molecules

MDPI

Article

Expectation-Maximization Model for Substitution of Missing Values Characterizing Greenness of Organic Solvents

Gabriela Łuczyńska [1,2], Francisco Pena-Pereira [3], Marek Tobiszewski [4] and Jacek Namieśnik [4,*]

1 Division of Applied Mathematics and Probability, Institute of Mathematics, Faculty of Mathematics, University of Gdansk, 8 J. Bażyńskiego St., 80-309 Gdańsk, Poland; gabluczy@student.pg.edu.pl
2 Department of Nonlinear Analysis and Statistics, Faculty of Applied Mathematics, Gdańsk University of Technology (GUT), 11/12 G. Narutowicza St., 80-233 Gdańsk, Poland
3 Department of Analytical and Food Chemistry, Faculty of Chemistry, University of Vigo, Campus As Lagoas-Marcosende s/n, 36310 Vigo, Spain; fjpena@uvigo.es
4 Department of Analytical Chemistry, Chemical Faculty, Gdańsk University of Technology (GUT), 11/12 G. Narutowicza St., 80-233 Gdańsk, Poland; marektobiszewski@wp.pl
* Correspondence: jacek.namiesnik@pg.edu.pl or chemanal@pg.edu.pl

Received: 13 April 2018; Accepted: 25 May 2018; Published: 28 May 2018

Abstract: Organic solvents are ubiquitous in chemical laboratories and the Green Chemistry trend forces their detailed assessments in terms of greenness. Unfortunately, some of them are not fully characterized, especially in terms of toxicological endpoints that are time consuming and expensive to be determined. Missing values in the datasets are serious obstacles, as they prevent the full greenness characterization of chemicals. A featured method to deal with this problem is the application of Expectation-Maximization algorithm. In this study, the dataset consists of 155 solvents that are characterized by 13 variables is treated with Expectation-Maximization algorithm to predict missing data for toxicological endpoints, bioavailability, and biodegradability data. The approach may be particularly useful for substitution of missing values of environmental, health, and safety parameters of new solvents. The presented approach has high potential to deal with missing values, while assessing environmental, health, and safety parameters of other chemicals.

Keywords: E-M algorithm; green analytical chemistry; missing data prediction; solvents; sustainability assessment

1. Introduction

Green Chemistry has been defined as the "design of chemical products and processes to reduce or eliminate the use and generation of hazardous substances" [1]. With this aim, Anastas and Warner, introduced, in 1998, the twelve principles of Green Chemistry that charted a path towards sustainability in chemical processes [1]. Several principles of Green Chemistry point out the need to eliminate or replace solvents by less harmful alternatives [2]. Particularly, the 5th principle of Green Chemistry specifically recommends the use of innocuous solvents when avoiding the use of solvents is not possible. The employment of harmful solvents is also indirectly discouraged, as can be deduced from additional principles, such as waste prevention (1st principle) and the prevention or minimization of potential chemical accidents (12th principle) that are associated to their use. Furthermore, advances toward the design of safer chemicals (4th principle) that at the end of their function can be transformed into innocuous non-persistent products (10th principle), and importantly, renewable feedstocks, rather than depleting non-renewable resources (7th principle), are highly recommended [3,4].

Organic solvents are increasingly used in scientific and technological activities, with an estimated worldwide consumption of roughly 30 million metric tons per year [5]. Apart from the steady increase of solvent consumption in the last years, especially worrisome is the fact that certain solvents of very high concern are still being widely used. Thus, the implementation of solventless process is strongly advisable from the point of view of Green Chemistry. While remarkable efforts have been made in certain areas toward the implementation of solventless processes (e.g., solventless sample preparation approaches [6] and greener reactions under solvent free conditions [7]), these strategies are, in general terms, still far from reaching the desired level of implementation. Alternatively, the minimization of solvent consumption and/or the replacement of hazardous solvents by cleaner alternatives are highly recommended strategies to reduce the risks that are associated to solvent usage [8,9]. In this vein, a number of solvent selection guides have been reported in the literature for a convenient selection of alternatives to harmful solvents [10–18]. However, the lack of relevant data, such as physicochemical properties and environmental impact, might hinder their implementation in scientific and technological processes [19]. The problem of missing data in solvents assessments is managed by default, substitution with value for nearest neighbor (homologue), and substitution with mean value for the entire chemical class [20].

