STUDY OF THE POP ATMOSPHERIC MOBILITY BY QSAR APPROACH

Paola Gramatica, Stefano Pozzi, Federica Consolaro and Roberto Todeschini*

QSAR Research Unit, Dep. Struct. Funct Biol., University of Insubria, via Dunant 3, 21100 Varese (Italy) ; http://andromeda.varbio.unimi.it/~QSAR; e-mail:paola.gramatica@unimi.it
* Milano Chemometric Res. Group, Dep. Environ. Sci., University of Milano-Bicocca, Milano (Italy)

Introduction

Persistent Organic Pollutants (POPs) particularly PAH, PCB, polychlorinated dibenzo-*p*-dioxins and pesticides, are chemicals dangerous to the environment. The hazard for these compounds is due to both their toxicity for different organisms and their environmental fate, mainly persistence, bioaccumulation, etc. POP environmental fate mainly depends on a variety of physical and chemical processes, that can be studied through properties such as vapour pressure (vp), Henry's law constant (H), n-octanol/water partition coefficient (K_{ow}), soil sorption coefficient (K_{oc}), water solubility (S), atmospheric half-life, etc. Unfortunately, for a large number of POPs the experimental data for several properties remain unknown, thus regression models, according to the QSAR/QSPR strategies, needed to be developed to predict the missing values.
A rank of POPs, according to their atmospheric mobility or environmental fate is possible by multivariate approaches (PCA and Multicriteria Decision Making) based on these predicted data.
A classification of POP in mobility classes by few molecular descriptors will allow a fast screening of large data bases.

Molecular descriptors and Methods

The experimental data of several physico-chemical properties are lacking for a large number of POPs, thus there is a need to develop statistical models to predict such physico-chemical properties (boiling point, melting point, log Kow, log Koc, Henry's law constant, TSA, Vmol, water solubility, vapour pressure) and the atmospheric half-life for these compounds; this can be done by the QSAR/QSPR approach, using different kinds of molecular descriptors for the structural representation of the studied compounds.
Molecular descriptors represent the way chemical information contained in the molecular structure is transformed and coded, taking different aspects of the chemical information into account for QSAR and QSPR studies. Among the theoretical descriptors the best known, obtained simply from the knowledge of the formula are: molecular weight and count descriptors (1D-descriptors, i. e. counting of bonds, atoms of different kind, presence or counting of functional groups and fragments, etc.). Graph-invariant descriptors (2D-descriptors including both topological and information indices), are obtained from the knowledge of the molecular topology. WHIM

molecular descriptors [1] contain information about the whole 3D-molecular structure in terms of size, symmetry and atom distribution. All these indices are calculated [2] from the (x,y,z)-coordinates of a three-dimensional structure of a molecule, usually from a spatial conformation of minimum energy: 37 non-directional (or global) and 66 directional WHIM descriptors are obtained. A complete set of about two hundred molecular descriptors has been obtained.
Being our representation of a chemical based on a lot of molecular descriptors, an effective variable selection strategy GA-VSS (Genetic Algorithm - Variable Subset Selection) was applied to the whole set of descriptors in order to set out the most relevant variables in modelling the POP properties by Ordinary Least Squares regression (OLS), maximising the predictive power (Q^2_{LOO}) [3].
Models with good predictive performances (Q^2_{LOO} = 78-96%) are obtained for all the physico-chemical properties and the atmospheric half-life, thus achieving reliable data for 87 compounds. Principal Component Analysis (PCA) is the multivariate technique used for data exploration. The MultiCriteria Decision Making (MCDM) strategy [4], particularly the desirability functions, are then applied to perform a POP mobility ranking.
Classification And Regression Tree (CART), Regular Discriminant Analysis (RDA) and K-Nearest Neighbours (K-NN) [5], are the classification strategies used to classify POPs according to their environmental fate. Stepwise Linear Discriminant Analysis (SLDA) is used for the variable subset selection in classification. To check the classification model prediction quality, the Misclassification Risk in prediction (MRcv%), obtained with the *leave-one-out* procedure, is used. A comparison with the No-model Misclassification Risk is used to evaluate the performances of the models.

