Ecotoxicology and Environmental Safety 147 (2018) 292–298

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

journal homepage: www.elsevier.com/locate/ecoenv

Pre-selection and assessment of green organic solvents by clustering chemometric tools

Marek Tobiszewski[a], Miroslava Nedyalkova[b,*], Sergio Madurga[c], Francisco Pena-Pereira[d], Jacek Namieśnik[a], Vasil Simeonov[b]

[a] Department of Analytical Chemistry, Chemical Faculty, Gdańsk University of Technology (GUT), 11/12 G. Narutowicza St., 80-233 Gdańsk, Poland
[b] Faculty of Chemistry and Pharmacy, University of Sofia "St. Kl. Okhridski", 1164 Sofia, J. Bourchier Blvd. 1, Bulgaria
[c] Materials Science and Physical Chemistry Department & Research Institute of Theoretical and Computational Chemistry (IQTCUB) of Barcelona University (UB), C/ Martí i Franquès, 1, 08028 Barcelona, Catalonia, Spain
[d] Analytical and Food Chemistry Department, Faculty of Chemistry, University of Vigo, Campus As Lagoas-Marcosende s/n, 36310 Vigo, Spain

ARTICLE INFO

Keywords:
Green chemistry
Solvents
Chemometrics
Bioconcentration factor
Greenness assessment

ABSTRACT

The study presents the result of the application of chemometric tools for selection of physicochemical parameters of solvents for predicting missing variables – bioconcentration factors, water-octanol and octanol-air partitioning constants. EPI Suite software was successfully applied to predict missing values for solvents commonly considered as "green". Values for logBCF, $logK_{OW}$ and $logK_{OA}$ were modelled for 43 rather nonpolar solvents and 69 polar ones. Application of multivariate statistics was also proved to be useful in the assessment of the obtained modelling results. The presented approach can be one of the first steps and support tools in the assessment of chemicals in terms of their greenness.

1. Introduction

Green chemistry is the concept introduce by Anastas and Warner (1998) with the publication of the twelve principles that are specific guidance on the introduction of sustainability to chemical science. Since then this concept has developed much and efforts were made to develop zero-waste technologies, design benign products that maintain their properties, find renewable and bio-based feedstock for chemicals or apply energy- efficient technologies. Also, solvents gained a lot of attention, the fifth principle of green chemistry states that solvents should not be applied if possible (Jessop, 2016) otherwise they should be as inert to the environment as possible.

As the application of solvents cannot be avoided in many technological processes, it is highly desired to use green solvents. The green solvent is characterised by preferential environmental, health and safety (EHS) parameters (Capello et al., 2007). The first, very basic information about solvent greenness can be obtained from its physicochemical parameters and phase distribution constants. For example, solvents with low boiling points are very volatile; therefore the exposure by inhalation is very likely. High values of octanol – water partitioning coefficients give initial information indicating that the compound can be accumulated in the animal tissues. More specific information on solvents greenness can be obtained from toxicological and ecotoxicological data, such as oral (Sathish et al., 2016) or inhalation toxicities, toxicity towards aquatic organisms or carcinogenicity (Tobiszewski and Namieśnik, 2015). Similarly, environmental persistence data such as biodegradability or hydrolysis potential give information about environmental related hazards. Remarkably, the European REACH Regulation (EC No 1907/2006) ("Regulation (EC) No 1907/2006 – REACH – Safety and Health at Work – EU-OSHA", 2017) has set as a priority the assessment of chemicals' bioaccumulative potential, which is the potential of a substance to accumulate in biota and, eventually, to pass through the food chain. A parameter that is widely used to measure a chemical substance's bioaccumulative potential is the bioconcentration factor (BCF). BCF is commonly defined as the ratio between the concentration of a chemical substance present in an aquatic organism and in the surrounding environment at thermodynamic equilibrium under controlled laboratory conditions (Arnot and Gobas, 2006).

One of the problems related to the assessment of solvents in terms of their greenness is the non-availability of data that are required (Alder et al., 2016). The missing data can be approximated with the average value for a given class of chemicals. However, usually, the obtained assessment estimation is characterised by high uncertainty. Therefore, it is desired to find reliable methods for the predictions of missing data values. It is especially important in the case of solvents that are

* Corresponding author.
E-mail address: mici345@yahoo.com (M. Nedyalkova).

http://dx.doi.org/10.1016/j.ecoenv.2017.08.057
Received 26 February 2017; Received in revised form 20 August 2017; Accepted 21 August 2017
Available online 14 September 2017
0147-6513/

relatively novel and still poorly characterised, like esters or ethers derived from renewable feedstocks (Pena-Pereira et al., 2015).

There are several methodologies for modelling the bioconcentration values of various chemicals (Pavan et al., 2008). Quantitative structure-activity relationship (QSAR) applied to a dataset of chemicals allowed to predict logBCF with the model with an r^2 = 0.491 (Petoumenou et al., 2015). QSAR supported by partial least square modelling allowed to obtain fit the model with an r^2 = 0.868 (Qin et al., 2009). Also, an extensive investigation was made of the physico-chemical properties of 152 solvents in a search for quantitative structure–property relationships (QSPR) (Gramatica et al., 1999). The artificial neural network followed by relatively simple generic model also allowed to establish the values of logBCF with acceptable accuracy (Fatemi et al., 2003). Application of linear and nonlinear models allowed to model logBCFs of 107 pesticides successfully (Yuan et al., 2016).

