Data Mining:
Evaluation of the German Dioxin Data Bank

Gerlinde Knetsch[*], Stefanie Schulz[**], Maximilian Swerev[**]

[*] Federal Environmental Agency,
Bismarckplatz 1, D-14191 Berlin, Germany, E-mail: gerlinde.knetsch@uba.de
[**] Bavarian Institute for Waste Research,
Am Mittleren Moos 46, D- 86167 Augsburg, Germany; E-mail: bifa@augsburg.baynet.de

Introduction

Polychlorinated dibenzo-p-dioxins (PCDD) and polychlorinated dibenzofurans (PCDF), often collectively known as dioxins, are the topics of intense ecological discussion. The compounds of this class of organic pollutants are toxic, found in all environmental compartments, persistent and tend to bioaccumulate being fat soluble. Dioxins formed many years ago may still contribute to contemporary exposure. The accident at Seveso in 1976 and the detection in fly ash from municipal waste incinerators in 1977 lead to a widespread attention and attracted a great deal of research on PCDD and PCDF.

So, over the last 20 years, the German Federation and the Federation States have initiated extensive measuring programs about the environmental pollution by dioxins. The results of these measuring programs are compiled in the German Dioxin Data Bank of the Federal Environmental Agency.

This German Dioxin Data Bank contains about 350 data sets about dioxin emission, atmospheric deposition and concentrations in soils as well as 50 data sets relating to biota and waste samples. Moreover, there exist about 10000 elder data sets relating to air, food and human milk. These data sets are very inhomogenous in contrast to the demand for an uniform documentation and evaluation.

In our research project we develop a strategy for an assessment of the available data sets concerning the contamination of the environment by PCDD/PCDF. This strategy can be the basis for a general assessment of the environmental situation.

For this purpose suitable data mining methods for the extraction of information about the complex data sets were selected. The characteristics and chemical behaviour of the single congeners of PCDD/PCDF are taken into account.

The aims of the research project are:

- Documentation of changes of environmental situation (e.g. after taking measures to reduce the emission of PCDD/PCDF)
- Demonstration of relationships between cause and effect and of non-linear relationships particularly with regard to sensible samples and target organisms
- Demonstration of pathways of PCDD/PCDF released into the environment and of their transfer behaviour, particularly of the pathways

  air → fly ash → human
  air → animal food → human food
  air → soil → animal/human food → human
- Orientation for further necessary measures to reduce the release of persistent organic pollutants (especially PCDD/PCDF) into the environment.

Methods

The research project is divided into the following three steps:

1. Stocktaking and description of quality of the available data sets

The first aim is the development of a scheme for classification of the data sets. For this purpose the existing data sets are classified in different groups in order to get information about:

- Data sets of high, medium and low analytical quality (for all matrices)
- Number of data sets of particular matrices like soil, sediment, water, air, fly ash, vegetation, food, products, waste...
- classification of the data sets by periods, e.g. by the year of sampling.

For the description of the quality of the available data sets the following factors are taken into consideration:

- Full characterisation of samples related to sampling (e.g. corn size, season, date of air samples)
- Congener specific analysis; use of different columns (polar/unpolar)
- Description of quality assurance and quality control measures
- Analytical requirements according to different regulations by law or other regulations
- Analytical laboratory with accreditation or certification
- Categorize in different period scales (year/decade/century)
- Geographical characterisation
- Definition of an qualifier for polluted/unpolluted (background) samples and check, if present limits or guidelines are exceeded.

The result of this evaluation will lead to a matrix with rapid and extensive overview of existing data material concerning to number, age, range or precision of analytical results. Missing data can be easily recognized. The main focus of the different measuring programs can be identified and differentiated in groups relating to data quality, preferred matrices, analysis periods, geographical distribution etc. Furthermore, the best data sets can be identified as basis for the comparison of the chosen data mining methods.

Based on this evaluation scheme the specification of future observation programs can be done.

2. Analysis of data sets using established methods

In the second step we compare the results of all methods which up to now have been successfully applied to the evaluation of PCDD/PCDF data sets.

The following statistical methods are used:

- ♦ Similarity Index (1, 2)
- ♦ Correlation Analysis (3, 4)
- ♦ Cluster Analysis (Hierarchical Cluster Analysis and K-Means Analysis) (4, 5, 6)
- ♦ Multidimensional Scaling (4)
- ♦ Principal Component and Factor Analysis (4, 7, 8, 9, 10).