Expectation-Maximization (E-M) algorithm [21] was developed in the 1970's and it is widely applied in different branches of sciences as the tool for the substitution of missing values [22]. It is applied to deal with missing values for the characterization of patients to predict breast cancer recurrence [23]. The algorithm is used to predict missing values in genetic arrays [24]. Application in chemistry include the prediction of missing data in environmental monitoring [25] or to construct regression models in the case of missing data in the raw dataset [26]. Other applications of E-M algorithm for predictions in the chemistry related field include the prediction of biomarkers essentiality [27] or the prediction of peptide bounding [28].

The aim of the study is to substitute missing values in the dataset characterizing organic solvents with the application of E-M algorithm, and to find relations between the characteristics from the estimated distribution. Toxicity, biodegradability, and bioavailability parameters are predicted with E-M algorithm.

2. Materials and Methods

2.1. Dataset

The dataset consists of 155 solvents that are described by 13 variables. The values of variables are extracted mainly from the Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals [29] and from material safety data sheets of solvents. Physicochemical properties include melting point, boiling point, vapor pressure, density, water solubility, Henry's law constant, logarithms of octanol-water partitioning coefficient, and logarithm of octanol-air partitioning coefficient. Also, toxicity towards rodents when being administered orally (Oral LD_{50}), toxicity towards rodents via inhalation exposure pathway (Inhalation LC_{50}), toxicity towards fish (Fish LC_{50}), half-life time needed for biodegradation, and logarithm of bioconcentration factors were taken as variables for analysis.

Solvents included in the dataset are compounds of different chemical classes–from hydrocarbons, terpenes, chlorinated solvents to alcohols, ketones, ethers, and esters and carboxylic acids. 85 out of the 155 solvents were fully characterized in terms of abovementioned variables, whereas the dataset contained at least one gap in case of the remaining solvents.

2.2. E-M Model

To complete the data we use E-M algorithm. This algorithm consists of two steps: an Expectation step or the E-step and a Maximization step or the M-step.

We observe a data $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$, where $\boldsymbol{y_i}$ are realizations of a random vector $\boldsymbol{Y}$, $i = 1, \ldots, n$. Let $\boldsymbol{Y}$ has the probability distribution function depending on a vector of unknown parameters $\boldsymbol{\Psi}$.

Let $\boldsymbol{X}$ be a k dimensional random vector corresponding to a complete-data $\boldsymbol{x} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$, where $\boldsymbol{x_i}$ are realizations of $\boldsymbol{X}$, $i = 1, \ldots, n$. We consider the case where the vector $\boldsymbol{X}$ has multivariate normal distribution, which means that $\boldsymbol{\Psi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a vector of means and $\boldsymbol{\Sigma}$ is a covariance matrix.

Suppose that there are G groups with distinct missing patterns. Then, the observed-data log likelihood can be expressed as

$$\log L(\boldsymbol{\Psi}) = \Sigma_{g=1}^{G} \log L_g(\boldsymbol{\Psi}). \tag{1}$$

The likelihood function for gth group formed from the observed data $\boldsymbol{y}_g = \left(\boldsymbol{y}_{1g}, \ldots, \boldsymbol{y}_{n_g g}\right)$ is, discarding a proportionality constant, given by

$$L_g(\boldsymbol{\Psi}) = \left|\boldsymbol{\Sigma}_g\right|^{-\frac{n_g}{2}} \exp\left\{-\frac{\mathbf{1}}{\mathbf{2}}\Sigma_{i=1}^{n_g}\left(\boldsymbol{y}_{ig} - \boldsymbol{\mu}_g\right)^T \boldsymbol{\Sigma}_g^{-\mathbf{1}}\left(\boldsymbol{y}_{ig} - \boldsymbol{\mu}_g\right)\right\} \tag{2}$$

An estimate $\hat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}$ can be obtained by solving the log likelihood equation

$$\frac{\partial \log L(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}} = 0 \tag{3}$$

The E-M algorithm approaches the problem of solving the incomplete-data log likelihood Equation (3) indirectly by proceeding iteratively in terms of complete-data log likelihood function $\log L_c(\boldsymbol{\Psi})$, where

$$L_c(\boldsymbol{\Psi}) = \left|(2\pi)^k \boldsymbol{\Sigma}\right|^{-\frac{n}{2}} \exp\left\{-\frac{\boldsymbol{1}}{\boldsymbol{2}}\Sigma_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}_i - \boldsymbol{\mu})\right\} \tag{4}$$

The E-step calculates the conditional expectation of the complete-data log likelihood, $E_{\boldsymbol{\Psi}^{(0)}}(\log L_c(\boldsymbol{\Psi})|\boldsymbol{y})$, given the observed data and the parameter estimates. Then, the M-step finds the parameter estimates to maximize the complete-data log likelihood from the E-step.

