Journal of Chromatography A, 1218 (2011) 5688–5693

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Are the results of customary methods for analyzing dioxin and dioxin-like compound congener profiles court-proof?

Steffen Uhlig*, Stefanie Eichler

quo data, Gesellschaft für Qualitätsmanagement und Statistik mbH, Kaitzer Str. 135, 01187 Dresden, Germany

ARTICLE INFO

Article history:
Received 12 April 2011
Received in revised form 21 June 2011
Accepted 23 June 2011
Available online 30 June 2011

Keywords:
Dioxin and dioxin-like compounds
Congener profiles
Clustering algorithm
Principle component analysis
Court-proof
Validity

ABSTRACT

The congener profile of samples contaminated with dioxin and dioxin-like compounds allows identifying sources of contamination. This article studies the statistical methods of congener profile analysis reported in the literature with respect to the reliability of obtained results. The performance of customary analysis methods regarding raw data transformation and applied TEF (toxic equivalency factor) values is discussed. In particular, the method of principal component analysis and *k*-means cluster is taken as an example and examined in detail. Reasons for occurring inconsistencies such as the dependence of results on raw data transformation and the disregard of measurement uncertainty are described, and it is shown that they also explain inconsistencies in other methods of cluster analysis such as hierarchical cluster analysis and neural networks. It is concluded that these methods cannot be employed to reach court-proof decisions, i.e. decisions which meet court evidentiary standards. An alternative approach to analyzing congener profiles based on mathematical statistics is briefly presented, allowing reliable, court-proof decisions.

1. Introduction

Dioxin and dioxin-like compounds (DLC) represent a severe risk to human health even at very low concentrations due to their high ability to accumulate in the organism and generate persistent effects such as carcinogenicity or teratogenicity. The majority of human DLC intake is from food of animal origin, e.g. meat, dairy products and fish. Contamination of food is the result of contaminated feed and, in particular for wildlife fish, from contaminated waters. Hence, food quality assurance necessitates the control of waters [1–3], soil [4], food [5–7] and feed [6,8].

DLC contaminated samples contain a mixture of dioxin and dioxin-like congeners, exhibiting a broad range of toxicity and bioaccumulation. To assess the risk to human health originating from a contaminated sample, the sum of toxic equivalents (TEQ) of 17 hazardous dioxin and furan congeners is determined [9], as first proposed by Eadon et al. [10]. Analyzed contaminated samples may vary in the concentration ratios of the congeners, i.e. they exhibit different congener profiles. The congener profile allows to identify sources of contamination and to draw conclusions about human exposure. However, the customary methods to analyze congener profiles based on multivariate statistics and neural networks, such as principal component analysis (PCA) and diverse clustering methods can only offer hints which cannot be verified due to lack of proof or evidence. Therefore the question arises whether results from clustering methods allow valid and court-proof decisions. In order to provide valid, court-proof conclusions, the basis of a method to analyze congener profiles should incorporate analytical uncertainty prior to the application of statistical inference.

2. Results and discussion

2.1. The effect of raw data transformation on the results of congener profile analysis

In the literature congener profiles are compared using PCA [11,12] and clustering algorithms such as hierarchical cluster analysis [4], *k*-means cluster or neural networks (e.g. Kohonen maps) [2,3,13,14]. All these methods include an initial transformation of raw data, i.e. the concentration of each congener from each sample is transformed before analysis. For example, the original congener concentration can be transformed into the ratio of the congener TEQ to the overall TEQ of the sample. Upon closer examination, it becomes clear that commonly used transformations are not uniformly used throughout the literature (e.g. compare [14] and [13]) and can only be justified phenomenologically.

In the present study we observe widely varying results from clustering analysis, which are due to different data transformations and to variations in the applied toxic equivalency factors (TEFs). This feature is examined in detail for PCA and *k*-means cluster.

Fig. 1 and Table 1 summarize the raw data of 16 samples from data of a study on sediment of a river. Using methods based on

* Corresponding author. Tel.: +49 351 40288670; fax: +49 351 402886719.
E-mail address: uhlig@quodata.de (S. Uhlig).

