	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **QSARINS (QSAR-INSUBRIA) model of log Koc by PaDEL descriptors Keywords: PaDEL-Descriptor; GA-OLS; Koc; External Validation; QSARINS; INSUBRIA**
	Printing Date: **9-mar-2015**

1.QSAR identifier

1.1.QSAR identifier (title):

QSARINS (QSAR-INSUBRIA) model of log Koc by PaDEL descriptors
Keywords: PaDEL-Descriptor; GA-OLS; Koc; External Validation; QSARINS; INSUBRIA

1.2.Other related models:

Gramatica P., Giani E., Papa E., Statistical external validation and consensus modeling: A QSPR case study for Koc prediction, Journal of Molecular Graphics and Modeling, 2007, 25, 755-766. [1]

1.3.Software coding the model:

PaDEL-Descriptor
A software to calculate molecular descriptors and fingerprints [2], version 2.18
Yap Chun Wei, email: phayapc@nus.edu.sg
http://padel.nus.edu.sg/software/padeldescriptor/index.html

QSARINS
Software for the development, analysis and validation of QSAR MLR models [3,4], version 1.2 (verified also with 2.2, 2015)
Paola Gramatica, email: paola.gramatica@uninsubria.it
http://www.qsar.it/

2.General information

2.1.Date of QMRF:

30/01/2015

2.2.QMRF author(s) and contact details:

[1]Paola Gramatica Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421573 paola.gramatica@uninsubria.it http://www.qsar.it/
[2]Stefano Cassani Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421439 stefano.cassani@uninsubria.it http://www.qsar.it/

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Stefano Cassani Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421439 stefano.cassani@uninsubria.it http://www.qsar.it/
[2]Paola Gramatica Insubria University, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100 Varese (Italy) +390332421573 paola.gramatica@uninsubria.it http://www.qsar.it/

2.6.Date of model development and/or publication:

Developed in 2013, Published in 2014

2.7.Reference(s) to main scientific papers and/or software package:

[1]Gramatica P., Giani E., Papa E., Statistical external validation and consensus modeling: A QSPR case study for Koc prediction, Journal of Molecular Graphics and Modeling, 2007, 25, 755-766. doi:10.1016/j.jmgm.2006.06.005

[2]Yap, C.W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. 2011, J.Comput.Chem. 32, 1466-1474 doi: 10.1002/jcc.21707

[3]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J. Comput. Chem. (Software News and Updates), 2013, 34 (24), 2121-2132. DOI: 10.1002/jcc.23361

[4]Gramatica P., et al. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, J. Comput. Chem. (Software News and Updates), 2014, 35 (13), 1036-1044. DOI: 10.1002/jcc.23576

2.8.Availability of information about the model:

Non-proprietary. Defined algorithm, available in QSARINS [3,4]. Training and prediction sets are available in the attached sdf files of this QMRF (section 9) and in the QSARINS-Chem database [4].

2.9.Availability of another QMRF for exactly the same model:

No

3.Defining the endpoint - OECD Principle 1

3.1.Species:

No information available

3.2.Endpoint:

QMRF 2. Environmental fate parameters QMRF 2. 6. Partition coefficient. Organic carbon-sorption partition coefficient (organic carbon; Koc)

3.3.Comment on endpoint:

The soil sorption partition coefficient is expressed as the ratio between chemical concentration in soil and in water, normalized on organic carbon (Koc). This parameter is an indicator of the sorption of chemicals by soils and sediments, thus providing an estimation of compound mobility and persistence in these compartments.The Koc experimental data for 643 heterogeneous organic compounds were collected from literature [5-7] and compiled into a single dataset. These three references were not the primary source of the experimental data, but a collection of previous literature data, which were already used to develope published and good-quality QSPR models.

3.4.Endpoint units:

Dimensionless

3.5.Dependent variable:

log Koc

3.6.Experimental protocol:

No information available

3.7.Endpoint data quality and variability:

As normal in QSAR studies, QSAR modelers take data from various sources and collect it into a single large dataset. As stated in section 3.3, we used data already satisfactorily modelled and verified for their goodness (data curation) from different authors [5-7]: previous results are a proof of data quality. If more than one Koc value was available for a single compound, the median of the values was used.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSPR - Multiple linear regression model (OLS - Ordinary Least Square)

4.2.Explicit algorithm:

Log Koc model (Split Model)

MLR-OLS method. Model developed on a training set of 93 compounds.