The steps are carried out until the value of

$$L\left(\boldsymbol{\Psi}^{(k+1)}\right) - L\left(\boldsymbol{\Psi}^{(k)}\right) \tag{5}$$

is smaller than arbitrarily amount in case of convergence of the sequence of likelihood values $\left(L\left(\boldsymbol{\Psi}^{(k)}\right)\right)_k$. The extended description of this procedure can be found in Supplementary Material. More details for normal distribution can be found in [30].

2.3. Dataset Preparation

We consider 155 solvents that are described by 13 attributes, which are listed in Table 1. Some chemical compounds do not have the value for the parameter "Inhalation LC_{50}", because they are not volatile, so there is no possibility to be intoxicated via inhalation. We decided to give them values 5001. This is the expert judgment caused by the threshold of danger to the environment. According to Globally Harmonized System of Classification and Labelling of Chemicals, the values above 5000 ppm are not characterized as "harmful if inhaled" [31].

Table 1. Basic statistics of the dataset.

Variable	Mean	Std. Dev.	Variance	Minimum	Maximum	*N*	*N* Missing
Melting point (°C)	−43.901	48.728	2374.453	−140	49.52	152	3
Boiling point (°C)	142.385	68.626	4709.488	20	323	155	0
Density (g cm^{-3})	0.952	0.214	0.046	0.62	1.68	155	0
Water solubility (mg dm^{-3})	116,796.63	244,339.68	5.97×10^{10}	0.000927	1,000,000	155	0
Vapor pressure (Pa)	11,901.626	28,631.781	8.2×10^{8}	0	241,900	155	0
Henry law constant (Pa m^3 mol^{-1})	60,736.714	267,946.02	$7.18 \times 1\,0^{10}$	8.03×10^{-6}	2,219,017	153	2
log K_{OW}	2.229	2.352	5.531	−2.32	8.73	155	0
log K_{OA}	4.434	1.999	3.995	1.451	12.101	152	3
Oral LD_{50} (mg kg^{-1})	3667.383	4658.48	21,701,436	5	31,500	120	35
Inhalation LC_{50} (ppm)	10,532.284	18,252.957	3.33×10^{8}	34	123,000	109	46
Fish LC_{50} (mg dm^{-3})	970.096	2813.093	7,913,490	0.1	16,700	98	57
BOD $t_{1/2}$ [days]	55.360	127.192	16,178	1	800	93	62
log BCF	1.154	1.016	1.032	−1.63	4.7	151	4

LD_{50}: lethal dose administered orally to rodents that kills half of population; LC_{50}: toxicity towards rodents via inhalation exposure pathway; Std. Dev.: Standard Deviation; K_{ow}: octanol-water partitioning coefficient; K_{OA}: octanol-air partitioning coefficient; BOD: biodegradation half-life; BCF: bioconcentration factor.

3. Results and Discussion

3.1. Basic Statistics

Table 1 shows the basic statistics of investigated dataset, including the number of missing data (N missing). For boiling point, density, water solubility, vapor pressure, and log K_{OW}, all of the values are available. The biggest problems with data availability are in case of Oral LD_{50}, inhalation LC_{50}, fish LC_{50}, and BOD $t_{1/2}$ (biodegradation half-life), as they are characterized by a big fraction of missing values.

3.2. Predictions with Bayesian Model

Application of E-M Algorithm

To complete the data set, we use the programming language SAS 4GL. There is a proper procedure in SAS, called PROC MI, which performs the E-M algorithm by function EM. The extensive description of this procedure can be found in [32].

We assume that the random vector X corresponding to a complete-data vector x has the multivariate normal distribution, $X \sim N(\mu,\Sigma)$ where μ is a vector of means and Σ is a covariance matrix. To approach that assumption, we use a logarithmic transformation of the properties Henry's law constant, Oral LD_{50}, Inhalation LC_{50}, fish LC_{50}, and BOD $t_{1/2}$.

In this case, we have to find a parameter $\Psi = (\mu, \Sigma)$ where μ is a vector of means and Σ is a covariance matrix of the unknown distribution. The initial estimates $\Psi^{(0)}$ are the means and the standard deviations from available cases. The correlations are set to zero.

On the prepared set, we carry out the E-M algorithm in SAS. To satisfy the convergence requirement the difference (5) has to be smaller than 0.0001. The method converges after 69 iterations. It means that

$$L(\Psi^{(69)}) - L(\Psi^{(68)}) < 0.0001 \quad (6)$$

due to (5). Thereby, we received the completion of the data set. The most important thing is that we received full information about the set by finding the distribution of the data, $N(\mu,\Sigma)$. The E-M algorithm evaluated the parameters μ and Σ. Despite the Mardia's kurtosis test has not shown that the data has a multivariate normal distribution we think that such a model gives a good approximation of relations. We will find these relations using principal component analysis (PCA) [33]. Table 2 shows the obtained results of first three principal components.

