Chemometrics and Intelligent Laboratory Systems 149 (2015) 123–131

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Expert system for monitoring the tributyltin content in inland water samples

M. Daszykowski [a], M. Korzen [b], B. Krakowska [a], K. Fabianczyk [c],*

[a] Institute of Chemistry, University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland
[b] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, 49 Zolnierska Street, 71-210 Szczecin, Poland
[c] Polcargo International, Poboznego Street, 70-900 Szczecin, Poland

ARTICLE INFO

Article history:
Received 1 May 2015
Received in revised form 20 October 2015
Accepted 22 October 2015
Available online 30 October 2015

Keywords:
Expert system
Discriminant models
Machine learning
Chemometrics

ABSTRACT

In this study we discuss an attempt to build an expert system that can support decision making by analytical chemists regarding the presence of tributyltin (TBT) in inland Polish water samples in detail. It is possible to conclude with at least a 0.93 probability that a sample is free of TBT using the expert system that was constructed (if a sample is analyzed in accordance with the European norm PN-EN ISO 17353:2006). This idea, which is based on the efficient use of the information that is stored in a chromatographic database, can easily be extended to monitor other priority substances in water samples. Our on-going research, which is focused on octylphenols in water samples, has provided very encouraging results and additionally supports this hypothesis. The proposed framework can also be attractive to other testing laboratories that have a similar scope of expertise and follow the same analytical protocols. Moreover, as a natural consequence of our research further efforts should lead to the development of a ready-to-use product that would offer testing laboratories validated chromatographic libraries along with the expert system(s) with the possibility of upgrading them with respect to an increasing pool of analyzed samples. Such a solution when implemented in a testing laboratory environment may have a wide economic impact on its further functioning and increase throughput efficiency, especially in a case in which monitoring priority substances in water is a major concern.

1. Introduction

Optimizing the costs of demanding analytical procedures while preserving their specificity and reliability is a very challenging task and is a core part of developing intelligent laboratory systems/procedures and management. In general, analytical procedures are time consuming and highly experienced analytical personnel and sophisticated equipment are required.

For decades, different researchers have been seeking appropriate analytical procedures that would enable the identification of substances in very complex analytical matrices, which would then lead to reliable results from an analytical point of view. From this perspective, environmental water samples are a typical example of complex and demanding samples that require separation techniques for their examination. The identification of a large group of chemical constituents in complex mixtures, including environmental water samples, can be done using chromatographic methods and is one of the major challenges in any testing laboratory. Apart from its high level of sophistication, chromatographic analysis is extremely susceptible to external factors, which may result in the shifting of peaks and/or their overlapping and thus complicate the ability to draw conclusions from chromatograms. That is why continuous efforts are undertaken to make chromatographic analysis as reliable and effective as possible. From a practical point of view, two concepts for processing chromatographic signals/data can be identified. The first one relies on the active use of the hardware configurations and options that are available due to the technological advancement of modern chromatographic devices. The second one, which follows a chemometric philosophy, involves the extraction of useful information from complex chromatographic data as a result of the development and implementation of advanced algorithms and their application during different steps of analytical workflow, see e.g., [1–4].

In this study, we contrasted different theoretical approaches with the aim of developing an efficient expert system that is based on machine learning and is proposed to support the detection of the contaminant tributyltin (TBT) in Polish marine and fresh water ecosystems. Its major advantage relies on incorporating the knowledge of experts and a diverse representation of environmental water samples.

Tributyltin, which is a biocide agent, has been extensively used as an ingredient of antifouling paint and is designed to prevent or slow down the growth of organisms on coated with a painted surface. Because it is extremely efficient as a biocide agent, TBT has been used for 40 years mainly in the shipping industry. Initially, it was considered to be environmentally safe, but it was proven that when TBT is released into the

* Corresponding author. Tel.: +48 91 434 33 68; fax: +48 91 488 20 36.
E-mail address: k.fabianczyk@polcargo.pl (K. Fabianczyk).

http://dx.doi.org/10.1016/j.chemolab.2015.10.008
0169-7439/© 2015 Published by Elsevier B.V.

environment, it exceeds acute and chronic toxic levels. For this reason, different international regulations have been issued to effectively prohibit the further use of TBT and thus to reduce the progression of water contamination with TBT and its degradation products [6]. Unfortunately, toxic effects can still be observed because TBT has a relatively long half-life that depends on its sources, degradation products and their accumulation in sediments, environmental conditions and other factors [7,8]. This is why the level of TBT, including the TBT that is found in different water bodies, is the subject of strict on-going monitoring.

In this study, environmental water samples were collected in the course of the large-scale environmental diagnostic monitoring of inland waters in Poland that was requested by the Chief Inspectorate of Environmental Protection, which was focused i.a. on the detection and quantification of TBT. The sampling campaign took place between 2011 and 2013. TBT and other organotin compounds can be quantified in water samples using the diverse set of available analytical techniques [5]. Among them there are, e.g., the GC–MS method, the headspace solid-phase microextraction-gas chromatography-pulsed flame-photometric detection [6] and the fluorescence technique followed by chemometric modeling using the second-order calibration, which helps in the quantification of TBT at parts-per-trillion levels [7]. In 1403 water samples, the TBT content (quantified as tributyltin cation) was determined using the GC-MS technique according to the European norm PN-EN ISO 17353:2006. The chromatographic fingerprints of water samples that were obtained, despite the rigorous protocols that were applied during chromatographic analysis, reflect all of the major sources of variability. They are rather noisy and contain a substantial baseline component. Moreover, from sample to sample chromatographic peaks are shifted and the overlapping of peaks is frequently observed. On the other hand, real GC-MS fingerprints of complex mixtures pose a real challenge for chemometric and machine learning approaches. As a result of a considerable analytical effort, a relatively large collection of diverse water samples was analyzed. The set of 1403 chromatograms that was obtained offered a unique opportunity to verify their usefulness as a database that would support the construction of an expert system to facilitate in the detection of TBT contaminants in water.