Although different logBCF prediction methods have been mainly applied to organic pollutants, efforts to create models for predictions of these values for solvents, according to authors' knowledge were not made.

The aim of this study is to interpret two sets of solvents described by several physicochemical and biological features by multivariate statistics. Also, we present the assessments of physicochemical properties such as logarithms of partitioning constants between octanol and air ($logK_{OA}$), octanol and water ($logK_{OW}$) and logBCF by using the Estimation Programs Interface (EPI) Suite. Results of the EPI Suite predictions of the physical properties of the solvents sets and comparisons with the available literature values are validated. The above-mentioned software is a valuable tool to be applied to experimental data for such kind of solvents properties when there is lack of sufficient data. In this way, it is possible to find the relationships that are the basis for modelling of solvents parameters that define their greenness. Easy but reliable prediction of hazards related to greener solvents introduction is highly desired. The paper presents the application of multivariate statistics tools in the prediction of unknown properties and assessment of their modelling results.

2. Materials and methods

2.1. Dataset

The dataset consists of 43 solvents – *non-polar and sparingly volatile solvents* and 69 solvents *a* priori categorised as *polar* solvents. In this study, the datasets are treated separately. This *a priori* classification is based on previous research (Tobiszewski et al., 2015), where large dataset of 151 solvents was analysed with cluster analysis using Melting and boiling point, density, water solubility, vapour pressure, Henry's law constant, $LogK_{ow}$, $LogK_{oa}$ and surface tension as physicochemical parameters for solvent classification. The identified third cluster – (conditionally determined as "*non-polar and volatile*" solvents) in Fatemi et al. (2003), is not investigated in the present study, as solvents grouped there (mainly chlorinated solvents or other causing EHS problems) are out of interest in the light of green chemistry. Not every process can be carried out with water miscible solvents. In fact, immiscible solvents are required in a wide number of processes (e.g., extraction and separation processes, synthesis, etc.), so their greenness evaluation is also of high relevance.

Each solvent in the present study was characterised by 10 parameters, namely melting point, boiling point, density, surface tension, water solubility, vapour pressure and Henry's law constant, as well as logarithms of partitioning constants between octanol and air ($logK_{OA}$) and octanol and water ($logK_{OW}$) and logarithm of bioconcentration factor (logBCF). The physicochemical parameters were extracted from material safety data sheets of chemicals and from Handbook On Physical-Chemical Properties And Environmental Fate For Organic Chemicals (Mackay et al., 2006). Three additional parameters were included for the chemometric analysis, namely, $logK_{OA}$calc., $logK_{OWcalc}$., and $logBCF_{calc}$. These parameters were calculated with models described in Section 2.3.

2.2. Multivariate statistics

Hierarchical clustering (HC) is a well-documented approach to the unsupervised pattern recognition (Massart et al., 1983; Massart and Kaufman, 1998). It aims to select groups of similar objects (clusters) within different data sets and to interpret the meaning of the clustering either between the objects of interest or between the parameters used for the description of the objects. Usually, the hierarchical cluster analysis requires several steps in performing the algorithm of clustering: standardisation of the raw data (in order to avoid the effect of the different dimensionality of parameters); determination of the distance between the objects for clustering (in order to introduce a similarity measure); procedure for linkage. The results are normally presented on a tree-like plot called dendrogram and in the final stage, a criterion for determination of the cluster significance is needed in order to improve the interpretation. The use of **c**hemometrics for the treatment of different data sets provides a valuable tool for objective decision-making (Hristov et al., 2016; Nedyalkova et al., 2017).

Principal component analysis (PCA) is one of the several multivariate methods that allows us to explore patterns in complex data sets allowing to classify the information and detect structure in a diffuse data set. In general, PCA is a mathematical treatment of the input data matrix (objects described by many features or variables) where the goal is to represent the variation present in many variables by a small number of factors or latent variables. A new space of the features is formed which it makes possible to visualise and project the multivariate nature of the data set.

The central task in PCA is to reduce the original dimension of the input matrix **X** to two parts – factor loadings (part **A** matrix) and factor scores (part F matrix). The first one includes the weights of each feature (variable) in each identified factor (new latent variable). The higher the weights the higher is the contribution of the original variable. Thus, this procedure allows us to identify which variables influence the objects.

If the objects have to be presented in the space of the new latent variables, then the factor scores matrix must be used. The specific rules for performing and interpreting PCA are presented, for instance (Einax et al., 1997).

2.3. Modelling – EPI Suite™

In the current work the following subprograms of EPI Suite™ version 4.10 were used: *KOAWIN*™, KOWWIN™ and BCFBAF™. EPI Suite™ is available from the US Environmental Protection Agency – (US EPA et al., n.d.) (Computer Program Estimation Programs Interface Suite™ for Microsoft® Windows Version 4.10, available on http://www.epa.gov/oppt/exposure/pubs/episuite.htm)

This KOAWIN™ program estimates the logarithm of the octanol-air partition coefficient (K_{OA}) of an organic compound using the compound's octanol-water partition coefficient (Kow) and Henry's Law constant (HLC). KOAWIN requires only a chemical structure to estimate K_{OA}. Structures are entered into KOAWIN through SMILES (Simplified Molecular Input Line Entry System) notations, which are also used by other estimation programs in EPA's EPI Suite. It is possible to estimate K_{OA} from the octanol-water partition coefficient (K_{OW}) and Henry's law constant (H) by the following equation: $K_{OA} = K_{OW}(RT)/H$, where R is the ideal gas constant and T is the absolute temperature. K_{OA} and K_{OW} are unitless values. H/RT is the unitless Henry's law constant, also known as the air-water partition coefficient (K_{AW}) (Meylan and Howard, 1995). Therefore, the equation to estimate K_{OA} is: $\mathbf{K_{OA} = K_{OW}/K_{AW}}$.