Table 2. The results of data treatment with principal component analysis. Dark red is for very negative values, yellow is for neutral values and green stands for positive values.

	Comp. 1	Comp. 2	Comp. 3
Melting point	−0.4448	−0.1465	−0.0451
Boiling point	−0.4963	−0.0883	0.0598
Density	−0.0607	−0.0709	−0.4854
Water solubility	0.1150	−0.3744	0.0708
Vapor pressure	0.3478	0.0295	−0.0888
log Henry law const	0.1247	0.5103	0.0289
log K_{OW}	−0.2462	0.4340	0.1635
log K_{OA}	−0.4352	−0.1795	0.1165
log Oral LD_{50}	−0.0528	0.0933	0.5678
log Inhalation LC_{50}	0.2287	0.0600	0.4662
log fish LC_{50}	0.1673	−0.2642	0.2638
log BOD $t_{1/2}$	0.0848	0.2915	−0.3085
log BCF	−0.2492	0.4202	0.0174

The first three components explain 65.9% of variability of the raw dataset. Fourth principal component explained 8.4% of variability, and it was decided not to include it in the assessment result. The first component consists of melting point, boiling point, vapor pressure, and log K_{OA}, and it explains 26.2% of the initial variability. This component can be identified as responsible for the characterization of solvents in terms of their basic physicochemical properties, especially volatility. The second component is loaded with Henry's law constant, log Kow, and log BCF, and it explains 24.4% of variability. This relation can be explained by the polarity of solvents. All three variables are related to interphase transfer and the ability to be transferred out of water. Apart from these three variables, weaker negative loading (−0.3774) is observed for water solubility, which additionally supports the polarity related origin of this group. The third component is formed by inhalation and oral toxicities and the negative loading of density. It carries 15.3% of initial dataset variability. It can be defined as toxicity relation and the presence of density in this group is due to the fact that more toxic solvents are usually more dense (i.e., chlorinated solvents).

The prediction of missing values are presented in Table 3—shaded are modeled with the E-M algorithm, whereas non-shaded are input data. Algorithm allows to substitute missing values for alkyl glycerol esters (numbers 15–29 in the Table 3), bio-based solvents originating from biodiesel production. Glycerol is formed as a byproduct during biodiesel production, and it is a platform molecule for the synthesis of its alkylated ester derivatives [34]. Because they originate from renewable resource, undergo biodegradation, and are low cost, they are potentially attractive from a Green Chemistry point of view. However, they are not fully characterized in terms of their toxicology or environmental fate related properties. The means of completed values are of 2017 mg kg^{-1} LD_{50} by oral administration (mean for solvents with available data—3667 mg kg^{-1}), 13945 ppm of LC_{50} by inhalation (mean for solvents with available data—4658 ppm), 158 mg dm^{-3} of LC_{50} towards fish (mean for solvents with available data—970 mg dm^{-3}), 1.96 days of biodegradability half-lifes (mean for solvents with available data—55 days), and 0.46 of logarithm of bioconcentration factors (mean for solvents with available data—1.15). According to Globally Harmonized System of Classification and Labelling of Chemicals, oral toxicity of >2000 mg kg^{-1} and inhalation toxicity of >5000 ppm indicate that they are chemicals of low acute toxicity. The predicted results show that alkyl glycerol esters may be toxic (especially towards fish) and they should be characterized in this manner to confirm their green status. Other predicted that missing values show that alkyl glycerol esters are biodegradable and they do not undergo bioaccumulation.

Table 3. Input and completed values (shaded) for solvents that were not fully characterized. Green color indicate predicted values.