Results and discussion

The biplot of the principal component analysis (Figure 1) for 87 POP (Tab. 1), described by the more relevant physico-chemical properties (boiling point, melting point, log Kow, log Koc, Henry's law constant, TSA, Vmol, water solubility, vapour pressure) and the atmospheric half-life, shows a particular distribution of compounds along the first component (PC1, EV = 70.4%), in good accordance with their mobility-class assigned by Mackay and Wania [6] by a univariate approach.

Consequently, it is possible to use PC1, where a linear combination of many properties is considered, to classify all 87 POPs in one of four *a priori* classes of mobility (high, relatively high, relatively low and low mobility) by a more correct multivariate approach.
Some factors must be considered for their influence on the atmospheric mobility and generally environmental fate, particularly the compounds half-life, represented only in the second component in PCA analysis, and their sorption in the atmospheric particles.

ID	NAME	CLA	ID	NAME	CLA	ID	NAME	CLA
1	indan	1	30	naphthacene	4	59	PCB-47	2
2	naphtalene	1	31	benz[a]anthracene	3	60	PCB-52	2
3	1-methylnaphthalene	1	32	benzo[b]fluoranthene	4	61	PCB-53	2
4	2-methylnaphthalene	1	33	benzo[j]fluoranthene	4	62	PCB-54	2
5	1.2-dimethylnaphthalene	2	34	benzo[a]pyrene	4	63	PCB-77	2
6	1.3-dimethylnaphthalene	2	35	benzo[e]pyrene	4	64	PCB-86	2
7	1.4-dimethylnaphthalene	2	36	perylene	4	65	PCB-87	2
8	1.5-dimethylnaphthalene	2	37	7.12-dimethylbenz[a]anthracene	4	66	PCB-101	2
9	2.3-dimethylnaphthalene	2	38	3-methylcholanthrene	4	67	PCB-104	2
10	2-6-dimethylnaphthalene	2	39	benzo[g,h,i]perylene	4	68	PCB-128	2or3
11	1-ethylnaphthalene	2	40	dibenz[a,h]anthracene	4	69	PCB-153	2or3
12	2-ethylnaphthalene	2	41	biphenyl (PCB-0)	1or2	70	PCB-155	3
13	1.4.5-trimethylnaphthalene	2	42	PCB-1	2	71	PCB-171	3
14	acenaphthene	2	43	PCB-2	2	72	PCB-194	3
15	acenaphthylene	2	44	PCB-3	2	73	PCB-202	3
16	fluorene	2	45	PCB-4	2	74	PCB-206	3
17	1-methylfluorene	2	46	PCB-7	2	75	PCB-209	3
18	anthracene	2	47	PCB-9	2	76	α-HCH	1
19	2-methylanthracene	3	48	PCB-11	2	77	y-HCH	1
20	9-methylanthracene	3	49	PCB-12	2	78	p,p'-DDT	3
21	9-10-dimethylanthracene	3	50	PCB-14	2	79	p,p'-DDE	3
22	phenanthrene	2	51	PCB-15	2	80	p,p'-DDD	3
23	1-methylphenanthrene	3	52	PCB-18	2	81	Chlordane	3
24	fluoranthene	3	53	PCB-26	2	82	Dieldrin	3
25	1-2-benzofluorene	3	54	PCB-28	2	83	2,3,7,8-tetraCl-dibenzo-*p*-dioxin	3
26	2-3-benzofluorene	3	55	PCB-29	2	84	1,2,3,4,7,8-hexaCl-dibenzo-*p*-dioxin	3
27	pyrene	3	56	PCB-30	2	85	octaCl-dibenzo-*p*-dioxin	3
28	chrysene	3	57	PCB-33	2	86	Pentachlorobenzene	1
29	triphenylene	3	58	PCB-40	2	87	Hexachlorobenzene	1

Table 1: list of studied compounds and assigned mobility-classes. (1:high mobility, 4:low mobility)

