	QMRF identifier (JRC Inventory): **Q17-26-0057**
	QMRF Title: **QSARINS model 2 for log Koc**
	Printing Date: **Apr 16, 2018**

1.QSAR identifier

1.1.QSAR identifier (title):

QSARINS model 2 for log Koc

1.2.Other related models:

1.3.Software coding the model:

DRAGON

software for the calculation of molecular descriptors, ver. 5.3 for Windows, 2005

R. Todeschini, V. Consonni, A. Mauri, M. Pavan

http://www.talete.mi.it/

MOBY DIGS

Software for multilinear regression analysis and variable subset selection by Genetic Algorithm, ver 1 beta for Windows, 2005

Todeschini Roberto, Talete srl, Milan (Italy)

http://www.talete.mi.it/

QSARINS

Software for the development, analysis and validation of QSAR MLR models, version 2.2, 2015

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

http://www.qsar.it/

2.General information

2.1.Date of QMRF:

19/01/2011

2.2.QMRF author(s) and contact details:

[1]Paola Gramatica University of Insubria, Varese +390332421573 paola.gramatica@uninsubria.it http://www.qsar.it/

[2]Stefano Cassani University of Insubria, Varese +390332421439 stefano.cassani@uninsubria.it http://www.qsar.it/

[3]Ester Papa University of Insubria, Varese +390332421573 ester.papa@uninsubria.it http://www.qsar.it/

2.3.Date of QMRF update(s):

28/01/2015

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Paola Gramatica University of Insubria, Varese +390332421573 paola.gramatica@uninsubria.it http://www.qsar.it/

[2]Elisa Giani University of Insubria, Varese +390332421573 elisagiani@libero.it http://www.qsar.it/

[3]Ester Papa University of Insubria, Varese +390332421573 ester.papa@uninsubria.it

http://www.qsar.it/

2.6.Date of model development and/or publication:

2007

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

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

[2]HYPERCHEM (2002). Release 7.03 for Windows in: Molecular Modeling System, Hypercube, Inc., Gainesville, FL, USA. http://www.hyper.com/

[3]DRAGON (2005). Software for the calculation of molecular descriptors, Version 5.3 for Windows, R. Todeschini, V. Consonni, A. Mauri, M. Pavan. http://www.talete.mi.it/

[4]MOBY DIGS (2005). Software for multilinear regression analysis and variable subset selection by genetic algorithm, Version 1 for Windows, Talete srl, Milan, Italy. http://www.talete.mi.it/

2.8.Availability of information about the model:

Non-proprietary. Defined and available algorithm [ref 2; sect 9.2].

Training and prediction sets are available in the Supporting Information of the related paper [ref 2; sect 9.2], in the attached sdf files in this QMRF (see Section 9.3) and in the QSARINS database [ref 3,4; sect 9.2].

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

None to date.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

3.2.Endpoint:

2.Environmental fate parameters 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 [ref 5-7; sect 9.2] 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 develop published and good-quality QSPR models.

3.4.Endpoint units:

dimensionless

3.5.Dependent variable:

log Koc

3.6.Experimental protocol:

3.7.Endpoint data quality and variability:

Data were taken from various sources and collected into a single large dataset. As stated in section 3.3, we used data already curated and modelled by various authors [ref 5-7; sect 9.2]:. 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:

LogKoc 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)

GA-OLS

Split Model equation (N. Training set=93):

LogKoc = -2.19 (±0.30) + 2.10 (±0.14) VED1 - 0.34 (±0.04) nHAcc - 0.31 (±0.05) MAXDP - 0.33 (±0.12) CIC0

Full Model equation (N. Training set=643):

LogKoc = -1.92 (±0.11) + 2.07 (±0.06) VED1 - 0.31 (±0.01) nHAcc - 0.31 (±0.02) MAXDP - 0.39 (±0.05) CIC0

The modeling descriptors are: VED1 (eigenvector coefficient sum from distance matrix), nHAcc (number of acceptor atoms for H-bonds), MAXDP (maximal electropological positive variation), CIC0 (complementary information content index (neighbourhood symmetry of 0 order)). See section 4.3 for a more detailed description of the four modeling descriptors.