	Solvent	CAS Number	Oral LD_{50} (mg kg^{-1})	Inhalation LC_{50} (ppm)	Fish LC_{50} (mg dm^{-3})	BOD $t_{1/2}$ (days)	log BCF
1	Cyclopentane	287-92-3	11,400	57,377	100	10.6	1.61
2	Octane	111-65-9	7930	25,260	100	13.7	3.289
3	Nonane	111-84-2	218	3200	6.5	16.4	2.651
4	Decane	124-18-5	5000	1369	500	40.0	2.158
5	Tridecane	629-50-5	5000	41	0.9	18.2	2.979
6	Tetradecane	629-59-4	15,000	5001	1000	38.7	3.036
7	Pentadecane	629-62-9	5000	5001	100.1	39.6	2.34
8	1-pentene	109-67-1	3197	21,800	90.7	17.0	1.349
9	1-hexene	646-04-8	10,000	32,000	5.6	10.7	1.91
10	1-heptene	592-76-7	5000	27,986	175	12.6	2.372
11	1-octene	111-66-0	10,000	8500	6.8	9.3	2.819
12	1-nonene	124-11-8	4390	7116	5.0	9.6	3.266
13	Pentanol	71-41-0	2200	6119	370	4	0.463
14	oleic alcohol	143-28-2	9604	13,049	46.7	9.0	2.623
15	1,3-di-iso-propoxy-2-propanol	13021-54-0	1267	2725	33.5	1.6	0.5
16	1,3-dimethoxypropan-2-ol		1393	3794	104.7	2.5	0.5
17	1,3-di-*n*-butoxy-2-propanol		1130	885	69.4	2.1	0.603
18	1-ethoxy-3-iso-propoxy-2-propanol		1256	1889	377.7	4.3	0.5
19	1-methoxy-3-(propan-2-yloxy)propan-2-ol		1498	2945	160.8	2.1	0.5
20	1-*n*-butoxy-3-ethoxy-2-propanol		2220	2347	232.2	1.8	0.5
21	1-*n*-butoxy-3-iso-propoxy-2-propanol		3047	4273	188.0	1.5	0.168
22	1-*n*-butoxy-3-methoxy-2-propanol		1883	2582	197.6	2.1	0.5
23	1-*tert*-butoxy-3-ethoxy-2-propanol		2568	4601	97.1	1.3	0.5
24	1-*tert*-butoxy-3-methoxy-2-propanol		1477	3305	35.3	1.4	0.5
25	3-butoxypropane-1,2-diol		3875	2818	203.4	1.5	0.5
26	3-ethoxypropane-1,2-diol		2538	2663	186.2	1.9	0.5
27	3-methoxypropane-1,2-diol		2081	1985	272.4	2.6	0.5
28	3-*n*-butoxy-1-*tert*-butoxy-2-propanol		5660	5167	51.7	1.4	0.517
29	Isopropylidene glycerol	100-79-8	7000	167,197	16,700	1.3	0.125
30	Methoxycyclopentane	5614-37-9	1500	5250	34.9	6.6	0.721
31	Benzyl ethyl ether	539-30-0	2428	2625	38.6	6.6	1.374
32	1,2,3-trimethoxypropane		1305	2815	135.8	5.3	0.5
33	1,2,3-*tri*-*n*-butoxypropane		4390	5001	261.2	4.5	2.276
34	2-methylfuran		1965	9352	94.3	16.0	0.725
35	2-methyltetrahydrofuran		4500	24,083	319.6	6.4	0.343
36	3-*n*-butoxy-1-*tert*-butoxy-2-methoxypropane		2392	1656	95.4	2.9	1.094
37	Isosorbide dimethyl ether		1545	18,269	213.8	4.9	0.5
38	Dioxolane	646-06-0	2833	3,7363	31.0	6.1	0.149
39	Benzaldehyde	100-52-7	1300	1304	1.07	10	1.1
40	gamma-valerolactone	108-29-2	2800	1186	756.6	7.8	0.5
41	Dihydrolevoglucosenone		2021	2916	59.3	4.4	0.5
42	1,8-cineole	470-82-6	2480	1000	102	26.4	1.41
43	3-carene	13466-78-9	4800	8800	17.9	28	2.673
44	Neryl acetate	141-12-8	4550	5001	41.7	8.7	2.365
45	Propionic acid	79-19-4	3500	5422	51	1	0
46	Ethyl formate		1850	9800	276.6	15	0.5
47	Butyl levulinate	2052-15-5	5000	5001	26.3	3.3	0.278
48	Ethyl levulinate	539-88-8	5000	4735	121.3	3.3	0.5
49	Glycerol triacetate	102-76-1	3000	5001	72.5	2.2	0.5
50	Methyl caprylate	111-11-5	10,800	9987	95	7.0	1.856
51	Methyl lactate	27871-49-4	5000	1350	828.6	11.8	0.5
52	Methyl levulinate	624-45-3	2051	2888	92.7	3.4	0.5
53	Methyl linoleate	112-63-0	3977	5001	4.5	20.4	3.051
54	Isopropyl myristate	110-27-0	8348	11,207	8.4	10.7	3.07
55	Methyl oleate	112-62-9	2000	5001	6.1	18.9	2.694
56	Methyl palmitate	112-39-0	4786	5001	1.8	9.4	2.789
57	Isopropyl palmitate	142-91-6	17,781	45,414	50.3	13.0	1.725
58	Methyl stearate	112-61-8	5237	5001	2.8	10.4	1.46
59	Tributyl 2-acetylcitrate	77-90-7	31,500	226,174	60	14	1.6
60	Benzyl benzoate	120-51-4	1700	665	6.2	5.3	2.357
61	*cis*-1,2-dichloroethene	156-59-2	1393	13,700	54.2	180	1.18
62	1,1-dichloroethane	75-34-3	725	13,000	100.0	154	1.24
63	1,1,1,2-tetrachloroethane	630-20-6	670	2100	20	134.0	1.559
64	1-chloropropane	540-54-5	2000	14,034	117.8	30	0.763
65	1-chlorobutane	109-69-3	2670	11,879	101.2	18.2	1.333
66	1-chloropentane	543-59-9	3379	11,804	27.8	10.5	1.402
67	Dimethyl sulphide	75-18-3	535	5156	87.1	10.7	0.561
68	Dimethyl sulfoxide	67-68-5	2758	4291	36.9	1.6	0.349
69	Diethylamine	109-89-7	540	4000	218.5	5.0	0.21
70	2-pyrrolidone	616-45-5	2030	1083	152.5	2.5	0.5