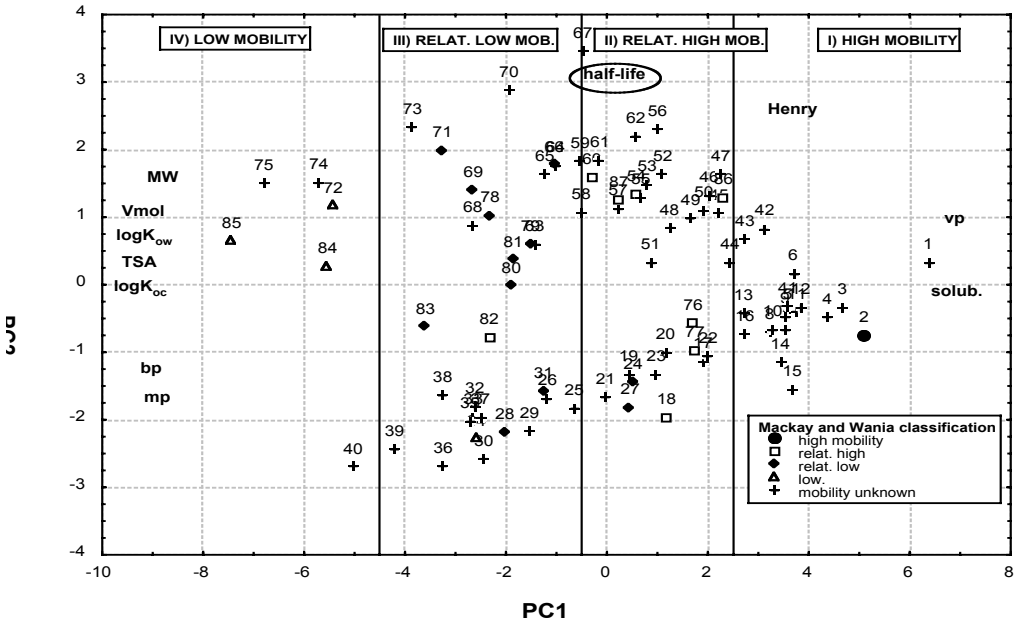

Figure 1: Principal component analysis on the physico-chemical data (EV=88.8%)

A chemometric strategy known as "Multicriteria Decision Making", in particular the desirability functions approach, was used for this purpose. POPs with low mobility are here considered as the most desirable. The used criteria are:

- the first principal component (PC1) score values as mobility indicator (optimum = low values)
- the logKoc as atmospheric particle sorption indicator (optimum = high values)
- the half-life values (optimum = low values)

In figure 2 the desirability values were plotted in function of the molecule number; the compound mobility trend seems to be very similar to the real world distribution.