4.3.Descriptors in the model:

[1]VED1 dimensionless eigenvector coefficient sum from distance matrix [8], related to molecular size. Is the most iimportant descriptor in the equation, with a positive sign, highlighting that the bigger compounds are more sorbed than leached

[2]nHAcc dimensionless number of acceptor atoms for H-bonds, related to electronegative atoms of molecules. Represent a way of taking into account the probability of bond formation between chemicals and groundwater: this descriptor is negative in sign (inversely related to logKoc) as high affinity for water precludes soil sorption of the chemicals

[3]MAXDP dimensionless maximal electropological positive variation [1], it takes into account the electronic distribution in the topological graph and 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]CIC0 dimensionless CIC0 dimensionless complementary information content index

(neighbourhood symmetry of 0 order) [ref 9,10; sect 9.2], it is related to molecular size.

4.4.Descriptor selection:

A total of 1079 molecular descriptors of differing types (0D, 1D, 2D, 3D) were calculated. 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 0.97 was removed to reduce redundant information), and a final set of 479 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 models were obtained. Then the GA was applied in order to explore new combinations of variables, selecting the variables by a mechanism of reproduction/mutation. The optimized parameter used was the cross-validated correlation coefficient R2CVor Q2LOO(leave-one-out). The GA-VSS, by Ordinary Least Squares regression (OLS), included in MOBYDIGS (and now reproduced in QSARINS [ref 4,5; sect 9.2]), was applied to select only the best combination of descriptors from the input pool: 4 descriptors selected from 479.

4.5.Algorithm and descriptor generation:

Multiple Linear Regression OLS method was applied to generate the model using the molecular descriptors calculated by the DRAGON software.

Descriptors were generated with DRAGON 5.3 from HYPERCHEM 7.03 optimized structures (*.hin files). Any user can re-derive the model calculating molecular descriptors with the DRAGON software (also from SMILES strings) and applying the given equation.

4.6.Software name and version for descriptor generation:

DRAGON (2005, version 5.3 for Windows)

Chemical structures, drawn with HyperChem 7.03 and used as input file (*.hin) for DRAGON 5.3, energy minimised using MM+ procedure. These structures are available in QSARINS (QSARINS-Chem module[ref 4; sect 9.2] enabling an end user to regenerate the descriptors for a new compound

Prof. R.Todeschini - distributed by Talete srl, via Pisani 13, 20124 Milano, Italy

http://www.talete.mi.it/

4.7.Chemicals/Descriptors ratio:

Split Model: 23.25 (93 chemicals / 4 descriptors)

Full Model: 160.75 (643 chemicals / 4 descriptors)

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 boundaries were 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 three standard deviation units) and chemicals very structurally influential in

determining model 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. Response and descriptor space:
Range of experimental LogKoc values: -0.31 - 6.33.
Range of descriptors values: VED1: 1.414 - 5.646; nHAcc: 0 - 11; MAXDP: 0 - 6.199; CIC0: 0.4 - 4.89.

5.2.Method used to assess the applicability domain:

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

5.3.Software name and version for applicability domain assessment:

QSARINS
AD re-verified with QSARINS, software for the development, analysis and validation of QSAR MLR models, ver. 2.2, 2015
Paola Gramatica, email: paola.gramatica@uninsubria.it
http://www.qsar.it/

5.4.Limits of applicability:

SPLIT Model Domain

Outliers for structure, hat value > 0.16 (h*): chlordecone (258), metasulfron methyl (628), thiameturon methyl (637).
Outliers for the response,standardised residuals > 3 standard deviation units:methylurea (330), 2,3,5-trimethylphenol (358), benfluralin (408), 2,6-dichlorobenzamide (427), 2,6-dinitro-n-propyltrifluoro-p-toluidine (432), toxaphene (499), dinitramine (556), and oxyfluoren (591) in the training set; trifluralin (394) in the prediction set.

FULL Model Domain:

Outliers for structure, hat value > 0.023 (h*): metasulfron methyl (628), thiameturon methyl (637).
Outliers for the response, standardised residuals > 3 standard deviation units: 2,3,5-trimethylphenol (358), 2,6-dichlorobenzamide (427) and toxaphene (499).

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 include 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:

Split model (N Training = 93): R^2 = 0.82

s = 0.539

F = 98.99

RMSE = 0.523

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

Split model (N Training = 93): Q^2LOO= 0.80

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

Q^2LMO was not calculated, since we calculated Q^2BOOT (see 6.11).