The KOWWIN™ program predicts the logarithm of the octanol-water partition coefficient. KOWWIN uses a "fragment constant" methodology to predict logP. In a "fragment constant" method, a structure is divided

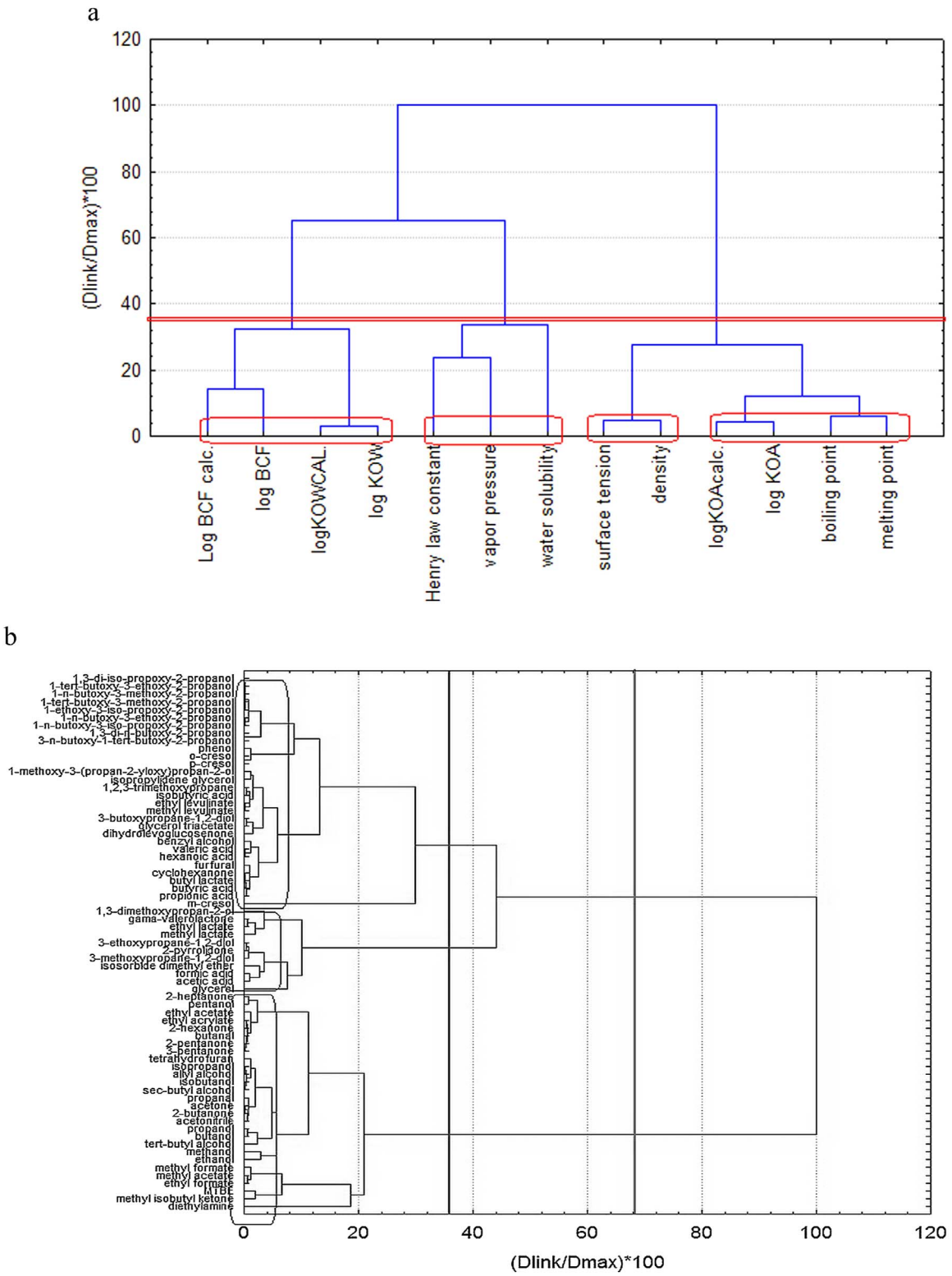

Fig. 1. a Hierarchical dendrogram for clustering of the variables for polar solvents (including those theoretically calculated). b Hierarchical dendrogram showing grouping of 69 polar solvents.

into fragments (atom or larger functional groups) and coefficient values of each fragment or group are summed together to yield the logP estimate. KOWWIN's methodology is known as an Atom/Fragment Contribution (AFC) method. Coefficients for individual fragments and groups were derived by multiple regression of 2447 reliably measured logP values. KOWWIN's "reductionist" fragment constant methodology (i.e. derivation via multiple regressions) differs from the "constructionist" fragment constant methodology of Hansch and Leo (1979).