LogKoc model (Full Model)

MLR-OLS method. Model developed on all the available experimental data (training set of 643 compounds).

Split model equation (N Training: 93): logKoc= 0.63 + 0.28 VP-0 - 0.26 nHBAcc + 0.09 nAromBond - 0.19 MAXDP

Full model equation (N Training: 643): logKoc= 0.87 + 0.26 VP-0 - 0.23 nHBAcc + 0.08 nAromBond - 0.19 MAXDP

The modeling descriptors, calculated in PaDEL-Descriptor 2.18, are: VP-0, nHBAcc, nAromBond, MAXDP. See section 4.3 for a more detailed description of the four descriptors.

4.3.Descriptors in the model:

[1]VP-0 dimensionless Valence path, order 0. It Is related to molecular size and has a positive sign, highlighting that the bigger compounds are more sorbed than leached

[2]nHBAcc dimensionless Number of hydrogen bond acceptors (using CDK HBondAcceptorCountDescriptor algorithm). It is related to electronegative atoms of molecules and represent a way of taking into account the probability of bond formation between chemicals and groundwater: as expected, this descriptor is negative in sign as high affinity for water precludes soil sorption of the chemicals

[3]nAromBond dimensionless Number of aromatic bonds, with negative sign in the equation (and less relevant descriptor)

[4]MAXDP dimensionless Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule). Using deltaV = (Zv-maxBondedHydrogens) /(atomicNumber-Zv-1). it takes into account the electronic distribution in the topological graph and it is related to molecule electrophilicity. It represent a way of considering the probability of bond formation between chemicals and groundwater: this descriptor is negative in sign (inversely related to logKoc) because high affinity for water precludes soil sorption of the chemicals

4.4.Descriptor selection:

A total of 681 molecular descriptors of differing types (0D, 1D, 2D) were calculated in PaDEL-Descriptor 2.18 [2]. Constant and semi-constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (one of any two descriptors with a correlation

greater than 0.98 was removed to reduce redundant information), and a final set of 169 molecular descriptors were used as input variables for variable subset selection by genetic algorithm (GA-VSS).The models were initially developed by the all-subset-procedure until two variable. Then the GA was applied in order to explore new combinations of variables, selecting the four variables by a mechanism of reproduction/mutation. The optimized parameter used was Q2LOO (leave-one-out). The GA-VSS, by Ordinary Least Squares regression (OLS), included in QSARINS, was applied to select only the best combination of descriptors from input pool: 4 modeling descriptors selected from 169.

4.5.Algorithm and descriptor generation:

Multiple linear regression (Ordinary Least Square method) was applied to generate the model.

Molecular descriptors were generated by PaDEL-Descriptor software (open source) [2]. The input files for descriptor calculation contain information on atom and bond types, connectivity, partial charges and atomic spatial coordinates, relative to the minimum energy conformation of the molecule, and were firstly obtained by the semi empirical AM1 method using the package HYPERCHEM ver. 7.03 [8]. Then, these files were converted by OpenBabel 2.3.0 [9] into MDL-MOL format and used as input for the calculation of descriptors in PaDEL-Descriptor. Any user can re-derives the model calculating the molecular descriptors by PaDEL-Descriptor 2.18 software (included in QSARINS 2.2) and applying the given equation (automatically done by QSARINS 2.2).

4.6.Software name and version for descriptor generation:

PaDEL-Descriptor
A software to calculate molecular descriptors and fingerprints, version 2.18
Yap Chun Wei, email: phayapc@nus.edu.sg
http://padel.nus.edu.sg/software/padeldescriptor/index.html

HyperChem
Software for molecular drawing and conformational energy optimization, version 7.03, 2002.
Phone: (352)371-7744
http://www.hyper.com/

OpenBabel
Open Babel: The Open Source Chemistry Toolbox. Used for conversion between HYPERCHEM files (hin) and MDL-MOL files, version 2.3.0, 2010
openbabel-discuss@lists.sf.net
http://openbabel.org/wiki/Main_Page

4.7.Chemicals/Descriptors ratio:

Split Model: 93 chemicals / 4 descriptors = 23.25
Full Model: 643 chemicals / 4 descriptors = 160.75

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model was verified by the leverage approach and fixed thresholds has been used to define both structural and response outliers (see section 5.4). The plot of leverages (hat diagonals) versus standardised residuals, i.e. the Williams plot, verified the presence of response outliers (i.e.compounds with cross-validated standardized residuals greater than 3 standard deviation units) and chemicals very structurally influential in determining model parameters parameters (i.e. compounds with a leverage value (h) greater than 3p'/n (h*), where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model).