Esters (mainly methyl or isopropyl esters of fatty acids) are characterized by low completed inhalation toxicity, rather low oral toxicity, but they some have low values (at the level of single mg dm^{-3}) of toxicity towards fish, which suggest they are potential threats to aquatic life. Their completed values suggest that they are biodegradable, with half-lifes at the level of few to 20 days. Gamma valerolactone is considered to be green solvent, and recently it gained much attention [35].

E-M algorithm showed that LC_{50} by inhalation is 1186 ppm and LC_{50} towards fish is 756.6 mg dm^{-3} and biodegradability half-life is 7.8 days. These values suggest that this compound is not very toxic and it undergoes biodegradation.

Chloropropane, chlorobutane, and chloropentane are characterized by mean predicted missing value of inhalation LC_{50} = 12572 ppm, which is in accordance with available values for cis-1,2-dichloroethene (13,700 ppm) and 1,1-dichloroethane (13,000 ppm). The completed values show that they are slightly less toxic than the mean of the entire dataset. Mean value of LC_{50} towards fish is 82 mg dm^{-3}, what shows they can be toxic to fish as the mean value of this parameters equals to 970 mg dm^{-3}. Biodegradability half-lifes are predicted to be 18.2 and 10.5 days for 1-chlorobutane and 1-chloropentane, respectively. This value for 1-chloropropane of 30 days is available in the original dataset. What is more, the completed value for 1,1,2,2-tetrachloroethane is equal to 134 days. Chlorinated solvents are not concerned as green solvents and predicted missing values confirm this statement.

To obtain the information about the standard errors bootstrap analysis is performed [36]. The results of analysis are presented in Table 4. Standard errors are small for almost all of the variables. Only those variables, which do not contain data gaps and are not transformed, have the high standard errors.

Table 4. Standard errors of predictions calculated with bootstrap.

Variable	Mean	Mean Error
Melting point (°C)	−43.1378	−0.0056
Boiling point (°C)	142.3852	−0.4245
Density (g cm^{-3})	0.9521	−0.0005
Water solubility (mg dm^{-3})	116,796.6328	−279.7044
vapor pressure (Pa)	11,901.6258	403.5914
Henry law constant (Pa m^3 mol^{-1})	2.4677	0.0955
log K_{OW}	2.2285	0.0125
log K_{OA}	4.4644	−0.0269
Oral LD_{50} (mg kg^{-1})	7.6122	−0.0093
Inhalation LC_{50} (ppm)	8.2734	0.0014
Fish LC_{50} (mg dm^{-3})	4.2248	−0.0069
BOD $t_{1/2}$ (days)	2.2431	0.0065
log BCF	1.1450	0.0076

4. Conclusions

E-M algorithm is useful in predicting of organic solvents missing parameters. Algorithm allows for completing 35 values of LD_{50} by oral administration to rodents, 46 values of LC_{50} by inhalation, 57 values of LC_{50} towards fish, 57 values of biodegradability, and 62 values of bioconcentration factors. Some of the solvents that are considered in this study are promising from the Green Chemistry point of view, even though they have not been fully characterized yet.

E-M algorithm can be useful in characterization of other novel, potential green alternatives of other chemicals. It is important for the characterization of chemicals for their rapid screening.

Supplementary Materials: The following are available online. E-M algorithm description.

Author Contributions: G.Ł. performed the calculations with PCA and E-M algorithm, and wrote substantial part of paper, F.P.-P. prepared the dataset and wrote substantial part of paper, M.T. prepared the dataset, interpreted the results and wrote substantial part of paper, J.N. was guiding the research.