6.10.Robustness - Statistics obtained by Y-scrambling:

$R^2_{Y\text{-}SC}$= 4.32

$Q^2_{Y\text{-}SC}$= 3.36.

The low values of Y-scrambled $R^2$and Q^2mean that the proposed model is not given by chance

6.11.Robustness - Statistics obtained by bootstrap:

Q^2_{BOOT}= 0.79.

The high value of Q^2BOOT means that the model is robust and stable.

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 external validation set ("prediction set") consists of 550 heterogeneous organic compounds with a range of logKoc values from 0 to 6.33. The training and prediction sets are structurally balanced, since the splitting was based on a 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 space [ref 11; sect 9.2].

7.7.Predictivity - Statistics obtained by external validation:

Q^2ext-F1[ref 12; sect 9.2] = 78.11

Q^2ext-F2[ref 13; sect 9.2] = 77.96

Q^2ext-F3[ref 14; sect 9.2] = 79.31

RMSE = 0.56

CCC [ref 15,16; sect 9.2]= 88.33

The high values of external Q2, calculated in different ways (see

references for more details), and CCC, show that the proposed model is predictive for new chemicals. In fact, the model shows good results when applied to the chemicals never seen during the model development (chemicals in the prediction set).

7.8.Predictivity - Assessment of the external validation set:

The validation set (prediction set) is large: in fact it is 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), allows for 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. In particular, regarding the structural representativity of training and prediction set, the range of descriptor values are as follows: VED1: training set [1.414 - 5.312], prediction set [1.414 - 5.646] nHAcc: training set [0 - 11], prediction set [0 - 11] MAXDP: training set [0.083 - 5.57], prediction set [0 - 6.199] CIC0: training set [0.4 - 4.810], prediction set [0.4 - 4.891] The applicability domain of the model on the prediction set was verified by the Williams plot: 8 compounds out of 550 of the prediction set are outliers for the response (not well predicted) and only 2 are structural outliers (extrapolated, even if, in this case, verified as good predictions). These results are indicative of the large applicability domain of the proposed model.

7.9.Comments on the external validation of the model:

In addition, to verify the external predictivity of the model, the experimental data set of 643 compounds was split into three different sets:

(a) a prediction set of 160 chemicals (25% of the total set), selected by activity sampling from the data set ordered by the response value, taking every fourth chemical from the set (splitting by ordered response);

(b) a training set of 307 chemicals on which to redevelop the model, selected by SOM (300 epochs, 10 x 10 map) (48% of the total set);

(c) a prediction set of 176 chemicals selected by SOM (27% of the total set).

Also redeveloping the model on the wider training set (b), the variables selected by GA in the best OLS model are VED1, nHAcc, MAXDP and CIC0. The model, validated with the two prediction sets (a, c), has good performances both in fitting and predicitvity (R^2=0.78, Q^2_{LOO}=0.78, $R^2_{pred\ (a)}$=0.77, $R^2_{pred\ (c)}$=0.82).

Good and externally predictive models are also obtained on the same variables even if scrambling is performed between the training and the different prediction sets.

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 of molecular descriptors (ordered by importance, according to the standardized coefficient values):

VED1: eigenvector coefficient sum from distance matrix, encoding the 2D molecular dimension (most important descriptor).

nHAcc: number of acceptor atoms for H-bonds.

MAXDP: the maximal electropological positive variation.

CIC0: a complementary information content index (neighbourhood symmetry of 0 order), related to the differences in the atomic distribution.

The nHAcc descriptor, which 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 descriptorsare negative in sign (inversely related to logKoc) as high affinity for water precludes soil sorption of the chemicals. The other two descriptors (VED1 and CIC0) are related to molecular size, but their relevance is very different: the more important VED1 has a positive sign, highlighting that the bigger compounds are more sorbed than leached, the less relevant descriptor CIC0 is probably useful only to improve model quality in order to adapt some particular chemicals.

8.3.Other information about the mechanistic interpretation:

No information available.

9.Miscellaneous information

9.1.Comments:

Given the 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.

As for all our models, 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=643).