0021-9673/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2011.06.080

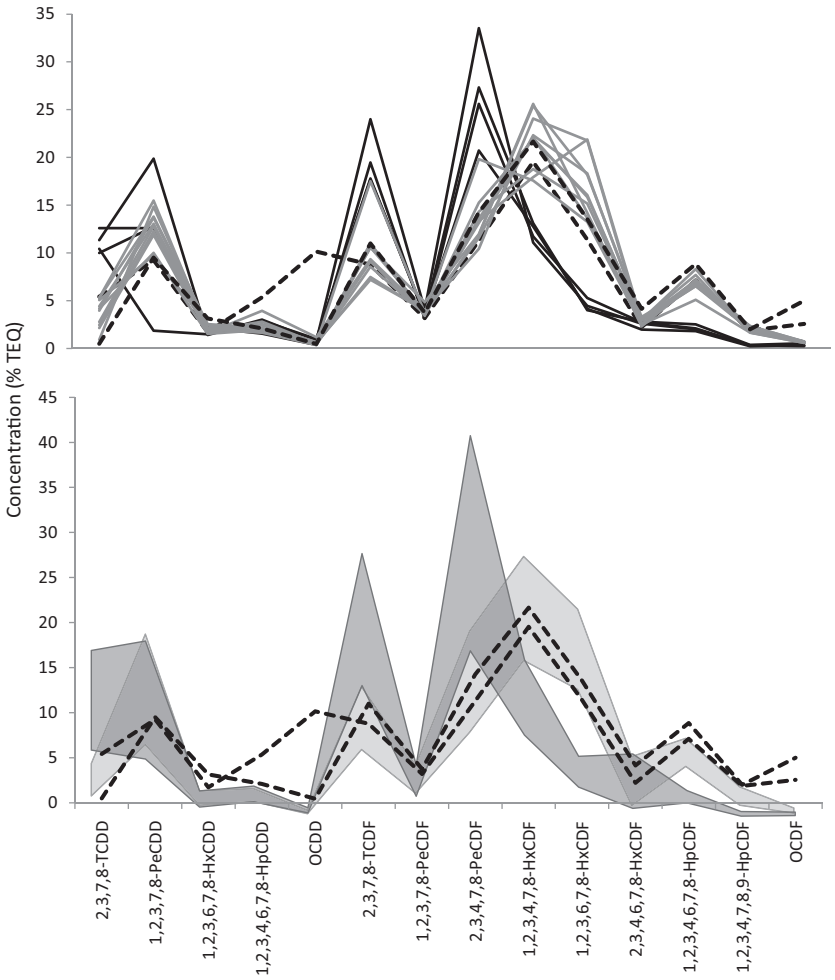

Fig. 1. *Upper panel*: Congener profiles (data transformed into relative TEQ applying the actual WHO TEF from 2005 [23]) of 16 samples from a study on sediment of a river. The statistically significant allocation of each sample to one of the two clusters is indicated by different lines (black solid, gray solid), the two outlier samples are depicted as black dashed lines. *Lower panel*: For each cluster the congener profile (relative TEQ) and the respective uncertainty range is shown, outlier samples are depicted as black dashed lines.

statistical inference we were able to allocate the 16 samples into two sub-populations regarding their congener profile and to identify two outlier samples (combined in a third sub-population). The obtained sample allocation is quite reasonable as it separates samples of downstream and upstream parts of the river into different sub-populations. The two outlier samples can be explained by the differing method of sampling (out of an helicopter) for the sample 324/I and by the geographical position of sample 678/II at the estuary mouth. By calculating the uncertainty range of the respective congener profile for each subpopulation and by determining the homogeneity within each sub-population, we are able to ensure a statistically significant sample allocation (see [15]). This approach was conducted by applying the web service hosted at [16].

In contrast to our congener profile analysis based on statistical inference, Table 2 presents the results of congener profile analysis by means of PCA and clustering using *k*-means cluster exemplary illustrated for one transformation in Fig. 2. PCA and *k*-means clustering were performed using the statistical computing language and environment R (version 2.12.2). PCA itself is not a clustering method but it provides plots (e.g. loading plots) which serve as a survey of the data and can possibly exhibit cluster structures. The allocation of samples into sub-populations is achieved by *k*-means cluster (number of clusters = 3) after reducing the dimensions of the dataset by PCA and the respective sample allocation is indicated in Table 2. Different results of these methods are obtained through application of different raw data transformations and of different TEFs. Moreover, it should be noted, that repeated application of the *k*-means cluster algorithm does not necessarily result in equal sample allocation to sub-populations due to the initial random choice of the cluster centers as a first step of the algorithm. The method of congener profile analysis based on statistical inference allows statistically verified allocation of samples to sub-populations and identification of outliers, whereas the congener profile analysis based on PCA and *k*-means cluster involves a distance determination disregarding any underlying uncertainty. The decision about the assignment of two samples, e.g. two points in a loading-plot obtained by PCA (similar to Fig. 2), to the same sub-population is based on the distance of the corresponding points in the plot. If their distance is small, it means that their transformed raw data is very similar and in consequence they are regarded as belonging to the same sub-population. Hence it immediately follows that the reason for different appearances of plots obtained by different raw data transformations is the fact that actually the transformed data is compared and not the raw data. However, for evaluation of one plot the same "ruler" is used to measure the distances between all points, implying that transformed data of all samples exhibit equal absolute errors. This assumption does not hold true for real data irrespective of the applied transformation, a fact reflected by the shape of the uncertainty ranges in Fig. 1 which