For new compounds without experimental data, leverage can be used as a quantitative measure for evaluating the degree of extrapolation: for compounds with a high leverage value (h > h*), that are structural outliers, predictions should be considered less reliable. In QSARINS the Insubria graph allows to identify for which chemicals the predictions are inter- or extrapolated by the model.

5.2.Method used to assess the applicability domain:

As it has been stated in section 5.1, the structural applicability domain of the model was assessed by the leverage approach, providing a cut-off hat value (h*=0.023). HAT values are calculated as the diagonal elements of the HAT matrix:

$H = X(X^TX)^{-1}X^T$

The response applicability domain can be verified by the standardized residuals in cross-validation greater than 3.0 standard deviation units

5.3.Software name and version for applicability domain assessment:

QSARINS

Software for the development, analysis and validation of QSAR MLR models, version 1.2 (verified also with 2.2)

Paola Gramatica, email: paola.gramatica@uninsubria.it

http://www.qsar.it/

5.4.Limits of applicability:

Split model domain: outliers for structure, hat>0.1613 (h*): Dibenzo(aj)anthracene (224-41-9), 1,2,5,6-dibenzanthracene (53-70-3), Hexabromobiphenyl (36355-01-8), Camphechlor (8001-35-2), Thifensulfuron-methyl (79277-27-3), Metsulfuron-methyl (74223-64-6), Indeno(1,2,3-cd)pyrene (193-39-5), Benzo(ghi)perylene (191-24-2), Tralomethrin (66841-25-6), C.I. Vat Yellow 4 (128-66-5), Chlordecone (143-50-0), Dibenzo(a,i)pyrene (189-55-9), Mirex (2385-85-5). Outliers for response, standardised residuals > 3 standard deviation units: Methylurea (598-50-5), Endothal (145-73-3), Pentachloroaniline (527-20-8), Esfenvalerate (66230-04-4), Quizalofop-ethyl (76578-14-8), Camphechlor (8001-35-2), Isoxaben (82558-50-7), Hexabromobiphenyl (36355-01-8).**FULL model domain**: outliers for structure, hat>0.023 (h*): Cyromazine

(66215-27-8), Thiophanate-methyl (23564-05-8), 1,2,5,6-dibenzanthracene (53-70-3), Dibenzo(aj)anthracene (224-41-9), Chlorsulfuron (64902-72-3), Indeno(1,2,3-cd)pyrene (193-39-5), Benzo(ghi)perylene (191-24-2), Hexabromobiphenyl (36355-01-8), Tralomethrin (66841-25-6), C.I. Vat Yellow 4 (128-66-5), Chlordecone (143-50-0), Camphechlor (8001-35-2), Chlorimuron-ethyl (90982-32-4), Dibenzo(a,i)pyrene (189-55-9), Sulfometuron-methyl (74222-97-2), Thifensulfuron-methyl (79277-27-3), Metsulfuron-methyl (74223-64-6), Mirex (2385-85-5). Outliers for response, standardised residuals > 3 standard deviation units: Isoxaben (82558-50-7).

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: Yes

INChI: No

MOL file: No

6.3.Data for each descriptor variable for the training set:

All

6.4.Data for the dependent variable for the training set:

All

6.5.Other information about the training set:

The training set of the Split Model consists of 93 organic compounds, with a highly heterogeneous chemical space, in fact the compounds includ almost all the principal functional groups. The chemicals are mainly pesticides, but also various organic pollutants are present. In addition the set has a very large range of logKoc values: -0.31 to 6.02. Training and prediction set are structurally balanced, being the splitting based on the structural similarity analysis, performed with Kohonen artificial neural network (K-ANN, or Self Organizing Maps, SOM) method included in KOALA software (Rel. 1.0 for Windows, 2001. R.Todeschini, V. Consonni, A. Mauri, Milan, Italy).

6.6.Pre-processing of data before modelling:

Transformation of Koc into logarithmic units (log Koc). If more than one value was available for a single compound, the average of the values was used. Only processed data are given.