The original estimation methodology used by the original BCFWIN program is described in Meylan and Howard (1995). The logBCF was regressed against log(K_{ow}), and chemicals with significant deviations from the line of best fit were analysed according to chemical structure. The BCFBAF method classifies a compound as either ionic or non-ionic.

The ionic substances were further divided into carboxylic acids, sulfonic acids and their salts, and quaternary N compounds. LogBCF for nonionic is estimated from log(K_{ow}) and a series of correction factors specific to each chemical (Meylan et al., 1996).

3. Results and discussion

3.1. Polar solvents

The first step was the calculation of the values for logBCF, logK_{OW} and logK_{OA}, as described in the previous section. To reveal the internal patterns existing in the group of polar solvents HC and PCA were applied. These techniques were used for clustering of the chemical variables and of the solvents themselves.

From Fig. 1a it can be easily read that three clusters have been formed – K1, consisting of log$BCF_{calc.}$, logBCF, log$K_{OWcalc.}$ and logK_{OW}; K2, formed by Henry law constant, vapour pressure, water solubility parameters and K3 that includes log$K_{OAcalc.}$, logK_{OA}, surface tension, density, boiling point and melting point. Grouping of calculated and their measured correspondents in one cluster indicates that the predictions could be accurate. The other important information is from which physicochemical factors, parameters of interest can be modelled. For example, Henry low constant, vapour pressure and water solubility are highly correlated.

PCA allowed to obtained similar clustering results to those obtained with HC, what can be read from the Table 1. The three latent factors explain over 70% of the total variance of the initial dataset. In the PCA interpretation, one could find that the variable logBCF (both experimentally found and calculated) is forming a separate latent factor not directly correlated with other variables. It indicates the specific importance of logBCF as a discriminant for the dataset. In fact, it is also indicated by HC where K1 could be conditionally subdivided into "logBCF" subcluster and logK_{ow}" subcluster.

Fig. 1b shows the grouping of polar solvents with HC. The objects (polar solvents) are clustered into three major groups. The mean values of physicochemical parameters for each group are presented in Table 2. The first group consists of alcohols with ether functional groups, aromatic alcohols and short-chain organic acids (apart from formic and acetic). Solvents in this group are less volatile, are characterised by slight water solubility and the highest values (but still low) of logK_{OW} and logBCF. Solvents present in this group are mainly novel, bio-based solvents. In the second group lactate esters, formic and acetic acids, glycerol and some alcohols with other functional groups are contained. These solvents are characterised by low volatility and very high water solubility. The third group consists mainly of "traditional" polar solvents, like short chain alcohols, ketones, aldehydes and esters. Its main discriminator is high volatility of solvents, reflected by low boiling points, high vapour pressures and Henry's law constants. These solvents rather do not undergo bioconcentration because of the low values of logBCF. The differences between clusters in terms of logK_{OW}, logK_{AO} and logBCF are not significant and are all low. This is an indication that solvents defined as polar ones do not undergo bioaccumulation, what is one of the parameters that define their greenness.

Table 1
Factor loadings for variables of the *polar* solvents dataset.

Variable	PC1	PC2	PC3
melting point	**0.848**	0.0508	0.220
boiling point	**0.902**	− 0.037	0.2110
density	**0.592**	− 0.573	0.374
water solubility	− 0.081	− **0.805**	− 0.128
vapour pressure	− **0.477**	− 0.020	− 0.022
Henry law constant	− **0.644**	0.169	0.274
logK_{OW}	0.097	**0.935**	0.178
logK_{OA}	**0.879**	0.205	0.022
surface tension	**0.750**	− 0.384	0.376
logBCF	0.099	0.165	**0.873**
logK_{OW}calc	0.005	**0.888**	0.257
logK_{OA}calc	**0.832**	0.316	0.006
LogBCF_{calc}	0.066	0.235	**0.727**
Expl. Var [%]	35.3	23.4	14.3

3.2. Non-polar solvents

Similarly, as in the case of polar solvents, the calculation of the values for logBCF, logK_{OW} and logK_{OW} was performed with Estimation Programs Interface (EPI). Then the clustering of variables and objects were performed. Here, we present the estimations of physicochemical properties such as octanol-air partition coefficients (logK_{OA}), octanol-water partition coefficient (K_{OW}), bioconcentration factor (BCF), using the Estimation Programs Interface (EPI) Suite. Predictions at room temperature were carried out for the all listed non-polar solvents. The EPI Suite requires only the chemical structure or the Chemical Abstracts Service (CAS) number to estimate the inquire properties. The BCF is estimated by the program by retrieving the BCF data in a file that contains information on measured BCF and other key experimental details. The logBCF was regressed against logK_{ow} and chemicals with significant deviations from the line of best fit were analysed according to chemical structure. Results of the EPI Suite predictions of the physical properties of the above non-polar solvents and comparisons with the available literature values are presented. It was interesting to compare the correlation between experimentally obtained and theoretically calculated indicators.

The clustering of the variables for non-polar solvents (Fig. 2a) shows a similar pattern as in the case of polar solvents (Fig. 1a). In Fig. 2a the

Table 2
Average values for variables of polar and nonpolar solvents grouped to respective clusters. The standard deviation (SD) is presented in parentheses.