Funding: This research received no external funding.

Acknowledgments: F.P.-P. thanks Xunta de Galicia for financial support as a postdoctoral researcher of the I2C program. Authors would like to express their gratitude to Karol Dziedziul for pointing to E-M algorithm and for his precious remarks and suggestions.

Conflicts of Interest: Authors declare no conflict of interests

References

1. Anastas, P.T.; Warner, J.C. *Green Chemistry: Theory and Practice*; Oxford Univerity Press: New York, NY, USA, 1998.
2. Anastas, P.; Eghbali, N. Green chemistry: Principles and practice. *Chem. Soc. Rev.* **2010**, *39*, 301–312. [CrossRef] [PubMed]
3. Gu, Y.; Jérôme, F. Bio-based solvents: An emerging generation of fluids for the design of eco-efficient processes in catalysis and organic chemistry. *Chem. Soc. Rev.* **2013**, *42*, 9550–9570. [CrossRef] [PubMed]
4. Pena-Pereira, F.; Kloskowski, A.; Namieśnik, J. Perspectives on the replacement of harmful organic solvents in analytical methodologies: A generation of eco-friendly alternatives. *Green Chem.* **2015**, *17*, 3687–3705. [CrossRef]
5. Linak, E.; Bizzari, S.N. *Global Solvents: Opportunities for Greener Solvents*; IHS Markit: London, UK, 2013.
6. Nerín, C.; Salafranca, J.; Aznar, M.; Batlle, R. Critical review on recent developments in solventless techniques for extraction of analytes. *Anal. Bioanal. Chem.* **2009**, *393*, 809–833. [CrossRef] [PubMed]
7. Cave, G.W.V.; Raston, L.; Scott, J.L. Recent advances in solventless organic reactions: Towards benign synthesis with remarkable versatility. *Chem. Commun.* **2001**, *21*, 2159–2169. [CrossRef]
8. Pena-Pereira, F.; Tobiszewski, M. *The Application of Green Solvents in Separation Processes*; Pena-Pereira, F., Tobiszewski, M., Eds.; Elsevier: Cambridge, UK, 2017.
9. Anastas, P.T. Green Chemistry as Applied to Solvents. In *Clean Solvents—Alternative Media for Chemical Reactions and Processing*; Abraham, M.A., Moens, L., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, USA, 2002; Volume 1991, pp. 1–9.
10. Curzons, A.D.; Constable, D.C.; Cunningham, V.L. Solvent selection guide: A guide to the integration of environmental, health and safety criteria into the selection of solvents. *Clean Technol. Environ. Policy* **1999**, *1*, 82–90. [CrossRef]
11. Jiménez-González, C.; Curzons, A.D.; Constable, D.J.C.; Cunningham, V.L. Expanding GSK's Solvent Selection Guide—Application of life cycle assessment to enhance solvent selections. *Clean Technol. Environ. Policy* **2005**, *7*, 42–50. [CrossRef]
12. Alfonsi, K.; Colberg, J.; Dunn, P.J.; Fevig, T.; Jennings, S.; Johnson, T.A.; Kleine, H.P.; Knight, C.; Nagy, M.A.; Perry, D.A.; et al. Green chemistry tools to influence a medicinal chemistry and research chemistry based organisation. *Green Chem.* **2008**, *10*, 31–36. [CrossRef]
13. Henderson, R.K.; Jiménez-González, C.; Constable, D.J.C.; Alston, S.R.; Inglis, G.G.A.; Fisher, G.; Sherwood, J.; Binks, S.P.; Curzons, A.D. Expanding GSK's solvent selection guide—Embedding sustainability into solvent selection starting at medicinal chemistry. *Green Chem.* **2011**, *13*, 854–862. [CrossRef]
14. Moity, L.; Durand, M.; Benazzouz, A.; Pierlot, C.; Molinier, V.; Aubry, J. Panorama of sustainable solvents using the COSMO-RS approach. *Green Chem.* **2012**, *14*, 1132–1145. [CrossRef]
15. Prat, D.; Pardigon, O.; Flemming, H.-W.; Letestu, S.; Ducandas, V.; Isnard, P.; Guntrum, E.; Senac, T.; Ruisseau, S.; Cruciani, P.; et al. Sanofi's solvent selection guide: A step toward more sustainable processes. *Org. Process Res. Dev.* **2013**, *17*, 1517–1525. [CrossRef]
16. Tobiszewski, M.; Tsakovski, S.; Simeonov, V.; Pena-Pereira, F. A solvent selection guide based on chemometrics and multicriteria decision analysis. *Green Chem.* **2015**, *17*, 4773–4785. [CrossRef]
17. Tobiszewski, M.; Namieśnik, J.; Pena-Pereira, F. Environmental risk-based ranking of solvents using the combination of a multimedia model and multi-criteria decision analysis. *Green Chem.* **2017**, *19*, 1034–1042. [CrossRef]
18. Prat, D.; Wells, A.; Hayler, J.; Sneddon, H.; Mcelroy, C.R.; Abou-shehada, S.; Dunn, P.J. CHEM21 selection guide of classical- and less classical-solvents. *Green Chem.* **2016**, *18*, 288–296. [CrossRef]
19. Byrne, F.P.; Jin, S.; Paggiola, G.; Petchey, T.H.M.; Clark, J.H.; Farmer, T.J.; Hunt, A.J.; Mcelroy, C.R.; Sherwood, J. Tools and techniques for solvent selection: Green solvent selection guides. *Sustain. Chem. Process.* **2016**, *4*, 1–24. [CrossRef]
20. Alder, C.M.; Hayler, J.D.; Henderson, R.K.; Redman, A.M.; Shukla, L.; Shuster, L.E.; Sneddon, H.F. Updating and Expanding GSK's Solvent Sustainability Guide. *Green Chem.* **2016**, *18*, 3879–3890. [CrossRef]
21. Do, C.B.; Batzoglou, S. What is the expectation maximization algorithm? *Nat. Biotechnol.* **2008**, *26*, 897–899. [CrossRef] [PubMed]
22. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef] [PubMed]

