RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

Correlation of Selected Molecular Properties and Recovery Values in Volatile Organic Compounds Analysis: Comparison of Two Water Matrices

Ivana Ivančev-Tumbas[1], Tatjana-Djaković Sekulić[1], Jelena Molnar, Aleksandra Tubić, Jasmina Agbaba , Jelena Tričković, Marijana Kragulj

University of Novi Sad Faculty of Sciences, Department for Chemistry, Biochemistry and Environmental Protection, Trg D. Obradovića 3, 21000 Novi Sad, Republic of Serbia

Abstract

This study investigates if certain molecular properties can influence the recovery of 18 volatile organic compounds (VOCs) in water under the applied analytical conditions, a purge and trap gas chromatographic method with mass spectrometric detection (P&T GC MS). Statistical and quantitative structure property relationship (QSPR) analyses were applied to find correlations between molecular parameters and analytical recoveries in two different water matrices (clean water (CW) produced in the laboratory and natural groundwater (GW)) at two different concentration levels. At 1µg/L, most compounds had higher recoveries in CW than in GW, whereas at 15µg/L, the recoveries were higher in the GW matrix. Polarity number and hydrophilic factor were correlated with the recovery differences at both concentration levels in GW. Polarity was significant in the distinction of recovery differences for CW and GW matrices at low concentration, while air diffusivity had an acceptable correlation with recovery differences for both matrices at the higher concentration. Further correlation of the recoveries themselves with the molecular properties was made by multivariate linear regression (MLR) resulting in a QSPR model. This was only possible for GW at the low 1 µg/L concentration. Partial least square analysis indicated that hydrophilic factor, polarity, and molecular weight were the most important properties investigated. No significant correlation was found in CW matrix or in the higher concentration level in GW matrix, which implies that the most significant properties might only be relevant for the VOCs recovery at low concentrations and only if the matrix contains other water constituents (DOC, salt).

Keywords: VOC; Purge & Trap-GC/MS; recovery; water matrix; stepwise multiple linear regression; validation

1. Introduction

Volatile organic compounds (VOCs) may be present in different types of water. In surface water they mostly originate from traffic pollution (benzene, toluene, ethylbenzene and xylene, e.g. BTEX compounds) and agricultural or industrial pollution (i.e., chlorobenzenes, 1,2-dichloroethane, etc.), whereas in drinking water they are mostly related to disinfection practices (i.e., trihalomethanes). In groundwater, their presence is usually related to pollution caused by waste leakage or accidental pollution. The importance of their presence is underlined in several publications.[1-4] These compounds are widely used as solvents, cosmetic products, fuels, furnishings, etc. They can easily end up in wastewaters and may pose a risk to the environment. Some VOCs are regulated by the Water Framework Directive[5], but not all[6]. A year long survey in 1998/99 by Nikolau et al.[7] showed that in rivers and lake waters in Greece, VOCs concentrations were usually of the order of several µg/L or lower. In wastewaters their concentrations were higher for some compounds (reaching tens or even hundreds of µg/L).

Nowadays, VOC analysis in water is routine. Gas chromatography is mostly applied in combination with different extraction techniques: head space, purge/trap and solid phase micro-extraction. Different types of detectors can be used, such as flame ionisation, electron capture or mass spectrometric detection.[8] The technique used in this research - purge and trap gas chromatographic method with mass spectrometric detection

[1] Corresponding authors:
ivana.ivancev-tumbas@dh.uns.ac.rs, Tel: + 381 21 485 27 46
tatjana.djakovic-sekulic@dh.uns.ac.rs, Tel. +381 21 485 27 42
University of Novi Sad Faculty of Sciences, Department for Chemistry, Biochemistry and Environmental Protection, Trg D. Obradovića 3, 21000 Novi Sad, Republic of Serbia; Fax: +381 21 454 065;

(P&T GC/MS), is recognised as highly sensitive and reliable for very low concentration ranges of various compounds.[6]

A survey of the available literature shows an abundance of data on the precision and accuracy of the different methods. However, almost no attempts have been made to compare recovery values for different matrices. One exception is a paper by Barco-Bonilla et al.[4] By comparing calibration curve ratios for different wastewater effluents and liquid chromatography/mass spectrometry grade (LC/MS) water (ratios were within the range of 0.8-1.2 for water spiked with 1 and 5 µg/L), they showed that there were no matrix effects. For this purpose, purge and trap (P&T) was coupled with gas chromatography (GC) triple quadruple mass spectrometry and was used for the analysis of a membrane bioreactor, extended aeration, maturation pond and anaerobic pond. In Standard Methods for the Examination of Water and Wastewater for Volatile Organic Compounds (Method 6200) published in 1998[9] one can find data related to recovery values at a concentration level of 0.5 µg/L in reagent water. They are within the range of 85-110% for the compounds-of-interest, with relative standard deviations (RSD) in the range 6-10%. For reagent and raw water, recovery values can be found in EPA 524.2.[10] (higher values for raw water for almost all analytes). Concentration levels are different and depend on the water type.

Comparison of recovery values for different matrices is important from the point of view of analytical efficiency. Changes in the recovery due to the matrix are influenced by water constituents and are not a means for controlling the accuracy and precision of a method. Recovery is the amount of a compound that reaches the GC for analysis relative to the amount that was originally present in the sample. The ideal recovery value is 100%. However, acceptable values range from 70-130% for most analytes in gas chromatography methods. In P&T analysis it is essential to vaporise the substances and partition them into the gas phase. The vapour pressure, solubility and extraction temperature affect the procedure. Once the procedure is established, we assume that the influence of the water matrix, if present, can be understood as a possibility to alter the transfer of the analyte into the gaseous phase due to interactions with the water constituents, while other analytical conditions are kept constant as defined by the method. The aim of this study was to investigate if certain molecular properties could influence recovery in the analysis of 18 volatile organic compounds (VOC) in two different water matrices (i.e., clean water (CW) produced in the laboratory, and natural groundwater (GW)) at two different concentration levels (1 and 15 µg/L)). Statistical analysis followed by quantitative structure property relationship analysis (QSPR) was applied. In general, QSPR models present relationships between properties of series of molecules and their structural characteristics (derived either from experiments or theoretically). Correlation of the selected molecular parameters and the recovery (assumed as a property under the given analytical conditions) was not studied with the aim to predict recovery but rather to find out if any of the selected molecular parameters could influence recovery under the applied analytical conditions.