2. Materials and methods

2.1. Sample collection and chromatographic analysis

The sampling plan, sampling frequency, as well as the protocols for sample collection regarding the determination of certain priority substances in the course of the diagnostic monitoring of Polish inland waters in 2011 and 2013 were prepared by the Chief Inspectorate of Environmental Protection in Poland. The following procedure was applied to the analysis of the TBT content in the water samples. Water samples (each 1000 ml) were collected in dark glass flasks. Samples were stored at 4 °C (also during transport). Further sample treatment and analysis was carried out in a specialized laboratory that has up-to-date certificates of accreditation issued by the Polish Center of Accreditation within the scope that includes the analysis of TBT done in accordance with the European norm PN-EN ISO 17353:2006.

Chromatographic fingerprints were registered using a gas chromatographic system (Agilent Technologies 7890A) with a single quadruple mass detector (Agilent Technologies 5975C) with electron ionization. Mixture components (1 μL of a water extract) were resolved using helium as the gas carrier and a DB-5 column (30 m × 250 μm × 0.25 μm). The temperature of the inlet was set to 250 °C and during the separation a temperature gradient was applied (from 60 °C up to 170 °C every minute by 12 °C, then 170 °C up to 280 °C every minute by 20 °C) in the oven. The following ions were monitored (scanning rate at 4.53 cycles/sec): the target ion of tributyltin (TBT) $m/z = 291$; the qualifier ion tri-n-propyltin (TprT) $m/z = 289$; the target ion $m/z = 249$ and the qualifier ion $m/z = 247$ (temperatures of the transfer line, MS source and MS quadruple were set to 300 °C, 230 °C and 150 °C, respectively and the energy of electrons was set to 69.0 eV).

2.2. Preprocessing of chromatographic fingerprints

A typical workflow for preprocessing chromatographic fingerprints that is carried out prior to multivariate data modeling usually includes the improvement of the signal-to-noise ratio. It consists of baseline removal, noise elimination and the alignment of chromatographic signals [1]. Correction of heteroscedasticity usually is done by transforming data, e.g., logarithm or power transformations [8]. In our study, baseline was corrected using the penalized least squares asymmetric least squares approach, PALS [9]. The alignment of signals was carried out using correlation optimized warping, COW [10]. More details about these preprocessing methods can be found in cited references.

2.3. Discriminant models

The aim of discriminant models is to assign a sample to one of the existing groups of samples based on its characteristics (e.g., a chromatographic fingerprint). In this study, we focused on the discrimination between two groups of water samples with and without TBT. This task essentially corresponds to the issue of identification — the presence or absence of TBT in water samples (confirmed for model set samples by the presence of characteristic mass spectra).

For the pilot discrimination of the groups of samples that were studied, a classic chemometric linear discriminant approach was used — partial least squares–discriminant analysis, PLS-DA. In addition, in order to confirm the presence of TBT in the environmental water samples, the following machine learning methods were used: logistic regression (LR) with the L^1 regularization, linear support vector machines (LK-SVM), ensemble methods including AdaBoost (AB) and random forest (RF), K-nearest neighbors (KNN) and the Parzen classifier (PC).

In the following sections, a brief characteristic of each machine learning technique will be provided.

2.3.1. Partial least squares-discriminant analysis

Partial least squares-discriminant analysis, PLS-DA, is a variant of the classic partial least squares regression model, in which a categorical dependent variable that indicates to which group a sample belongs, is modeled [11]. Any discrimination between the groups of samples is achieved by the construction of a linear separation hyper plane in the space of a few latent variables, which are also called latent or PLS factors. They are mutually orthogonal and maximize the covariance between the set of latent variables and the response variable (in a simple PLS variant with one response, PLS-1). Owing to the construction of orthogonal latent factors, the construction of the PLS model is not hampered by the presence of collinear explanatory variables. In fact, PLS-DA and linear discriminant analysis share the same objective — minimizing within group variance and maximizing between group variance [12].

2.3.2. Logistic regression and family of linear classifiers

PLS-DA is one example from a large group of linear discriminant methods. Others, which are commonly used family of methods within this group, are methods that are based on penalized logistic regression. Similar to PLS-DA, in logistic regression the likelihood function with a penalizing term is used instead of the least squares cost function.

$$\mathrm{Q}(\mathbf{X},\ \mathbf{y};\mathbf{b}) = \mathrm{Loglikelihood}(\mathbf{X},\mathbf{y}) + \mathrm{P}(\mathbf{b}) \tag{1}$$

Depending on the type of regularization approach that is selected, a considerably different behavior of a learning machine can be obtained. For instance, the L^2 regularization (or ridge, $\mathrm{P}(\mathbf{b}) = ||\mathbf{b}||_2$) leads to a grouping effect of the correlated variables, the L^1 regularization (or