23. Vazifehdan, M.; Moattar, M.H.; Jalali, M. A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. *J. King Saud Univ. Comput. Inf. Sci.* **2018**. [CrossRef]
24. Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520–525. [CrossRef] [PubMed]
25. Miller, L.; Xu, X.; Wheeler, A.; Zhang, T.; Hamadani, M.; Ejaz, U. Evaluation of missing value methods for predicting ambient BTEX concentrations in two neighbouring cities in Southwestern Ontario Canada. *Atmos. Environ.* **2018**, *181*, 126–134. [CrossRef]
26. Stanimirova, I.; Serneels, S.; Van Espen, P.J.; Walczak, B. How to construct a multiple regression model for data with missing elements and outlying objects. *Anal. Chim. Acta* **2007**, *581*, 324–332. [CrossRef] [PubMed]
27. Wei, G.; Margolin, A.A.; Haery, L.; Brown, E.; Cucolo, L.; Julian, B.; Shehata, S.; Kung, A.L.; Beroukhim, R.; Golub, T.R. Chemical genomics identifies small-molecule MCL1 repressors and BCL-xL as a predictor of MCL1 dependency. *Cancer Cell* **2012**, *21*, 547–562. [CrossRef] [PubMed]
28. Chang, K.Y.; Suri, A.; Unanue, E.R. Predicting peptides bound to I-Ag7 class II histocompatibility molecules using a novel expectation-maximization alignment algorithm. *Proteomics* **2007**, *7*, 367–377. [CrossRef] [PubMed]
29. Mackay, D.; Shiu, W.-Y.; Ma, K.-C.; Lee, S.C. *Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2016.
30. Dellacherie, C.; Meyer, P.-A. *Probabilities and Potential*; North-Holland Mathematics Studies; North-Holland Publishing Co.: Amsterdam, The Netherland, 1978.
31. Schafer, J.L. *Analysis of Incomplete Multivariate Data*; Chapman and Hall: New York, NY, USA, 1997.
32. OSHA. *Chemical Hazard Classification and Labeling: Comparison of OPP Requirements and the GHS*; OSHA: Washington, DA, USA, 2004.
33. Mardia, K.V.; Kent, J.T.; Bibby, J.M. *Multivariate Analysis*; Academic Press: London, UK, 1979.
34. García, J.I.; García-Marín, H.; Mayoral, J.A.; Pérez, P. Green solvents from glycerol. Synthesis and physico-chemical properties of alkyl glycerol ethers. *Green Chem.* **2010**, *12*, 426–434.
35. Qi, L.; Mui, Y.F.; Lo, S.W.; Lui, M.Y.; Akien, G.R.; Horvaáth, I.T. Catalytic conversion of fructose, glucose, and sucrose to 5-(hydroxymethyl) furfural and levulinic and formic acids in γ-valerolactone as a green solvent. *ACS Catal.* **2014**, *4*, 1470–1477. [CrossRef]
36. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Cambridge University Press: Cambridge, UK, 1997.

Sample Availability: Samples of the compounds are not available from the authors.