2. Materials and methods

2.1. Water matrices

The clean water matrix (CW) was produced in the laboratory by LABCONCO (WaterPro RO/PS Station, Kansas City, USA) system (water of ASTM (American Society for Testing and Materials) Type I quality). It was then boiled for 15 min and purged for 1 hour by nitrogen. It was stored in a glass bottle and prepared daily. Dissolved organic carbon (DOC) content was less than the practical quantitation limit of 0.5 mg/L, and electrical conductivity was 0.055 µS/cm**.** Measurements of DOC were performed after sample acidification with concentrated hydrochloric acid to pH=2, according to method SRPS ISO 8245:2007.[11] pH measurement was carried out on a WTW InoLab (Weilheim in Oberbayern, Germany) portable instrument.

The natural groundwater (GW) matrix was taken from the Danube riparian area (25-30 m depth) free from VOC. Its characteristics were: 3.6 mg/L DOC, pH 7.02, electrical conductivity 770±101 µS/cm, dry residue of 415 mg/L (at 105°C, according to standard method 2540 B[9]).

2.2. Chemicals and reagents

Standards with concentrations of 2000 µg/mL in methanol as Volatile Organic Compounds Mix 7 (Dr. Ehrenstorfer GmbH. Germany) (chloroform, 1,1,1-trichloroethane (1,1,1-TCE), 1,2-dichloroethane (1,2-DCE), benzene, trichloroethene, bromodichloromethane (BDCM), dibromochloromethane (DBCM), bromoform, 1,4-dichlorobenzene), Volatile Organic Compounds Mix 8 (Supelco) (chlorobenzene, ethylbenzene, o-xylene, m-xylene, p-xylene, tetrachloroethene, toluene, 1,2-dichlorobenzene), vinyl chloride (Supelco) and internal

standard (IS) fluorobenzene (Supelco) were used to prepare working solutions. Dilution was done with 99.9% methanol (J.T. Baker, Avantor Performance Materials B.V., Deventer, The Netherlands).

2.3. Purge and trap conditions

A Tekmar Dohrmann 3100 Sample Concentrator with Vocarb 3000 trap (Carbopack B, 10 cm/Carboxen-1000, 6 cm/Carboxen 1001, 1 cm, *Supelco,* Sigma-Aldrich Co., St. Louis, USA) was used. 5mL aliquots of sample were dispensed into the 5 mL purging device with a gas tight syringe. The sample was purged with a stream of helium at 37.4 mL/min for 11 min at ambient temperature.

Before each sample analysis, the purge and trap system was baked (270°C for 3 min). This was followed by blank analysis of CW. After sample loading, desorption by heating the Vocarb 3000 trap was carried out at 250°C for 2 min. The injector was set to split mode (30:1).

2.4. Chromatographic and MSD conditions

An Agilent (USA) Gas chromatograph (GC) 7890A with 5975C Mass selective detector (MSD) was used. The inlet temperature was 110°C. An Agilent J&W Scientific DB-5MS column (30 m x 0.25 mm x 0.25 µm) was used. The oven temperature was set at 35°C (5 min) and raised to 100°C at 15°C/min and to 225°C at 25°C/min for 3 min, with a total run time of 20 min. Electron ionization was carried out at 70 eV, the source temperature and GC interface were set at 230°C and 280°C respectively. Acquisition was performed in selected ion monitoring/ scanning mode from 50-300 amu (SIM/SCAN). Quantification was carried out based on the target ions given in Table 1. Compound confirmation was performed simultaneously based on the qualifier ions (Table 1). The ions were selected from among the total ion chromatograms obtained by full mass range scans. The GC carrier gas was helium of 99.999% purity, supplied by Messer (Novi Sad, Serbia).

Table 1. Target ions and qualifiers in GC/MS analysis of VOCs

Compounds	Target ion	Qualifier ions
Vinyl chloride	62	64
Chloroform	83	85
1,1,1-TCE	97	99
1,2-DCE	62	98
Benzene	78	77
Trichloroethene	95	130, 132
BDCM	83	127, 129
Toluene	91	92
DBCM	129	208, 173
Tetrachloroethene	166	168, 129
Chlorobenzene	112	77
Ethylbenzene	91	106
m+p- xylene	106	91
o- xylene	106	91
Bromoform	173	171, 252
1,2-dichlorobenzene	146	148, 252
1,4-dichlorobenzene	146	148, 252
Fluorobenzene (IS)	96	97

2.5. Calibration and quantification

Calibration curves were obtained by spiking CW matrix with 17 VOCs mixture in methanol by gas-tight syringe, in accordance with the instructions given in Standard Methods 6200B[9], with further sample processing of the samples as explained in Standard Methods 6200B[9]. 9 point calibration curves were made using concentrations of 0.4, 0.5, 1.0, 2.0, 5.0, 8.0, 10, 14 and 25 µg/L and for vinyl chloride within the range of 0.2-25 µg/L. The internal standard concentration was 10 µg/L. For the curves in the 0.4-25 µg/L concentration range, the coefficients of determination (R^2) were within the range of 0.991-0.998.

As a part of routine quality control, a calibration verification standard (CVS) from the same source and a laboratory control standard (LCS) from an independent source were used for verification. A criterion of ±15% of difference in comparison to the initial calibration was accepted in accordance with EPA 8000B.[12]. The obtained results indicated that all the measured values were within the range of ±15% of the expected value.

RSDs for 50 measurements of the calibration verification standard (4 µg/L CVS) were collected and for all compounds were lower than 10%. LCS measurements ranged from 70% to130% of the expected concentration, in accordance with EPA 8000B.[12]

2.6. Method performance

EPA Methods 5030B [13] and 8260B [14] were used to develop an internal laboratory procedure for the analysis. Method detection limits were determined along the guidelines given by Glase et al. [15] (analyte added in a concentration which is 1-5 times the estimated method detection level (MDL)). The MDL was determined based on the RSD of 6 measurements of both spiked CW and GW matrix at a concentration level of 0.4µg/L, except for vinyl chloride (0.2µg/L). The practical limit of quantitation (PQL) was calculated as 5 x MDL according to EPA Method 8260B.[14] Repeatability for this concentration level was determined as a RSD of 6 consecutive measurements.

Precision was assessed as the RSD of recovery values determined for four samples analysed in 6 series in duplicates: clean water matrix at concentrations of 1 µg/L and 15 µg/L (CW_1 and CW_{15}) and groundwater matrix at the same concentrations (GW_1 and GW_{15}). One series consisted of a blank sample (non- spiked CW) and duplicates of the four mentioned samples. Additionally, the groundwater matrix (GW) was checked for the presence of VOC.