Properties	POLAR SOLVENTS			NON POLAR SOLVENTS		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Melting point [°C]	− 2.7 (26)	− 14.4 (36)	− 81.1 (31)	− 66 (45)	5 (47)	− 15 (35)
Boiling point [°C]	196.18 (28.5)	175 (69)	85 (29)	127 (56)	209 (56)	197 (58)
Density [g cm^{-3}]	0.99 (0.08)	1.12 (0.08)	0.82(0.05)	0.75(0.08)	0.80(0.04)	0.9 (0.1)
Water solubility [mg dm^{-3}] at 25 °C	$1.9 \cdot 10^5$ ($1.3 \cdot 10^5$)	$7.4 \cdot 10^5$ ($3.3 \cdot 10^5$)	$2.3 \cdot 10^5$ ($2.8 \cdot 10^5$)	25.4 (50)	$2.6 \cdot 102$ ($9 \cdot 10^2$)	$1.5 \cdot 103$ ($2 \cdot 10^3$)
Vapour pressure [Pa] at 25 °C	101 (203)	$8.4 \cdot 10^3$ ($1.7 \cdot 10^3$)	$2.2 \cdot 10^4$ ($4.8 \cdot 10^4$)	$1.2 \cdot 10^4$ ($2.4 \cdot 10^4$)	384.60 (1328.0)	43.75 (70)
Henry law constant [Pa m^3 mol^{-1}]	0.42 (1.6)	2.5 (4.6)	10 (14)	$1.7 \cdot 106$ ($2.7 \cdot 10^5$)	$4.3 \cdot 105$ ($8 \cdot 10^4$)	52 ($1.02 \cdot 10^2$)
logK_{OW}	0.79 (0.80)	− 0.77 (0.60)	0.54 (0.60)	4.5 (1.1)	8 (1.9)	3.5 (1.5)
logK_{OA}	6.4 (1.2)	4.3 (1.7)	3.2 (0.7)	3.14 (0.82)	7.1 (1.9)	6.2 (2.2)
Surface tension [dyne cm^1]	33.4 (4.10)	3.1 (10)	23 (2.4)	24 (3)	29.4 (2.3)	31.5 (5.1)
logBCF	0.65 (0.57)	0.38 (0.20)	0.25 (0.70)	2.2 (0.8)	2.6 (0.7)	1.6 (0.7)
LogK_{OWcalc}	0.74 (0.89)	− 0.64 (0.7)	0.60 (0.5)	4.2 (0.9)	7.6 (1.7)	3.2 (1.2)
LogK_{OAcalc}	6.3 (1.8)	4.45(1.9)	3.4 (0.7)	3.0 (0.8)	7.0 (1.9)	6.3 (2.0)
LogBCF_{calc}	0.62 (0.45)	0.50 (0.5)	0.52 (0.5)	2.3 (0.4)	3.1 (0.7)	1.6 (0.6)