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

LogKoc = -1.92 (±0.11) + 2.07 (±0.06) VED1 - 0.31 (±0.01) nHAcc - 0.31 (±0.02) MAXDP - 0.39 (±0.05) CIC0

N = 643; R^2 = 0.79; Q^2 = 0.79; Q^2_{BOOT}

= 0.79; s = 0.547; RMSE = 0.545; RMSE $_{LOO}$ = 0.550 All the modelling descriptors were verified, and reproducible, in

the free on-line version of the DRAGON software (http://www.vcclab.org/lab/edragon/).

In addition, a **Consensus model**, calculated by averaging the predicted values from the best 10 individual models (N $_{TR}$=93, N $_{P}$=505), was also developed and proposed in the paper [ref 2; sect 9.2]

Fitting (R^2 = 0.82) and predictive ability (verified by R^2ext = 0.80) for the Consensus model are better than for any individual model. Finally, our models were compared with the EPI Suite model for Koc (KOCWIN), characterized by lower values of R^2 (0.78) and higher RMSE for the prediction set chemicals (0.635). Comparing the residuals between the experimental and predicted logKoc values, our model showed better results if compared to KOCWIN, with lower mean and maximum residuals, and lower number of chemicals with residual > 1.5 (See Tables 2 and 4 the paper [ref 2; sect 9.2]).

9.2.Bibliography:

[1]Gramatica P, Corradi M & Consonni V (2000). Modelling and prediction of soil sorption coefficients of non- ionic organic pesticides by molecular descriptors. Chemosphere 41, 763–777. DOI:10.1016/S0045-6535(99)00463-4

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

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

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

[5]Sabljic A, Gusten H, Verhaar H & Hermens J (1995). QSAR modeling of soil sorption. Improvements and systematics of log Koc vs. log Kow correlations. Chemosphere 31, 4489–4514. DOI:10.1016/0045-6535(95)00327-5

[6]Tao S, Piao H, Dawson R, Lu X & Hu H (1999). Estimation of organic carbon normalized sorption coefficient (KOC) for soils using the fragment constant method. Environmental Science & Technology 33, 2719–2725. DOI:10.1021/es980833d

[7]Huuskonen J (2003). Prediction of soil sorption coefficient of a diverse set of organic chemicals from molecular structure. Journal of Chemical Information and Computer Sciences 43, 1457–1462. DOI:10.1021/ci020342j

[8]Balaban AT, Ciubotariu D & Medeleanu M (1991). Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. Journal of Chemical Information and Computer Sciences 31, 517–523. url not available

[9]Bonchev D (1983). Information Theoretic Indices for Characterization of Chemical Structures. , RSP/Wiley, Chichetser (UK). url not available

[10]Magnuson VR, Harriss DK & Basak SC (1983). Studies in Physical and Theoretical Chemistry. Elsevier, Amsterdam, The Netherlands, in King RB (Ed.), 178–191. url not available

[11]Gramatica P, Pilutti P & Papa E (2004). Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training Test Sets and Consensus Modeling. Journal of Chemical Information and Computer Sciences 44, 1794-1802. DOI:10.1021/ci049923u
[12]Shi LM, Fang H, Tomg W, Wu J, Perkins R, Blair RM, Branham WS, Dial SL, Moland CL & Sheenan DM (2001). QSAR Models Using a Large Diverse Set of Estrogens. Journal of Chemical Information and Computer Sciences 41, 186–195. DOI:10.1021/ci000066d
[13]Schuurman G, Ebert RU, Chen J, Wang B & Kuhne R (2008). External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. Journal of Chemical Information and Modeling 48, 2140-2145. DOI:10.1021/ci800253u
[14]Consonni V, Ballabio D & Todeschini R (2009). Comments on the Definition of the Q2 Parameter for QSAR Validation. Journal of Chemical Information and Modeling 49, 1669-1678. DOI:10.1021/ci900115y
[15]Chirico N & Gramatica P (2011). Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, Journal of Chemical Information and Modeling 51, 2320-2335. DOI:10.1021/ci200211n
[16]Chirico N & Gramatica P (2012). Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection, Journal of Chemical Information and Modeling 52, 2044–2058. DOI:10.1021/ci300084j

9.3.Supporting information:

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q17-26-0057

10.2.Publication date:

2017-09-27

10.3.Keywords:

DRAGON;Koc;octanol-carbon;sorption;QSARINS;INSUBRIA;

10.4.Comments:

old# Q47-19-49-477