2.7. Recovery comparison

The recovery comparison for the two water matrices was carried out by statistical analysis of 18 VOCs for 12 replicates at both concentration levels, i.e. 1 and 15 µg/L. The concentration levels selected are similar to the VOC levels defined as relevant environmental quality standards for surface waters (Directive 2008/105/EC)[16].

Nine molecular parameters (descriptors) (water solubility (S, mg/L), Henry's law constant (HLC, atm-m^3/mol), octanol-water partition coefficient (K_{ow}), air ($D_{i,a}$) and water diffusivity at 25°C ($D_{i,w}$) taken from USEPA,[17] as well as molecular mass (MW), polarity number (Pol), hydrophilic factor (Hy) and molar refractivity calculated by Dragon [18]) were selected for the study. Descriptor values are given in Supplementary Table S2. The influence of molecular specific parameters on differences in recovery, i.e. its median value, was analysed by finding correlations between differences in recovery for the following: two concentrations in clean water (CW_{1-15}), two concentrations in groundwater (GW_{1-15}), a lower concentration in two different matrices $(CW\text{-}GW)_1$ and a higher concentration in two different matrices $(GW\text{-}CW)_{15}$ for compounds 1-18. In addition, the recovery values themselves were correlated with the molecular descriptors.

In order to visualize similarities and differences in water matrices, hierarchical clustering analysis (HCA) was used (Ward's linkage method). Furthermore, linear regression followed by stepwise multiple linear regression (MLR) (forward selection method) was applied[19, 20]. QSPR models for median recovery values based on a few deliberately selected explanatory variables (molecular descriptors) were constructed. The data were organized in matrices **X**(18x9) where the rows represented the 18 investigated compounds (VOC), and the columns corresponded to the 9 molecular parameters . The independent variables were the 9 selected descriptors while the dependent variables were the observed median values of recovery for 12 measurements of each of the 18 VOCs. Additionally, the partial least squares (PLS) method was used since it is a well-known approach for the analysis of multidimensional data sets.[21-25]

All the calculations were carried out by STATISTICA v. 10.0.[27] The data were mean-centred (subtracting the mean and dividing by the standard deviation[26]) before any statistical operation in order to prevent the highly abundant components from dominating in the final result over the components present in much smaller quantities.

3. Results and discussion

3.1. Method performance

All the recovery values at the 0.4 µg/L concentration level (0.2µg/L for vinyl chloride) were within the acceptable range (70-130%), with the exceptions of chloroform and toluene, due to the laboratory contamination. This contamination affected only the low 0.4 µg/L concentration level. This caused the

determined MDL and PQL values for these analytes to be slightly higher (i.e. PQLs of 1.60 and 1.06 µg/L respectively). Generally speaking, the calculated MDL and PQL values were within the range found in Standard Methods [9] and EPA Method 8260B[14]. In the case of BTEX, chlorobenzenes, chloroform, toluene, trichloroethylene and tetrachloroethylene, the values were somewhat higher. Details on the method performance and corresponding values from the literature are given in supplementary material S1. Based on the calculation procedure explained in 2.6, different MDLs were obtained for different matrices. The MDL values in CW were within the range of 0.027-0.32 µg/L, while for GW the range was somewhat narrower (0.026-0.18 µg/L), and for most of the compounds the MDLs were lower (except vinyl chloride, 1,2-DCE and 1,1,1-TCE where there were no differences between the samples), whereas for benzene it was higher. Repeatabilities for the 0.4 µg/L concentration level (0.2 µg/L for vinyl chloride) were determined as the RSD of 6 consecutive measurements and were within the range of 1.2-13.2%.

It should be noted that the values from table S1 (given for evaluation of method performance) were not the basis for further recovery comparison at concentration levels of 1 and 15µg/L. The recovery values for concentrations of 1 and 15µg/L in different matrices were collected and further evaluated by QSPR. Average values for 12 replicates of each sample were in the range of 83-123% for CW matrix and in the range of 72-108% in GW matrix. Method precision at these concentrations was within the range of RSD 1.0-8.7% for CW matrix and within the range of RSD 2.0-9.4% for GW matrix.

3.2. Comparison of recoveries

Box plots of recovery values for the 12 replicates of each substance at concentration levels of 1 and 15µg/L and both matrices grouped by compound are given in Figure 1and Figure 2. A parallel presentation of all the performed measurements allows comparison of recovery values of individual samples. Boxes themselves and their width indicate the degree of dispersion and skewness in the data. Central tendency (median) is connected by line in order to facilitate comparison of the results obtained in different experimental conditions. In addition, the quartiles are given as an indication of the dispersion of the recovery values. More precisely, the bottom and the top of the box are the lower (Q_1) and upper (Q_3) quartiles, respectively. The ends of the whiskers represent 1.5 interquartile ranges (IQR=Q_3-Q_1). The values below the lower whiskers (Q_1-1.5 IQR) and above the upper whiskers (Q_3+1.5 IQR) are outliers. Consequently, extreme outliers are values beyond 3 IQR from Q_1 and Q_3. The asymmetrical position of the median in the boxes in Figures 1 and 2 is an indication that the recovery data do not follow a normal distribution (most probably due to the limited number of data points) and hence convenient descriptive statistics, such as mean and standard deviation, are not appropriate.