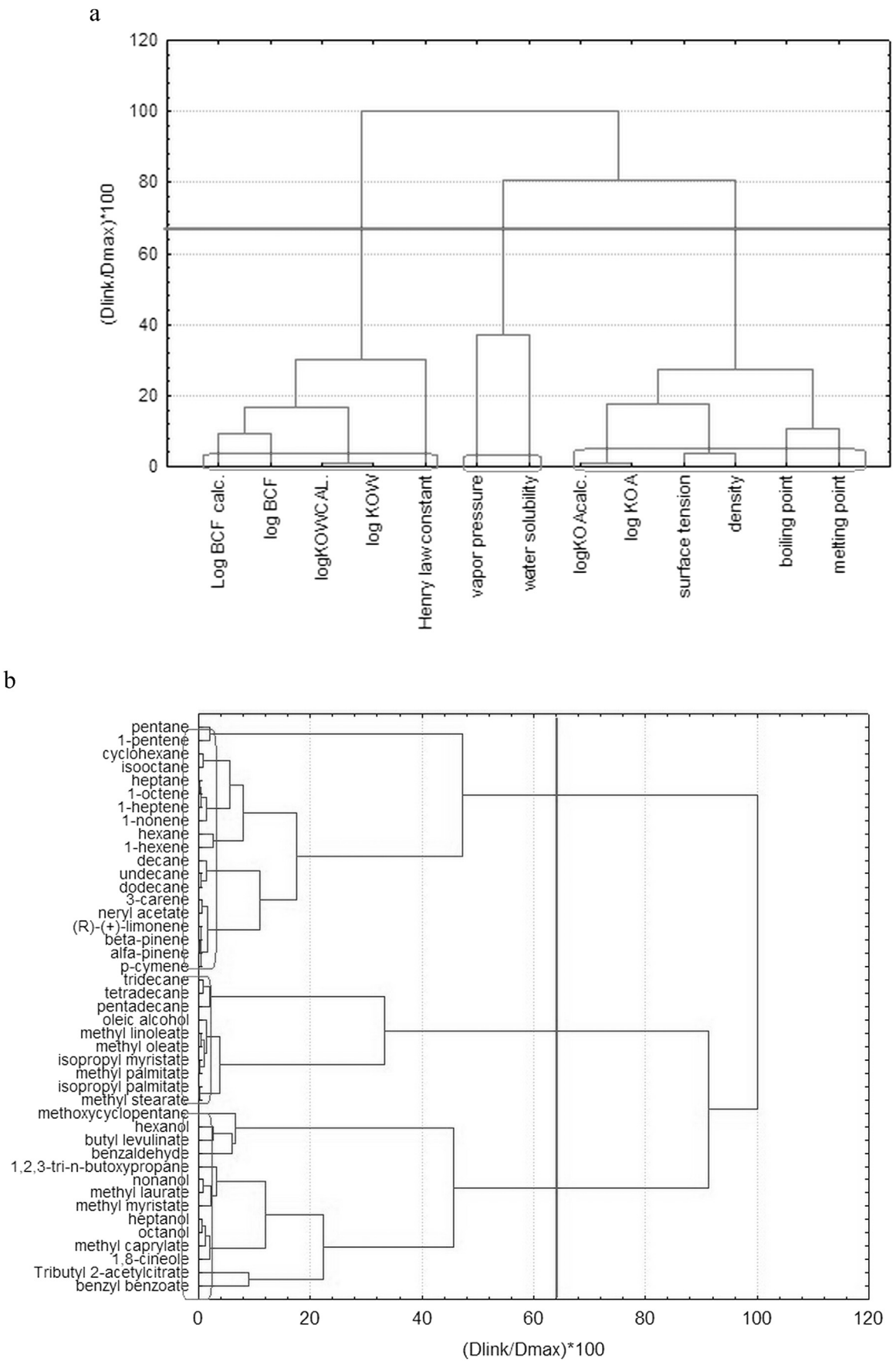

Fig. 2. a Hierarchical dendrogram for clustering of variables for non-polar solvents. b Hierarchical dendrogram presenting clusters of 43 non-polar solvents.

hierarchical dendrogram of clustering of variables is shown (z-transformed input data, squared Euclidean distances as a similarity measure, Ward's method of linkage and Sneath's criterion for cluster significance). The clustering of the theoretically calculated and experimentally existing values for logBCF, $logK_{OW}$ and $logK_{OA}$ match very well (they are joint together in the clusters) and it leads to the practically important conclusion that the calculating approach used could be used when there are missing data in the data set for the indicators in consideration. However, to know the obtained precision of each individual indicator requires to analyze individually. The obtained clusters confirm the relationship between parameters like logBCF and $logK_{OW}$ with the Henry law constant. The variable $logK_{OA}$ is correlated with a whole group of physicochemical parameters like surface tension, density, boiling and melting point.

In Table 3 the factor loadings values (Varimax rotation mode of PCA) are presented. The clustering of variables results is generally confirmed by PCA results.

Three latent factors explain over 78% of the total variance. The first latent factor PC1 (contribution of 33.7% of the total variance) indicates the strong correlation between a big group of physicochemical indicators (melting point, boiling point, density, surface tension) with the experimental and calculated values of $logK_{OA}$. Thus, it coincides entirely with cluster K3 and could be conditionally named “physicochemical factor”.

PC2 also explains a significant part of the total variance (30.2%) and resembles cluster K1 showing a strong relationship between the theoretical and experimental values of $logK_{OW}$ and logBCF. It is readily seen that the water solubility is negatively correlated to the above-mentioned parameters and this is a difference to the clustering in K1. But this relationship does not seem unusual and this latent factor could be conditionally named “solubility or polarity factor”.

In PC3 (contribution of 14.8% of the total variance) one finds a negative correlation between vapour pressure indicator and Henry law constant. Having in mind the big differences in the values of Henry law constant and vapour pressure for the non-polar solvents found in the literature there is no surprise for such a connection. In the hierarchical dendrogram (Fig. 1b) the vapour pressure and water solubility are linked together for the level of significance 66.67% of D_{max} but at level 33.33% of $D_{max,}$ such a linkage does not exist. The Henry law constant appears to be linked to logBCF and $logK_{OW}$ but at quite a high level of linkage.

The more significant aspect of the cluster analysis was to reveal relationships between the different non-polar solvents and possible markers making the difference within the seemingly homogeneous factor of non-polarity.

In Fig. 2b the hierarchical dendrogram of clustering of 43 nonpolar solvents is shown. Three major clusters are very clearly indicated with the level of significance 66.67% of D_{max}. The first one contains 19 out of 43 solvents, the second one – 10 out of 43 and the third one – the rest of 14 solvents. It is obvious that the formation of three different patterns of non-polar solvents requires identification of specific markers for each one of the groups of similarity. In Table 2 the average values for each one of the 13 variables used for solvents clustering and for each one of the clusters found are presented.

The clustering is based on the specific discriminators being present in the initial dataset. In K1 are included nonpolar solvents (like pentane, cyclohexane, heptane, decane, etc.) with the lowest melting point, lowest boiling point, lowest density, highest vapour pressure, lowest surface tension, and lowest $logK_{OA}$ (both experimental and theoretically calculated). This group is formed by volatile and rather nonpolar solvents. The second cluster K2 consists of solvents having on average lowest water solubility and vapour pressure, the highest Henry law constant and $logK_{OW}$, logBCF and $logK_{OA}$. Cluster K2 consists of a group of non-polar solvents, which are not water soluble. The third cluster K3 is characterised by highest density, water solubility solvents. All 43 solvents defined as non-polar ones are characterised by the much higher potential for bioaccumulation than solvents defined as polar ones. However, from a practical point of view, it is important to develop and assess green less polar solvents, as many processes require solvent that is not miscible with water.