First of all relatively homogenous data sets will be analysed using this methods. In a second step other data sets will be added.

For the comparison of the different mathematical-statistical methods the following questions are considered:

- How can we handle not detected congeners?
- What kind of normalization is reasonable for the different methods?
- Do the methods account for the analytical uncertainty depending on the degree of chlorination?

For a recommendation of suitable methods the easiness of use and the necessary mathematical know-how are also taken into consideration.

3. Analysis of data sets using innovative methods

In the last step we will test the use of innovative methods, first regarding their applicability and secondly applying to selected data sets from the German dioxin data bank. Up to our knowledge these methods have not been applied to PCDD/PCDF.

The following methods will be tested:

- ♦ Fuzzy Cluster Analysis (Fuzzy c Means, GK analysis according to Gustafson and Kessel, GG analysis according to Gath and Geva, Grid Cluster) (11)
- ♦ Neural Networks (12).

Conclusion

The presented research project serves for classification of different and inhomogenous data sets and for the development of necessary measures for further uniform documentation and evaluation. This evaluation can be the basis for a general assessment of environmental situation for all persistent organic pollutants.

Furthermore, the comparison of the different data mining methods allows to choose suitable ways for extract valuable information from complex and inhomogenous data sets concerning different patterns, time trends and distribution of persistent organic pollutants.

Acknowledgement

This research project is supported by the Federal Environmental Agency (Project-Number 216 06 137).

References

1. Buchert, H., Ballschmiter, K.: Mustererkennung von polychlorierten Biphenylen (PCB) in Umweltproben. *Fresenius Z. Anal. Chem.* **1985**, 320, 709.
2. Schreitmüller, J., Vigneron, M., Bacher, R., Ballschmiter, K.: Pattern analysis of polychlorinated biphenyls (PCB) in marine air of the Atlantic Ocean. *Intern. J. Environ. Anal. Chem.* **1994**, 57, 33.
3. Sadler, J.C., Campbell, I.: A simple method for dioxin congener pattern comparison. *Organohalogen Compounds* **1994**, 19, 61.
4. Backhaus, K., Erichson, B., Pinke, W., Weiber, R.: *Multivariate Analysemethoden.* Springer Verlag, Berlin, 8. Auflage **1996**.
5. Hagenmaier, H., Lindig, C., She, J.: Correlation of Environmental Occurence of Polychlorinated Dibenzo-p-Dioxins and Dibenzofurans with Possible Sources. *Organohalogen Compounds* **1993**, 12, 271.
6. Bauer, K.M., Stanley, J.S., Remmers, J., Breen, J.J., Schwemberger, J., Schultz, B., Kang, H.K.: Pattern recognition analysis of VA/EPA PCDD and PCDF data. *Organohalogen Compounds* **1990**, 2, 91.
7. Stalling, D.L., Norstrom, R.J., Smith, L.M., Simon, M.; *Patterns of PCDD, PCDF, and PCB contamination in Great Lake fish and birds and their characterization by principal component analysis*, Chemosphere **1985**, 14, 627.
8. Brakstad, F.: A comprehensive pollution survey of polychlorinated dibenzo-p-dioxins and dibenzofurans ba means of principal component analysis and partial least squares regression. *Chemosphere* **1992**, 25, 1611.
9. She, J., Hagenmaier, H.: Use of principal component analysis in the source identification of PCDDs and PCDFs. *Organohalogen Compounds* **1993**, 12, 187.
10. Tysklind, M., Fängmark, I., Marklund, S., Lindskog, A., Thaning, L., Rappe, C.: Atmospheric transport and transformation of polychlorinated dibenzo-p-dioxins and dibenzofurans. *Environ. Sci. Technol.* **1993**, 27, 2190.
11. Höppner, F., Klawonn, F., Kruse, R.: Fuzzy-Clusteranalyse. Vieweg Verlag, Wiesbaden **1997**.
12. Nauck, D., Klawonn, F., Kruse, R.: Neuronale Netze und Fuzzy-Systeme. Verlag Vieweg, Wiesbaden, 2. Auflage **1996**.