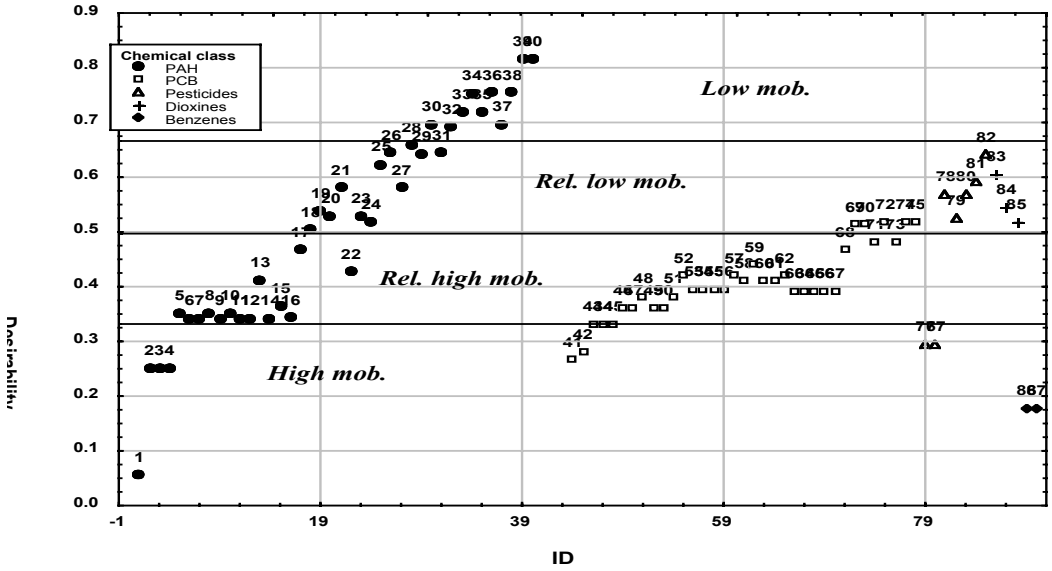

Figure 2: Plot of desirability values in function of the molecule number

POP classification according to their environmental fate was finally made by three different classification methods (Classification And Regression Tree, CART, K-Nearest Neighbours, K-NN, and Regular Discriminant Analysis, RDA). The prior classes were obtained from the desirability values: high mobility (1) < 0.33, relatively high (2) = 0.330-0.5, relatively low (3) = 0.5- 0.67, low mobility (4) > 0.67.

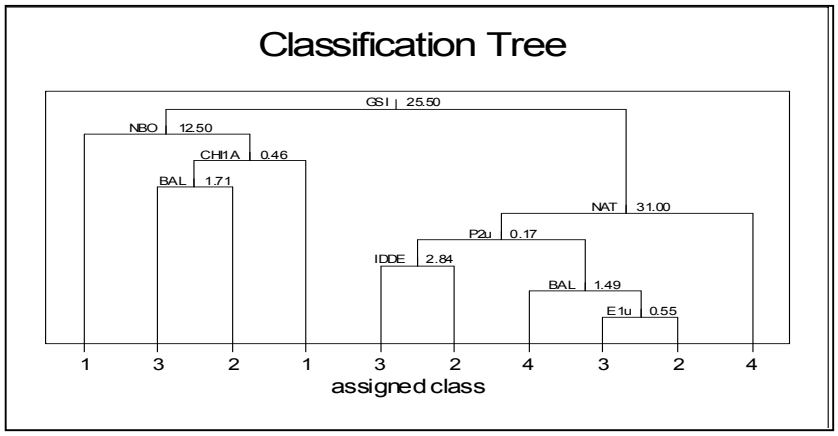

Figure 3: Mobility classification Tree obtained with the CART method (MR_{cv} = 12.6%).

All the classification methods give models with satisfactory and congruent predictions. The simplest model, and consequently the most directly applicable, is the one developed with CART (Figure 3): the selected descriptors are mainly related to molecular size.

In conclusion, different objectives have been realized in this paper:
- The values of the principal physico-chemical properties, not experimentally available for all the compounds, are here predicted by validated regression models in a Quantitative Structure-Property Relationships (QSPR) approach, based on different molecular descriptors.
- A rank of POPs according to their tendency towards the atmospheric mobility and generally their environmental fate is realized by two multivariate approaches: a) Principal Component Analysis of predicted physico-chemical data; b) Multicriteria Decision Making, taking into account also the atmospheric half-life and the sorption in the atmospheric particles.
- Finally POPs are classified in 4 mobility classes by several Classification methods, thus POP environmental fate can be easily assessed from the molecular structure alone.

Acknowledgements

We thank Prof. Davide Calamari for the reviewing of this paper and for his helpful suggestions.

References

[1] Todeschini R. and Gramatica P.; *Quant.Struct.-Act.Relat*. **1997**, 16, 113-119

[2] Todeschini R. *WHIM-3D / QSAR - Software for the calculation of the WHIM descriptors*. rel. 4.1 for Windows, Talete srl, Milan (Italy) **1996.** Download: http://www.disat.unimi.it/chm.

[3] Todeschini R. *Moby Digs - Software for Variable Subset Selection by Genetic Algorithms*. Rel. 1.0 for Windows, Talete srl, Milan (Italy) **1997**.

[4] Hendriks M.M.W.B., De Boer J.H., Smilde A.K. and Doornbos D.A.; *Chemom. Intell. Lab. Syst.* **1992**, 16, 175-191

[5] *SCAN - Software for Chemometric Analysis*, Minitab Inc. (USA), **1995**.

[6] Wania F. and Mackay D.; *Environ. Sci. Technol*. **1996**, *30*, 390-397.