The defined as a non-polar solvent group could be divided into three subcategories like volatile, water nonsoluble and slightly water soluble solvents. Grouping can be helpful in the studies searching theoretical relationship between solvent chemical structure and bioconcentration. The close resemblance between experimentally found and theoretically calculated parameters makes it possible to use the approach of filling missing data in data set comprising physicochemical and bioconcentration variables. EPI Suite predicts physicochemical properties and is a relatively convenient means of studying organic materials. When experimental data are not available to assess environmental risk, a possible way to estimate the necessary values is the use of estimation models. The EPI Suite was developed to help environmental scientists to prepare profiles for a wide array of chemical profiles. The fact that the program simply requires the chemical structure or Chemical Abstract

Table 3
Factor loadings for PCA analysis of *non-polar* solvents set.

Variable	PC 1	PC 2	PC 3
Melting point	**0.571**	0.228	0.561
Boiling point	**0.680**	0.256	0.640
Density	**0.860**	− 0.263	0.074
Water solubility	0.153	− **0.813**	0.160
Vapour pressure	− 0.490	− 0.060	− **0.621**
Henry law constant	− 0.348	0.287	**0.704**
$LogK_{OW}$	0.230	**0.876**	0.245
$LogK_{OA}$	**0.918**	0.243	0.012
Surface tension	**0.900**	− 0.130	0.179
LogBCF	0.022	**0.713**	0.336
$LogK_{OWcalc.}$	0.219	**0.890**	0.229
$LogK_{OAcalc.}$	**0.929**	0.211	0.026
$LogBCF_{calc.}$	− 0.049	**0.899**	0.174
Expl. Var%	33.7	30.2	14.8

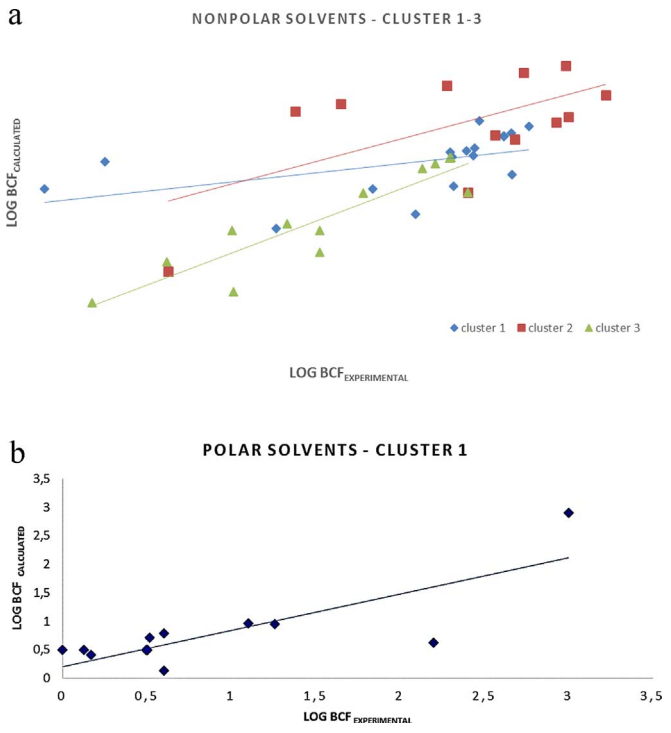

Fig. 3. a – The plot of experimental vs. predicted logBCF values for clusters 1–3 – non-polar solvents group and b – plot of experimental vs. predicted logBCF values for cluster 1 – polar solvents group.

Table 4
Summarised data for logBCF with correlation coefficients (predicted vs. experimental logBCF values) for both groups of solvents.

Solvent	*Cluster*	R^2 *(the square correlation coefficient between predicted and experimental values)*
Nonpolar	*1*	*0.25*
	2	*0.36*
	3	*0.83*
Polar	*1*	*0.67*
	2	*0.00*
	3	*0.007*

Service (CAS) number to generate all the predicted and experimental values has simplified its use.

In Fig. 3a an effort is made to indicate the internal relationship between clusters of non-polar solvents 1–3, based on the correlation coefficients found (predicted vs. experimental values for logBCF). In Table 4, the gradual increase of the correlation coefficient is observed with the highest value for cluster 3. This cluster includes the solvents whose logBCF values were predicted also by EPI suit. For the second group of solvents (polar solvents), only cluster 1 shows a reasonable correlation coefficient (predicted vs. experimental logBCF values). Probably, the correlation coefficient could be used as another discriminant factor for the solvents studied: when the variability between experimental and predicted values of logBCF is significant the correlation coefficients are with higher values (e.g. non-polar solvents). The lesser variability in logBCF leads to a lower correlation which is the case with polar solvents.

4. Conclusions

The simple classification with well-known chemometric tools is an important preliminary step in the selection of an optimal set of parameters for proper theoretical predictions. Here, cluster analysis and PCA were used to group solvents according to their similarity. Variables were grouped with principal component analysis and cluster analysis to assess and identify from which properties missing values can be predicted. The results show that values of logBCF for organic solvents can be modelled with EPI Suite software. Thus, these estimations will allow identifying novel green solvents for which experimental logBCF values are not yet available.

Acknowledgements

The support of an H2020 program of the European Union (H2020-TWINN-2015 – project “Materials Networking”, Project ID: 692146) is gratefully acknowledged by M. Nedyalkova, V.Simeonov and S. Madurga.

S. Madurga acknowledged for the financial support from Generalitat de Catalunya (Grants 2014SGR1017 and XrQTC).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ecoenv.2017.08.057.