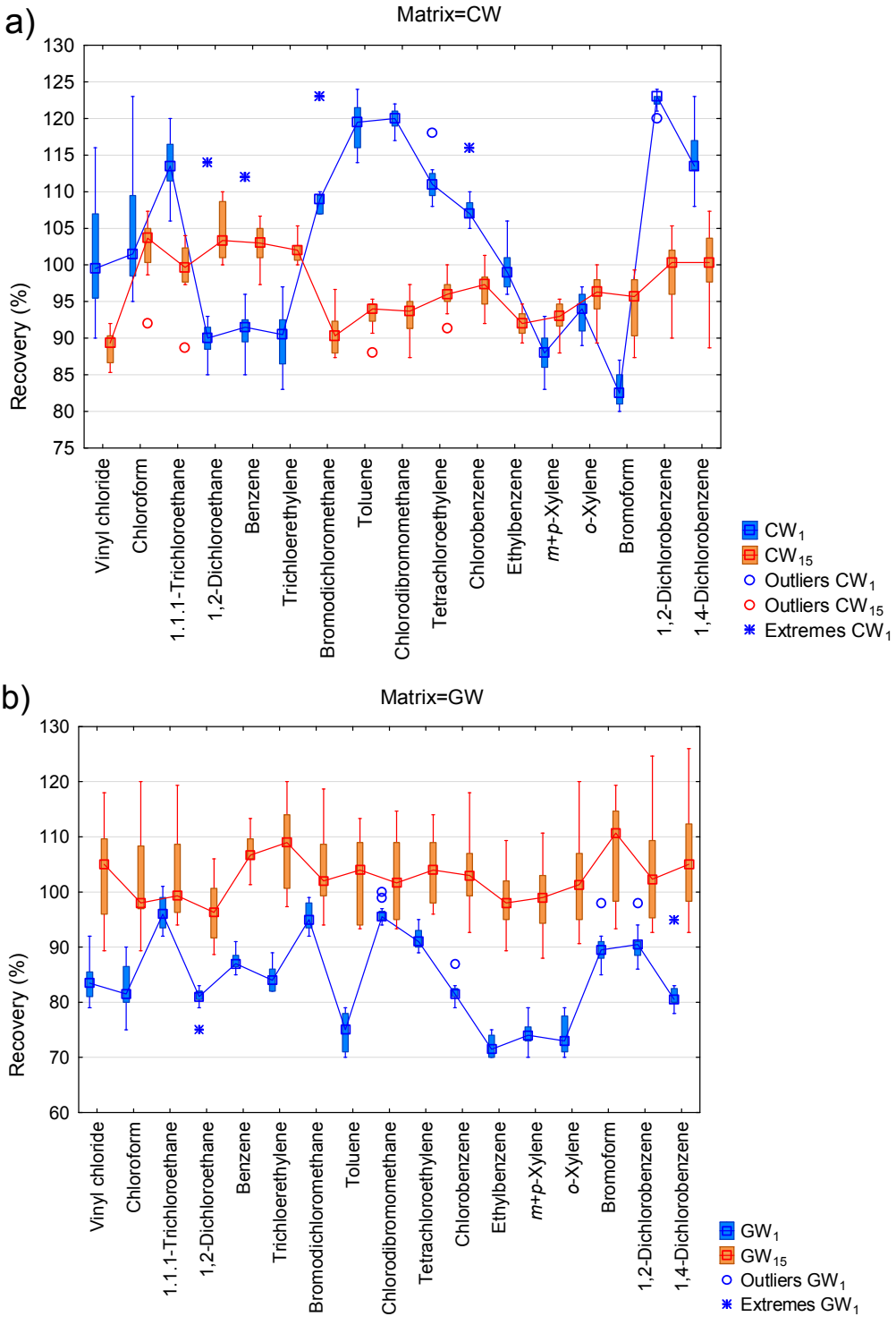

Fig 1. Box plot of recovery data for 12 replicates of 18 VOCs for clean water matrix at concentration levels of 1 µg/L (CW_1) and 15 µg/L (CW_{15}) (a) and for groundwater matrix at concentration levels of 1 µg/L (GW_1) and 15 µg/L (GW_{15}) (b).

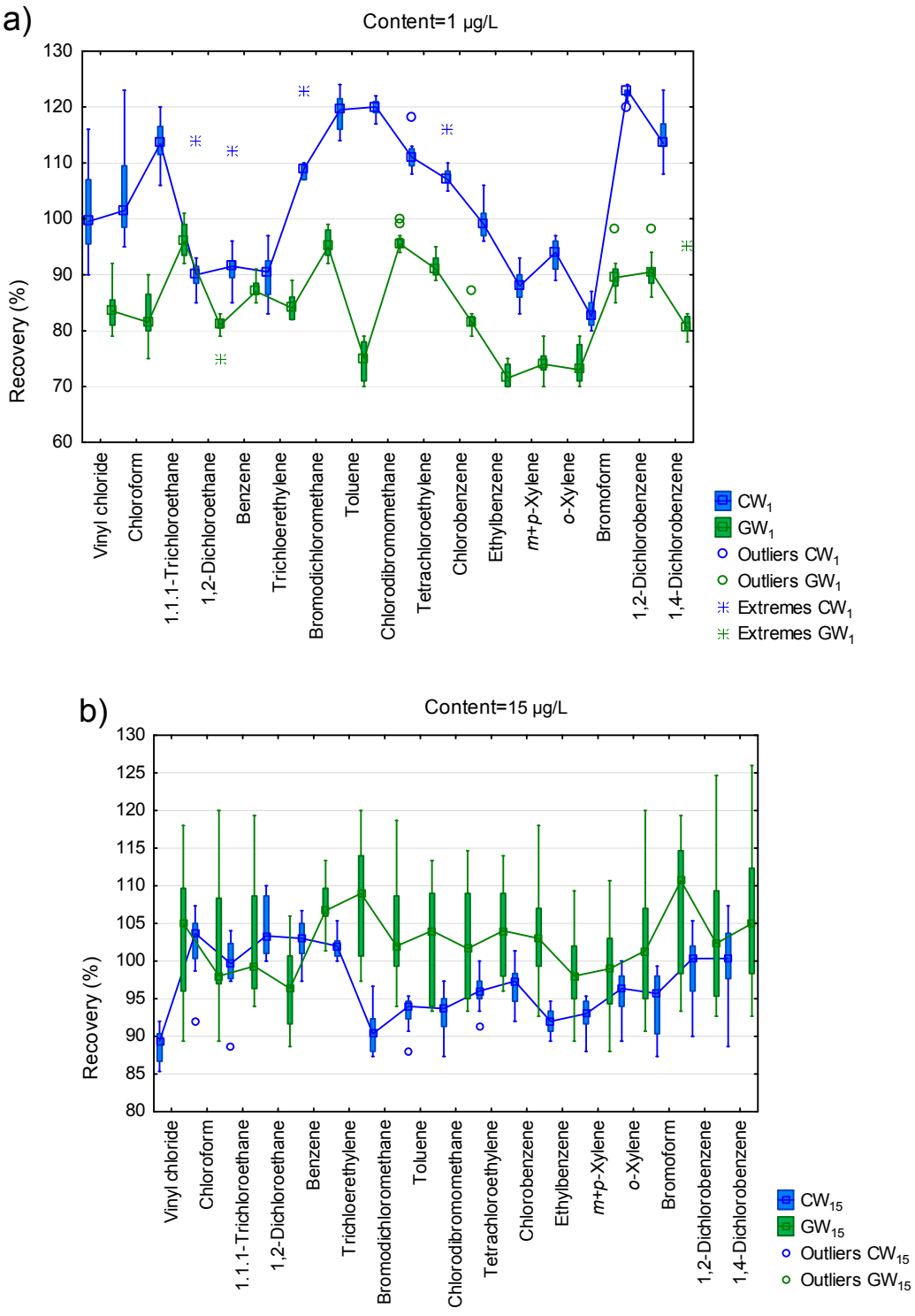

Fig 2. Box plot of recovery data for 12 replicates of 18 VOCs for for both clean and groundwater matrices at the concentration level of 1 µg/L (a) and 15 µg/L (b).