6.7.Statistics for goodness-of-fit:

R^2= 0.84; CCCtr [10,11]=0.91; RMSE= 0.49

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

Q^2LOO= 0.82; CCCcv=0.90; RMSEcv= 0.52

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

$Q^2LMO_{30\%}$= 0.82. High value of Q^2LMO

(average value for 2000 iterations, with 30% of chemicals put out at

every iteration) means that the model is robust and stable.

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2y-sc = 0.04. Very low value of scrambled R^2(average value for 2000 iterations, in where the Y-responses are
randomly scrambled), means that the model is not given by
chance-correlation.

6.11.Robustness - Statistics obtained by bootstrap:

No information available (since we have calculated Q^2LMO)

6.12.Robustness - Statistics obtained by other methods:

No information available

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

CAS RN: Yes
Chemical Name: Yes
Smiles: Yes
Formula: Yes
INChI: No
MOL file: No

7.3.Data for each descriptor variable for the external validation set:

All

7.4.Data for the dependent variable for the external validation set:

All

7.5.Other information about the external validation set:

The prediction set consists of 550 heterogeneous organic compounds with a range of logKoc values from 0 to 6.33. Training and this prediction set are structurally balanced, being the splitting based on the structural similarity analysis performed by SOM (as stated in section 6.5).

7.6.Experimental design of test set:

The splitting of the original data set (643 compounds) into a training set of 93 compounds and a prediction set of 550 compounds was realized by Kohonen artificial neural network (K-ANN or Self Organizing Maps, SOM), using the software KOALA (as reported in sections 6.5 and 7.5). Through its clustering capabilities, SOM ensures that both sets are homogeneously distributed within the entire area of the descriptor space; in this case the chemicals in both sets, selected to maximize the coverage of the descriptor space (i.e. representativity), represent the structural variety of the studied data set in a balanced way. The selected training chemicals are those with the minimal distance from the centroid of each cell in the top map. In this case, the representative points of the prediction set are close (in the same cell of the top map) to representative points of the training set in the multidimensional structural descriptor

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [12]= 0.78; Q^2extF2 [13]= 0.78; Q^2extF3 [14]= 0.79; CCCex=0.89; RMSE= 0.57.

The high values of external $Q^2$and concordance correlation coefficient-CCC (threshold for accepting the external Q^2F1-F2-F3 is 0.70, threshold for CCC is 0.85, [11]), show that the proposed model is highly predictive, when applied to 550 chemicals never seen during the model development.

7.8.Predictivity - Assessment of the external validation set:

The prediction set is very large: in fact it is really rare in QSAR modeling that an original data set of 643 chemicals is split in such a way: only 93 for training (to find the best modeling descriptors) and 550 for verify the predictivity on chemicals not used in model development. The splitting methodology, based on similarity analysis (explained in section 7.6), is a guarantee of the selection of a meaningful training set and a representative prediction set. Training and prediction set are balanced according to both structure and response. In particular, for response the range of logKoc values are [-0.31 - 6.02] and [0 - 6.33] respectively for training and prediction set. As much as concern structural representativity, the range of descriptors values are:

VP-0: training set (2.27 / 18.17), prediction set (1.16 / 22.53)

nAromBonb: training set (0 / 26), prediction set (0 / 29)

MAXDP: training set (0.008 / 5.66), prediction set (0 / 6.82)

nHBAcc: training set (0 / 10), prediction set (0 / 10)

The applicability domain of the model on the prediction set has been verified by the Williams plot: only 8 compounds on 550 of the prediction set are outliers for the response (not well predicted) and 11 are structural outliers (extrapolated). These results are a guarantee of the large applicability domain of the proposed model.

7.9.Comments on the external validation of the model:

No information available

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic basis for this physico-chemical property was set a priori, but a mechanistic interpretation of molecular descriptors was provided a posteriori (see 8.2).

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation:

The equation of the QSPR full model for the prediction of Koc is:

logKoc= 0.87 + 0.26 VP-0 - 0.23 nHBAcc + 0.08 nAromBond - 0.19 MAXDP

where VP-0: Valence path, order 0

nHBAcc: Number of hydrogen -bond acceptors

nAromBond: Number of aromatic bonds

MAXDP: Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule).