References

Alder, C.M., Hayler, J.D., Henderson, R.K., Redman, A.M., Shukla, L., Shuster, L.E., Sneddon, H.F., 2016. Updating and further expanding GSK's solvent sustainability guide. Green Chem. 18, 3879–3890.

Anastas, P.T., Warner, J.C., 1998. Green Chemistry: Theory and Practice. Oxford University Press, New York. ⟨http://greenchemistry.yale.edu/green-chemistry-green-engineering-defined⟩.

Arnot, J.A., Gobas, F., 2006. A review of Bioconcentration Factor (BCF) and Bioaccumulation Factor (BAF) assessments for organic chemicals in aquatic organisms. Environ. Rev. 14, 257–297.

Capello, C., Fischer, U., Hungerbühler, K., 2007. What is a green solvent? A comprehensive framework for the environmental assessment of solvents. Green Chem. 9, 927–934.

Einax, J.J., Heinz, W.Z., Geiss, S., 1997. Chemometrics in Environmental Analysis. Wiley-VCH Verlag Gmb.

Gramatica, P., Navas, N., Todeschini, R., 1999. Classification of organic solvents and modelling of their physico-chemical properties by chemometric methods using different sets of molecular descriptors. TRAC 18, 461–471.

Fatemi, M.H., Jalali-Heravi, M., Konuze, E., 2003. Prediction of bioconcentration factor using genetic algorithm and artificial neural network. Anal. Chim. Acta 486, 101–108.

Hansch, C., Leo, A., 1979. Substituent Constants for Correlation Analysis in Chemistry and Biology. Wiley, New York.

Hristov, H., Nedyalkova, M., Madurga, S., Simeonov, V., 2016. Boron oxide glasses and nanocomposites: synthetic, structural and statistical approach. J. Mater. Sci. Technol. http://dx.doi.org/10.1016/j.jmst.2016.07.016.

Jessop, P.G., 2016. The use of auxiliary substances (e.g. solvents, separation agents) should be made unnecessary wherever possible and innocuous when used. Green Chem. 18, 2577–2578.

Mackay, D., Shiu, W.-Y., Ma, K.-C., Lee, S.C., 2006. Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals, second ed. CRC/Taylor & Francis, New York.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., 1998. Book Series: Data Handling in Science and Technology. Elsevier, Amsterdam.

Massart, Desiré L., Kaufman, L., 1983. The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis. Wiley, New York.

Meylan, W.M., Howard, P.H., 1995. Atom/fragment contribution method for estimating octanol-water partition coefficients. J. Pharm. Sci. 84, 83–92.

Meylan, W.M., Howard, P.H., Boethling, R.S., 1996. Improved method for estimating water solubility from octanol/water partition coefficient. Environ. Toxicol. Chem. 15, 100–106.

Nedyalkova, M.A., Donkova, B.V., Simeonov, V.D., 2017. Chemometrics expertise in the links between ecotoxicity and physicochemical features of silver nanoparticles: environmental aspects. J. AOAC Int. http://dx.doi.org/10.5740/jaoacint.16-0413.

Pavan, M., Netzeva, T.I., Worth, A.P., 2008. Review of literature-based quantitative structure – activity relationship models for bioconcentration. QSAR Comb. Sci. 27, 21–31.

Pena-Pereira, F., Kloskowski, A., Namieśnik, J., 2015. Perspectives on the replacement of harmful organic solvents in analytical methodologies: a framework toward the implementation of a generation of eco-friendly alternatives. Green Chem. 17, 3687–3705.

Petoumenou, M.I., Pizzo, F., Cester, J., Fernández, Alberto, Benfenati, Emilio, 2015. Comparison between Bioconcentration Factor (BCF) data provided by industry to the European Chemicals Agency (ECHA) and data derived from QSAR models. Environ. Res. 142, 529–534.

Qin, H., Chen, J., Wang, Y., Wang, B., Li, X., Li, F., Wang, Y., 2009. Development and assessment of quantitative structure-activity relationship models for bioconcentration factors of organic pollutants. Chin. Sci. Bull. 54, 628–634.

Regulation (EC) No 1907/2006 – REACH – Safety and Health at Work – EU-OSHA, 2017. ⟨https://osha.europa.eu/es/legislation/directives/regulation-ec-no-1907-2006-of-the-european-parliament-and-of-the-council⟩. (Accessed February 9).

Sathish, M., Silambarasan, S., Madhan, B., Raghava Rao, J., 2016. Exploration of GSK'S solvent selection guide in leather industry: a CSIR-CLRI tool for sustainable leather manufacturing. Green Chem. 18, 5806–5813.

Tobiszewski, M., Namieśnik, J., 2015. Scoring of solvents used in analytical laboratories by their toxicological and exposure hazards. Ecotoxicol. Environ. Saf. 120, 169–173.

Tobiszewski, M., Tsakovski, S., Simeonov, V., Namieśnik, J., Pena-Pereira, F., 2015. A solvent selection guide based on chemometrics and multicriteria decision analysis. Green Chem. 174773–174785.

US EPA, OCSPP, OPPT, n.d. EPI SuiteTM-Estimation Program Interface.

Yuan, J., Xie, Ch, Zhang, T., Sun, J., Yuan, X., Yu, S., Zhang, Y., 2016. Linear and nonlinear models for predicting fish bioconcentration factors for pesticides. Chemosphere 156, 334–340.