In clean water matrix (Figure 1a) for 1,2-dichloroethane, benzene, trichloroethylene and bromoform, recoveries are higher at the higher concentration level, while for vinyl chloride, 1,1,1-trichloroethane, bromodichloromethane, toluene, chlorodibromomethane, tetrachloroethylene, chlorobenzene, ethylbenzene, 1,2-dichlorobenzene and 1,4-dichlorobenzene, they are lower at the higher concentration. In ground water matrix (Figure 1b), all the compounds have higher recoveries at the higher concentration level.

Comparison of recovery values at the low concentration level for different matrices (Figure 2a) showed that the significant differences were not observed for benzene, trichloroethylene and bromoform, while for all the other compounds the values were higher in CW than in GW. In contrast, at the higher concentration level (Figure 2b), recoveries were higher in GW.

In all further consideration the medians were used, since they are not influenced by extreme outliers in the data set, which is not the case with arithmetic mean values.

HCA was performed in order to visualize similarities (clusters formation) and differences (linkage distances between clusters) of the recovery values for the two matrices and two concentration levels. The resulting dendrogram is presented in Figure 3. The y-axis represents the corresponding linkage distances between the two objects or clusters which are merged. The dendrogram reveals two distinct clusters: cluster A with recoveries related to GW_1 and cluster B that included recoveries for CW_1 together with one sub-cluster with recoveries observed in matrices spiked with higher concentration levels (GW_{15} and CW_{15}). One can hypothesize that the recovery values for GW_1 might be influenced by other factors than in the other three cases since they are grouped in two separate clusters- A and B. Furthermore, the difference between the two matrices at the higher concentration level is lower than at the lower concentration level of 1 µg/L.

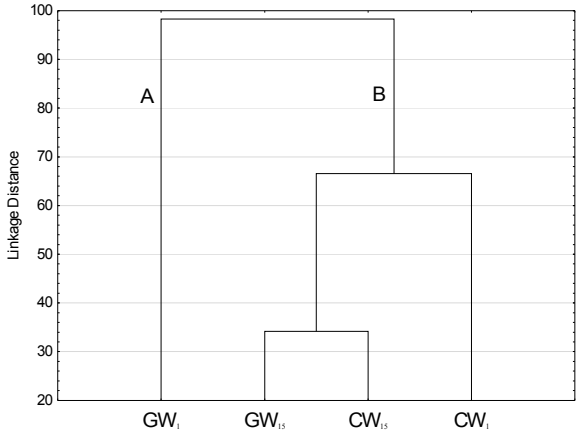

Fig 3. Dendrogram of two water matrices at two concentration levels (GW_1- groundwater matrix with the 1µg/L concentration level, GW_{15}- groundwater matrix with the 15µg/L concentration level, CW_1- clean water matrix with the 1µg/L concentration level, CW_{15}- clean water matrix with the 15µg/L concentration level) in the space of 18 VOC compounds obtained by the Ward linkage method using Euclidean distance

3.2.1. Influence of chemical-specific molecular parameters on recovery

The differences evident in the recovery values obtained for the two matrices as well as the two concentration levels may result from certain molecular characteristics that define the general behaviour of the compounds during analysis, assuming that all the other analytical conditions are held constant and defined by the method. These different intrinsic properties of molecules are usually expressed by various molecular descriptors. Their selection is governed by general knowledge of the features that influence the transfer from water to air and hence have a decisive influence on analytical recoveries while all the other analytical conditions are kept constant. For example, transfer from water to air might be influenced by interactions between analytes and water constituents. Our assumption is that water DOC is not extracted, nor purged nor injected into the GC during P&T GC/MS analysis, since it is known from the literature that in natural groundwater a negligible portion of the DOC is usually volatile. This *a priori* excludes the kind of influence by sorptive interaction of DOC with the relevant instrument surfaces such that the only possible interactions are those of solutes and constituents in the

water phase. Correlations were therefore used as an exploratory tool to indirectly assess if the selected molecular properties might influence analytical recoveries under controlled analytical conditions.

The influence of chemical molecular descriptors on recovery values was investigated for different matrices, as well as for the different concentration levels for each matrix. Initially, the relationship between the differences in recovery for the different concentrations in clean water (CW_{1-15}), for the different concentrations in groundwater (GW_{1-15}), for the lower concentration in the two different matrices $(CW\text{-}GW)_1$ and for the higher concentration in the two different matrices $(GW\text{-}CW)_{15}$ for VOCs and the selected chemical-specific parameters was examined. The intention was to find out which parameter describes the observed differences with the best fit. According to Pearson's correlation (Table 2), linear correlations for the observed differences in recovery and molecular descriptors appear in a wide range, but not higher than 0.68. According to Bevington [28] the limiting value for linear-correlation coefficients for an acceptable correlation of the 18 compounds at a probability level of 0.05 is a minimum 0.468. Correlations exceeding this limit are bolded in Table 2 and are significant (high enough for acceptable linear correlation). Compared to other molecular descriptors, polarity number (Pol) and hydrophilic factor (Hy) had a significant linear correlation coefficient for the difference observed at the two concentration levels in GW matrix. Pol is significant for the difference observed between CW and GW matrix at the lower concentration level (bold in Table 2). At the higher concentration level, differences between the CW and GW matrix showed significant linear correlation with air diffusivity ($D_{i,a}$).

Table 2. Significance of molecular specific parameters for recovery differences between two matrices and two concentration levels expressed by Pearson's correlation ($p < 0.05$, N=18).