The nHBAcc descriptor, that is related to electronegative atoms of molecules, and MAXDP, related to molecule electrophilicity, represent different ways of taking into account the probability of bond formation between chemicals and groundwater: as expected, these descriptors are negative in sign as high affinity for water precludes soil sorption of the chemicals. The other two descriptors (VP-0, the most important in the QSAR equation, and nAromBond) are related to molecular size and have positive signs: VP-0 highlights that the bigger compounds are more sorbed than leached, while the less relevant descriptor nAromBond is a counter for aromatic bonds.

8.3.Other information about the mechanistic interpretation:

No other information available

9.Miscellaneous information

9.1.Comments:

Given the good results of the external validation, this model has a large applicability domain and therefore unsuccessful applications are probably very reduced. Anyhow, the check of outliers by the Williams plot and the Insubria graph for chemicals without experimental data (see section 5.1) will allow to verify the model applicability.

To predict logKoc for new chemicals without experimental data, it is suggested to apply the equation of the **Full Model**, developed on all the available chemicals (N Training=643).

The equation (reported also in section 4.2) and the statistical parameters of the full model are:

logKoc= 0.87 + 0.26 VP-0 - 0.23 nHBAcc + 0.08 nAromBond - 0.19 MAXDP

N Training set= 643; R^2= 0.79; Q^2LOO= 0.79; $Q^2LMO_{30\%}$= 0.79; CCC = 0.89; CCCcv = 0.88 ;RMSE= 0.543; RMSEcv = 0.547

9.2.Bibliography:

[1]Gramatica P., Giani E., Papa E., Statistical external validation and consensus modeling: A QSPR case study for Koc prediction, J. Mol. Graphics Modell., 2007, 25, 755-766. doi:10.1016/j.jmgm.2006.06.005

[2]Yap, C.W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J.Comput.Chem. 2011, 32, 1466-1474. doi: 10.1002/jcc.21707

[3]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J. Comput. Chem. (Software News and Updates), 2013, 34 (24), 2121-2132. doi: 10.1002/jcc.23361

[4]Gramatica P., et al. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, J. Comput. Chem. (Software News and Updates), 2014, 35 (13), 1036-1044. doi: 10.1002/jcc.23576

[5]Sabljic A., et al. QSAR modeling of soil sorption. improvements and systematics of log Koc vs. log Kow correlations, Chemosphere 1995, 31, 4489–4514. doi:10.1016/0045-6535(95)00327-5

[6]Tao S., et al. Estimation of organic carbon normalized sorption coefficient (KOC) for soils using

the fragment constant method, Environ. Sci. Technol. 1999, 33, 2719–2725. doi: 10.1021/es980833d

[7]Huuskonen J., Prediction of soil sorption coefficient of a diverse set of organic chemicals from molecular structure, J. Chem. Inf. Comput. Sci. 2003, 43, 1457–1462. doi: 10.1021/ci020342j

[8]HyperChem 7.03, 2007 http://www.hyper.com/

[9]OpenBabel 2.3.0, 2010 http://openbabel.org

[10]Chirico N. and 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. doi: 10.1021/ci200211n

[11]Chirico N. and Gramatica P., Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection, J. Chem. Inf. Model. 2012, 52, 2044–2058 doi: 10.1021/ci300084j

[12]Shi L.M. et al. QSAR Models Using a Large Diverse Set of Estrogens, J. Chem. Inf. Comput. Sci. 2001, 41, 186–195. doi: 10.1021/ci000066d

[13]Schuurman G. et al. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean, J. Chem. Inf. Model. 2008, 48, 2140-2145. doi: 10.1021/ci800253u

[14]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 2009, 49, 1669-1678 doi: 10.1021/ci900115y

9.3.Supporting information:

Training set(s)

Log Koc Training Set.sdf	file:///C:\Documents and Settings\lab-qsar\Desktop\QMRF to send 2015\Koc PaDEL\Log Koc Training Set.sdf

Test set(s)

Log Koc Prediction set.sdf	file:///C:\Documents and Settings\lab-qsar\Desktop\QMRF to send 2015\Koc PaDEL\Log Koc Prediction set.sdf

Supporting information

Log Koc full.sdf	file:///C:\Documents and Settings\lab-qsar\Desktop\QMRF to send 2015\Koc PaDEL\Log Koc full.sdf

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC