental Chemistry and Ecotoxicology, University of Gdańsk, Sobieskiego 18, Gdańsk 80-952, Poland, e-mail: puzi@pcb.chem.univ.gda.pl

**Introduction**

Polychlorinated naphthalenes (chloronaphthalenes, PCNs, CNs) are widely dispread pollutants.[1-2] There are published many studies on theirs possible anthropogenic and natural sources, environmental concentrations, toxic modes of action as much as various physical and chemical properties.[1-4] One of the most important property of the mentioned compounds is theirs persistency in different environmental media, such as air, water and soil. According to UN-POPs Protocol, those chemicals, for which half-lives ($t_{1/2}$) in the individual environmental compartments passed the established thresholds, are qualified as 'persistent', and they are subjected to the restricted regulations of the protocol.[5] Because of the fact, that the half-lives have never been experimentally determined for all of CN congeners, the QSPR strategy was used first time to calculate them. Quantitative Structure – Property Relationships (QSPR) model express the modeled activity or property as a mathematical function of the molecular structure. There are also many examples of successful implementation of this methodology in case of prediction such properties for POPs.[6-8]

One of the crucial steps of the model identification is a correct choice of the statistical method of modeling. Estimation of missing data in such studies is often done with the statistical approaches, such as: multiple linear regression (MLR), principal component regression (PCR), partial least square regression (PLS), partial least square regression with initial elimination of the uninformative variables (UVE-PLS), partial least square regression with variable selection using a genetic algorithm (GA-PLS).[9] In this study a comparison of efficiency of these five approaches was made, and, in this way, the optimal method of half-life determination for PCNs was chosen.

**Materials and methods**

Initially, a set of 26 structural descriptors was computed for each of 75 chloronaphthalene congener at the level of the density functional theory (B3LYP hybrid functional) in the 6-311++G** basis set.[10] Simultaneously, data on $t_{1/2}$ in air, water and soil available for the other, structurally similar persistent organic pollutants were collected from previous published papers and critically evaluated.[11-19] After evaluation, a set of 92 compounds was selected for further modeling and divided into a training and a validation subsets. For all of these selected compounds the same structural descriptors at the level of B3LYP/6-311++G** were also calculated. Next the models were constructed in turn for each of medium: air, water, soil separately by means of the five following statistical approaches: MLR, PCR, PLS, UVE-PLS, and GA-PLS. Each model was developed using the same training set and each validated using the same validation set. Applicability domains of the individual models were verified by means of principal components ranges. In case of each modeled property quantitative comparison of the models was done using the values of root mean square error in the validation set (RMSEP) as a measure of predictive ability and the number of used descriptors as a criterion of complexity of the studied models. These comparisons led to the final choice of optimal models for predictions in each case.

**Results and discussion**

The results of the quantitative comparison of the studied models are presented in figures 1 and 2. As it could be observed, GA-PLS (for $t_{1/2}$ in air and soil) and MLR (for $t_{1/2}$ in water) models were characterized by the best predictive ability. The value of RMSEP for the GA-PLS model of $t_{1/2}$ in water was also relatively low. Analyzing complexity of the models it could be clearly state, that the MLR models in each case were characterized by the lowest number of input variables (only one), while PCR and PLS used all 26 molecular descriptors.

Implementation of UVE and GA algorithms for initial variable selection in PLS model resulted in significant reduction of the lowest informative descriptors.

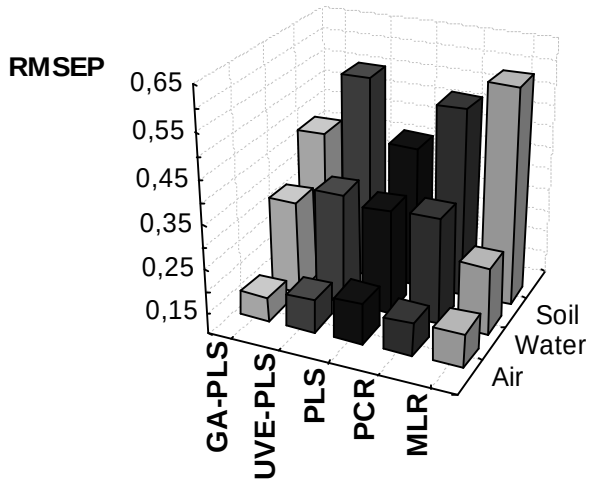


Figure 1. Predictive ability of the models (value of RMSEP)

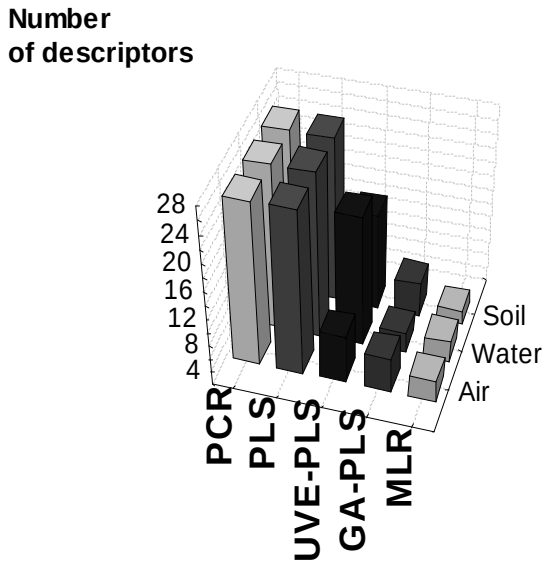


Figure 2. Complexity of the models (number of independent variables).

Finally, the GA-PLS approach was proposed for estimation of half-lives of PCNs in the individual environmental media. This methodology led to the significant reduction in descriptors and, simultaneously, it improved the prognostic quality of the PLS model. The disadvantage of MLR method is the necessity of choosing *a priori* the best set of explanatory variables as well as the sensitivity to interconnection between the descriptors. From the other hand, its advantage is an ability to obtain quite good model using a simple and good mechanistic interpretable procedure.

Comparing the types of molecular descriptors chosen in each case by the GA-PLS model it could be stated that they seem to be connected with the main physical-chemical processes of degradation in the individual medium. The descriptors such as: energy of the highest unoccupied molecular orbital (LUMO), total energy of the molecule (Et), solvent accessible surface in water (SASw), solvent accessible volume in water (SAVw), total electrostatic energy of solvation in water (TEESolw), solvent accessible volume in octanol (SAVo), cavitation energy in octanol (CEo), total non-electrostatic energy of solvation in octanol (TNEo) were the most important for half-lives prediction in air. They are corresponding to the photochemical and radical degradation, partitioning to the air-born organic particles and, wet and dry deposition. Degradation of chloronaphthalenes in water was described by descriptors correlated with solubility in water, polarity and photolytic degradation, such as: dipole moment (D), polarized solute-solution interaction energy in water (PolSSw), LUMO, TEESolw, PolSSw and CEw. Descriptors: energy of the highest occupied molecular orbital (HOMO), LUMO, SAVw, PolSSw, SASo, SAVo and TNEo used in the model for soil probably correspond to microbial and photolytic degradation, partitioning between organic phase and water as much as vaporization. Reliability of the descriptors used by the model and possibility of its mechanistic interpretation additionally confirmed high usefulness of the GA-NN approach in this exercise.

**Acknowledgments**

Dr. Tomasz Puzyn is the recipient of a fellowship from the Foundation for the Polish Science. Computations were carried out using computers in the TASK - Academic Computer Center in Gdańsk. The research project was funded by the University of Gdańsk (grant no. BW-8000-5-0304-6).

**References**

1. Falandysz J. *Food Addit Contam* 2003;20:995.
2. Falandysz J. *Environ Pollut* 1998;10:7790.
3. Harner T, Bidleman TF, Jantunen LMM, Mackay D. *Environ Toxicol Chem* 2001;20:1612.
4. Wania F. *Environ Sci Technol* 2003;37:1344.
5. Lerche D, van de Plassche E, Schwleger A, Balk F. *Chemosphere* 2002;47:617.
6. Puzyn T, Falandysz J. *Atmos Environ* 2005;39:1439.
7. Puzyn T, Falandysz J. *J Environ Sci Health* 2003;38A:1761.
8. Falandysz J, Puzyn T, Szymanowska B, Kawano M, Markuszewski M, Kaliszan R, Skurski P, Błażejowski J. *Pol J Environ Std* 2001;10:217.
9. Mazerski J. *Podstawy chemometrii*, Wydawnictwo Politechniki Gdańskiej, Gdańsk, Poland, 2000.
10. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima Y, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. GAUSSIAN 03. Gaussian Inc. Pittsburgh, 2003.
11. Beyer A, Mackay D, Matthies M, Wania F, WebsterE. *Environ Sci Tech* 2000;34:699.
12. Gramatica P, Consolaro F, Pozzi S. *Chemosphere* 2001;43:665.
13. Hirai Y, Sakai S, Watanabe N, Takatsuki H. *Chemosphere* 2004;54:1383.
14. Chen J, Peijnenburg WJGM, Quan X, Chen S, Martens D, Schramm KW, Kettrup A. *Environ Pollut* 2001;114:137.
15. PAN Pesticides Database, US EPA, availavle on-line: http://www.pesticideinfo.org
16. Gouin T, Harner T . *Environ Internat* 2003;29:717.
17. Pennington DW. *Chemosphere* 2001;44:1617.
18. Pennington D.W. *Chemosphere* 2001;44:1589.
19. Cousin IT, Jones KC. *Environ Pollution* 1998;102:105.