Recovery difference	S (mg/L)	log Kow	$D_{i,a}$ (cm^2/s)	$D_{i,w}$ (cm^2/s)	MW	Pol	Hy	molar refractivity	HLC (atm-m^3/mol)
CW_{1-15}	0.2251	-0.1768	-0.2485	0.1754	-0.1490	-0.0672	-0.0763	-0.1065	-0.1059
GW_{1-15}	-0.4063	0.3954	0.3499	-0.2330	-0.4185	**0.5713**	**-0.6843**	0.3802	-0.0030
$(CW\text{-}GW)_1$	-0.3489	0.4364	0.1945	-0.1774	-0.2437	**0.4831**	-0.3691	0.3545	-0.0209
$(GW\text{-}CW)_{15}$	0.3603	-0.0589	**-0.4974**	0.3589	-0.2752	-0.0532	-0.0053	-0.0408	0.2821

S- water solubility, mg/L; HLC- Henry's law constant, atm-m^3/mol; K_{ow}- octanol-water partition coefficient; $D_{i,a}$- air and $D_{i,w}$- water diffusivity at 25°C taken from USEPA,[17] MW- molecular mass; Pol- polarity number; Hy- hydrophilic factor

Subsequently, an attempt was made to correlate not the differences in recoveries for certain samples, but the recovery values themselves and more than one molecular descriptor by MLR. The goal was not to predict the recovery values but to investigate which among the selected molecular descriptors might influence them and under what conditions.

Statistical parameters for the models calculated by stepwise regression on standardized data are shown in Table 3. The developed models were internally validated. The cross-validated regression coefficient (Q^2) is defined as:

$$Q^2 = 1-\Sigma(Y_{pred} - Y_{exp})^2/\Sigma (Y_{exp} - Y_{mean})^2 \quad (1)$$

where Y_{pred}, Y_{exp} and Y_{mean} are the predicted, experimental, and mean values of the target property (recovery).

Q^2 higher than 0.5 is a necessary condition for the model to have predictive power, but still does not automatically imply high predictability.[29] Additional important cross–validation parameters accounting for a good estimate of the real predictive error of the model are PRESS (predictive residual error sum of squares), SSY (sum of squares of deviation of the dependent variable values from their mean) and PRESS/SSY that should be smaller than 0.4.[30] In Table 3, Model 1 shows the best statistics, although it is a borderline case for the PRESS/SSY ratio. Model 3 failed because it does not fulfil the statistical criteria ($Q^2 > 0.5$). The performance of model 4 cannot be accepted as it is over-parameterized (the ratio n/descriptors is >5).

Table 3 . Internal validation statistics[1] for obtained QSPR models

	Molecular descriptor included in model	Multiple - R	Q^2	Adjusted - R^2	F	p	PRESS	PRESS/SSY	Model No.
GW_1	MW, Pol	0.7589	0.5760	0.5194	10.1869	0.0016	473.4190	0.424042	1
	$D_{i,a}$, Pol	0.7484	0.5601	0.5014	9.5481	0.0021	491.1586	0.439931	2

GW_{15}	S, Pol	0.5721	0.3274	0.2377	3.6500	0.0511	174.1626	0.672648	3
CW_{15}	S, HLC, $D_{i,a}$, MW, Pol	0.8893	0.7909	0.7038	9.0776	0.0009	74.8705	0.289164	4

[1] R-correlation coefficients ,R^2_{adj} - adjusted square of the correlation coefficient ,p- significance level (set to $p \leq 0.05$), and Fisher test for significance of the equation (F-test). The acceptance level for the individual independent variable was set to 95% significance level.

In Figure 4 the best constructed stepwise regression model describing recovery for the tested compounds in GW_1 (model 1) is presented.

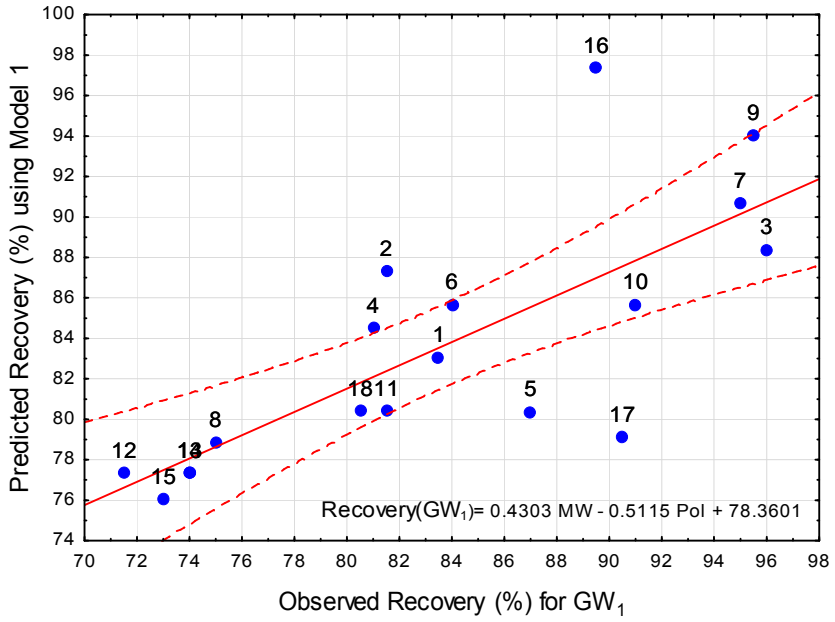

Fig 4. Observed vs. predicted values of Recovery according to QSPR Model 1 from Table 3 (1 – vinyl chloride; 2 – chloroform; 3 – 1,1,1-trichloroethane; 4 - 1,2-dichloroethane; 5 – benzene; 6 – toluene; 7 – bromodichloromethane; 8 – tetrachloroethylene; 9– chlorodibromomethane; 10 – trichloerethylene; 11 – chlorobenzene; 12 – ethylbenzene; 13 - *m*-xylene; 14 - *p*-xylene; 15 - *o*-Xylene; 16 – bromoform; 17 - 1,2-dichlorobenzene; 18 - 1,4-dichlorobenzene) with indicated 95% confidence interval band to the regression line.

Recovery values for compounds 2 (chloroform), 5 (benzene), 16 (bromoform) and 17, (1,2-dichlorbenzene))are not well described by the model proposed (see Figure 3). Sorting the values for each descriptor in increasing order one could find that these substances may be grouped only by HLC (all of them belong to the subgroup which has values from 5.4 10^{-4} to 5.6 10^{-3} atm-m^3/mol). However, six more substances belong to that group (o-xylene, chlorobenzene, 1,4-dichlorobenzene, bromodichloromethane, 1,2-dichloroethane and chlorodibromomethane, numbered 15, 11, 18, 7, 4 and 9 respectively) but fit well with the proposed model. The rest of the compounds have higher HLC values up to 2.7 10^{-2} atm-m^3/mol.

MLR can be a good method for data analysis in cases when the descriptors are few in number, not significantly collinear and if the equation has a good fit. However, if the number of descriptors is too large (in our case more than 3 for 18 substances, like in model 4) MLR becomes inappropriate. Therefore we applied PLS regression. The difference between MLR and PLS is that PLS fits the model simultaneously for all descriptors, while MLR fits descriptors separately.[31] On the other hand, comparison between MLR and PLS models is more abstract which makes it difficult to understand and interpret. PLS regression is an extension of the multiple linear regression model. The main purpose is to build a linear model transforming the original variables into the new orthogonal variables, maximizing the description of a covariance between X and Y[26] .

It was only possible to construct a PLS model for GW_1, which is in accordance with findings related to results presented in Figure 3 .Statistical data for the PLS model are given in Table 4. 63.33% of the sum of squares of the dependent variables is explained by two PLS components. Figure 5 shows the observed vs the predicted

recovery values according to the PLS regression model for the tested compounds in GW_1. Compared to Figure 4, Figure 5 does not seem very much improved particularly for compounds 5 (benzene), 9 (chlorodibromomethane), 16 (bromoform) and 17 (1,2-dichlorobenzene). Improvement was achieved for compounds 4 (1,2-dichloroethane) and 10 (trichloerethylene).

Table 4. Statistical data for PLS model related to GW_1

	R^2X(Cumul.)	Eigenvalues	R^2Y(Cumul.)	Q^2(Cumul.)	Significance	Iteration
1	0.4254	3.5865	0.5517	0.4080	S	1
2	0.6522	1.6407	0.6333	0.1449	NS	1

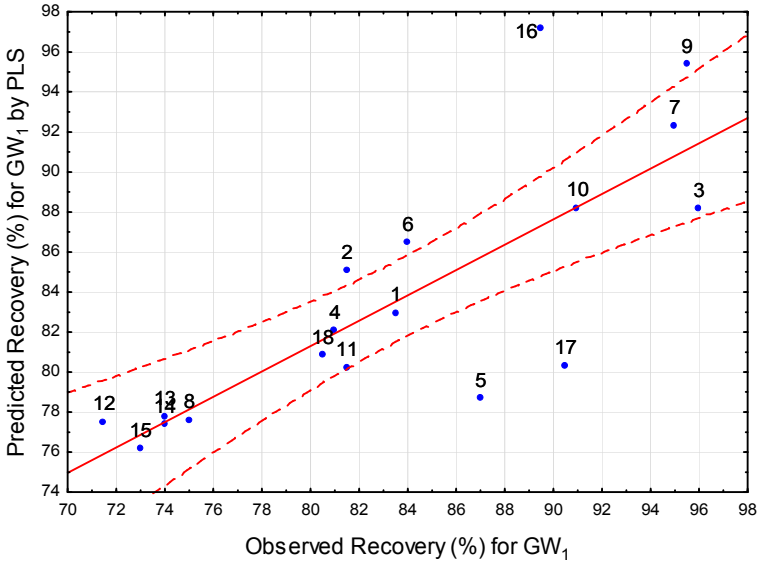

Fig 5. Observed vs. predicted Recovery values for PLS model obtained for groundwater at concentration level of 1µg/L (GW1) (1 – vinyl chloride; 2 – chloroform; 3 – 1,1,1-trichloroethane; 4 - 1,2-dichloroethane; 5 – benzene; 6 – toluene; 7 – bromodichloromethane; 8 – tetrachloroethylene; 9– chlorodibromomethane; 10 – trichloerethylene; 11 – chlorobenzene; 12 – ethylbenzene; 13 - *m*-xylene; 14 - *p*-xylene; 15 - *o*-xylene; 16 – bromoform; 17 - 1,2-dichlorobenzene; 18 - 1,4-dichlorobenzene) with indicated 95% confidence interval band to the regression line

In order to find out the significance of molecular specific parameters in the PLS model the variable importance for the projection (VIP) is given in Fig 6. This allows quick identification of the explanatory variables (descriptors) that contribute most to the model. The variables with higher VIP scores (e.g. power) are considered more relevant.

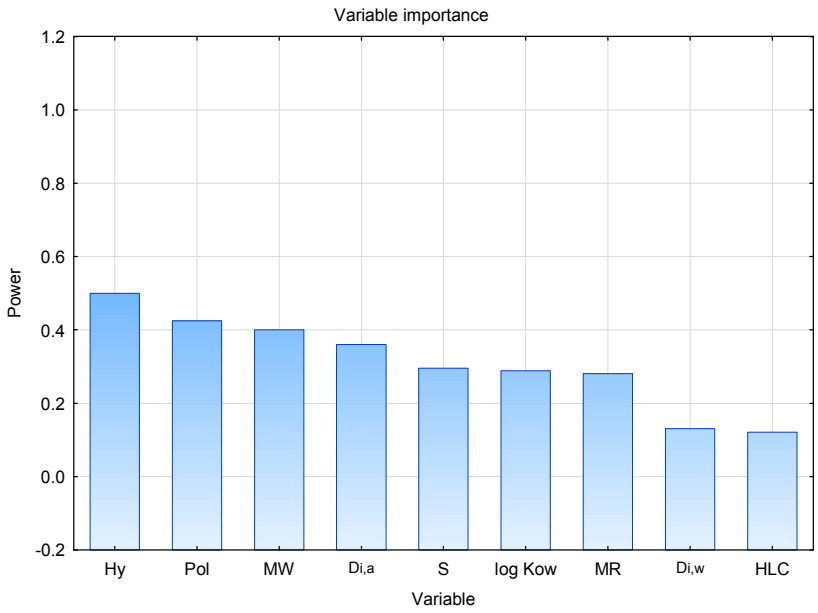

Fig 6. The VIP scores for molecular descriptors used in the PLS calculation for water matrix GW1

Thus the results obtained by PLS calculation show that the hydrophilic factor has the highest VIP score followed by polarity, molecular weight and Di,a which were already indicated by the MLR.

Based on these results one can conclude that recovery values can only be correlated with some of the selected molecular properties in the case of groundwater which contains DOC, analyte at a low concentration level (1µg/L) and salts (GW_1). A possible explanation might be related to the potential for physical or chemical interaction of analytes and water constituents to the extent that could influence the recovery in an indirect way. In the case of CW there is most probably no possibility for such interactions to occur so it is only logical that the selected molecular properties have no influence on the recovery under the analytical conditions applied. As for the groundwater matrix at higher concentrations of analytes (GW_{15}) and the same DOC and salt concentrations as the low concentration level of spiked VOCs (GW_1), one possible reason for the absence of correlation with the selected descriptors and recoveries might be a limited capacity for relevant interactions to alter the transfer of analyte into the gas phase. Thus the influence might become negligible in comparison to GW_1. Testing this hypothesis would require further work, at more concentration levels and in different natural matrices with varying amounts of DOC amount and other characteristics, which is beyond the scope of this current work.

4. Conclusions

Statistical analysis of the dataset showed that the recovery values for VOCs have different behaviours depending on the matrix type and concentration level. Comparison of the recovery values at the low concentration level (1µg/L) showed that for the majority of compounds the recovery values are higher in CW than in GW matrix. In contrast, at the higher concentration level (15µg/L), recovery values are higher in GW matrix.

Using QSPR analysis, polarity number and hydrophilic factor were found to have an acceptable linear correlation with differences between the recovery values at two concentration levels in GW matrix. Polarity was significant in the distinction of recovery differences for CW and GW matrix at the low concentration level, while air diffusivity had an acceptable value in the distinction of recovery differences for the two matrices at the higher concentration.

Correlation between the molecular properties and the recovery values by MLR resulted in a QSPR model for the groundwater matrix at the low VOC concentration. The QSPR includes molecular weight and substance polarity as molecular parameters. This result was further supported by PLS, although with somewhat different results for potentially important molecular parameters based on VIP scores: hydrophilic factor > polarity> molecular weight. No significant correlation was found in either CW matrix or in GW matrix at the higher concentration

level. This implies that theese most significant properties of VOC molecules might be relevant for the recovery only at low VOC concentration levels and only if the matrix contains other water constituents (i.e. DOC, salts).

Acknowledgements

This research is funded by Ministry of Education, Science and Technological Development of Republic of Serbia (grant ON 172028). The authors would like to thank prof. Zagorka Lozanov-Crvenković for valuable comments and constructive suggestions during the preparation of this manuscript.

References

1. S. Lacorte, L. Olivella, M. Rosell, M. Figueras, A. Ginebreda and D. Barcelo, *Chromatographia*, 2002, **56**, 739.
2. E. Martinez, S. Lacorte, I. Llobet, P. Vianna and D. Barcelo, *J. Chromatogr A*., 2002, **959,** 181.
3. A. Ikem, *J. Food Compos. Anal.*, 2010, **23,** 70.
4. N. Barco-Bonilla, P. Plaza- Bolaños, J.L. Fernandez-Moreno, R. Romero-Gonzales, A.G. Frenich and J.L. Martinez Vidal, *Anal. Bioanal. Chem*., 2011, **400,** 3537.
5. WFD, Water Framework Directive 2000/60/EC, OJ L 327/1, 2000.
6. M. Rosell, S. Lacorte, A. Ginebreda and D. Barcelo, *J. Chromatogr. A*, 2003, **995**,171.
7. Nikolau A.D., Golfinopoulos S.K., Kostopoulou M.N., Kolokythas G.A., lekkas T.D. (2002) *Water Research* 36, 2883-2890.
8. P. Lepom, B. Brown, G. Hanke, R. Loos, P. Quevauviller and J. Wollgast, *J. Chromatogr. A*, 2009, **1216,** 302.
9. AWWA-APHA-WEF Standard Methods for the Examination of Water and Wastewater, 20th ed., American Public Health Association/American Water Works Association/Water Environment Federation, Washington, DC, 1998.Volatile organic compounds (6200 B), pp 6-25
10. USEPA Method 524.2 Measurement of pureable organic compounds in water by capillary column gas chromatography/mass spectrometry, Revision 4, 1992.
11. SRPS ISO 8245:2007 Guidelines for determination of total organic carbon (TOC) and dissolved organic carbon (DOC) in water, 2007.
12. USEPA Method 8000B Determinative chromatographic separations, Revision 2, 1996b.
13. USEPA Method 5030B Purge-and-trap for aqueous samples, Revision 2, 1996a.
14. USEPA EPA Method 8260B Volatile organic compounds by gas chromatography/mass spectrometry (GC/MS), Revision 2, 1996c.
15. A.A. Glase, D.L. Foerst, G.D. McKee, S.A. Quave and W.L. Budde WL, *Environ. Sci. Tech.*, 1981, **15,** 1426.
16. Directive 2008/105/EC, OJ L348/84, 24.12.2008
17. USEPA Supplemental guidance for developing soil screening levels for superfund sites. United States Environmental Protection Agency Solid Waste and Emergency Response, 2002. http://www.epa.gov/superfund/health/conmedia/soil/pdfs/part_5.pdf. Accessed 15 June 2012.
18. Talete srl DRAGON (Software for Molecular Descriptor Calculation) Version 6.0 - 2011 - http://www.talete.mi.it, 2011.
19. D. Cox, *J. Royal Stat. Society: Series B*, 1972, **34**, 187.
20. S. Bergante, G. Facciotto and G. Minotta, *Central Europe Journal of Biology*, 2010, **5,** 522.
21. D. Massart, B.Vandeginste, L. Buydens, S. DeJong, P. Lewi and J. Weyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A,* Elsevier: Amsterdam, 1997.
22. H. Martens and T. Naes, *Multivariate calibration by data compression*. In: Williams PC, Norris K (eds) Near-infrared Technology in Agricultural and Food Industries, American Association of Cereal Chemists: St. Paul, Minnesota, 1987.
23. H. Martens and T. Naes, *Multivariate Calibration*. John Wiley & Sons: New York, 1989.
24. H. Wold, *Soft modeling: the basic design and some extensions*. In: Jöreskog K, Wold H (eds) Systems Under Indirect Observation: Causality, Structure, Prediction, North-Holland: Amsterdam, 1981.
25. S. Wold, H. Martens and H. Wold, *The Multivariate Calibration Problem in Chemistry Solved by the PLS Method.* Lecture Notes in Mathematics. Springer Verlag: Heidelberg, 1983.
26. P. Geladi, B. R. Kowalski, *Analytica Chimica Acta*, Volume 185, 1986, Pages 1-17
27. StatSoft Inc. STATISTICA (data analysis software system), version 10. www.statsoft.com, 2011, Tulsa, OK USA
28. P. R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw Hill Book Co., New York, 1969.

29. A. Golbraikh and A. Tropsha, J. Mol. Graph. Model, 2002, **20**, 269.
30. S. Wold, *Mol. Inform*., 1991, **10**, 191.
31. T. Dijkstra, *J. Econometrics*, 1983, **22**, 67.