DATA MINING OF WATER QUALITY DATA BY CHEMOMETRICAL METHODS

B.G.M. VANDEGINSTE, Unilever Research Laboratorium Vlaardingen, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands.

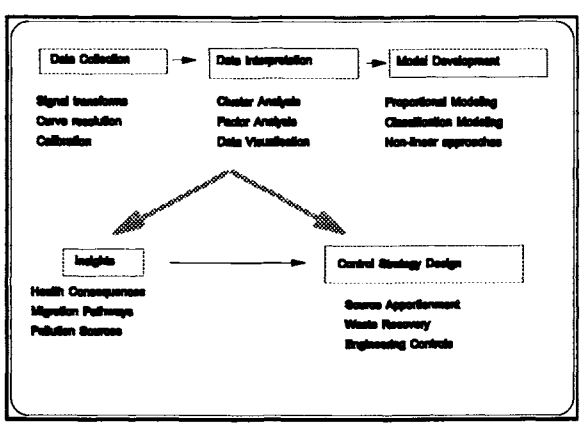
1. Setting the scene

In the field of water quality monitoring, increasingly complex sets of data are measured, which require advanced data processing technologies in order to extract the information hidden in the data.
In increasing order of complexity we may think of time dependent concentration levels of several constituents at one location or at several locations. This respectively results in a table or matrix of data (time * concentration) or a cube of data (time * position * concentration). A complication specifically with environmental data is that some constituents are not monitored with the same frequency, nor at the same time.

Complexity may even be larger if, instead of concentrations, we consider the output of e.g. a sensor array. If such a sensor array is positioned at several locations in a process, the operator receives consecutive frames of data of the dimension (time * number of sensors * number of locations). A quick judgement of the data in terms of control actions is then required. If these data are stored in a database, the mining of such historical data may be of paramount importance for process modelling.

Obviously, visualisation and modelling techniques should assist the operator in interpreting and modelling the data (Fig. 1). Methods which found a widespread application in environmetrics are pattern recognition, class modelling and factor analysis [1].

Fig. 1 Issues in environmental assessment (adapted from [1])



In some instances the environmental chemist may use complex measurement technologies, such as hyphenated techniques (GC-MS, LC-MS, LC-DAD, etc.). These instruments produce two-way data tables per sample, for instance a table of spectra (rows) and chromatograms (columns), which contains information on the concentration levels of several constituents. Advanced chemometric techniques are available for the analysis of such data tables. Specifically when the chromatographic resolution is insufficient, it is possible to retrieve the pure spectra and pure elution profiles by the application of special factor analysis methods, such as Orthogonal Projection Analysis (OPA) [2]. Tables of hyphenated data obtained at several location points and/or at several time slots form a stack of tables, so-called multiway data tables. Advanced chemometrical techniques are available to decompose these tables in pure spectra, pure chromatographic profiles and concentration profiles. This is a relatively new application area, which still has to demonstrate its capabilities to the environmental chemist.

Above methods heavily rely on the principle of linear additivity of signals or linear separability of clusters or categories. Particularly challenging are the non-linear systems which can be visualised and classified by non-linear PCA and QDA. Methods from natural computing techniques, such as Neural Networks have already proven their usefulness in modelling non-linear systems.

A step further is to relate measurements (up to complex multiway data sets) or arrays of results to a certain property. The traditional approach is to apply multivariate calibration (PCR, PLS or MLR) to model these relationships. Recently Neural Networks have been introduced to model complex relationships [3].

In water quality monitoring the operator relies on the quality of the measurements. This calls for intelligent measurement devices which contain internal checking procedures for system suitability tests and internal consistency.

Summarizing, the massiveness and complexity of data sets obtained in water quality monitoring require the application of advanced chemometric and natural computing techniques, preferentially combined in a system of data base management tools, and tools to process heuristic knowledge (if..then rules). Such systems are referred to as data mining for which software is becoming readily available.

2. Current situation

Table 1 summarizes the information which is wanted by the environmental researcher together with the technologies which are generally available to obtain that information. Many of those methods yet do not belong to the toolkit of the end-user and are still in the stage of research and development (techniques indicated with an *)

The wanted information falls into five categories: (1) the analysis of historical (space-time) data (2) methods for improving/cleaning the data (3) techniques to derive concentrations from

complex data structures (sensors, hyphenated methods) (4) process monitoring and control (5) thermodynamical/environmental modelling.

Table 1: Data mining techniques relevant to the environmental researcher

Information from data	Data mining techniques
Analysis of historical data	
Visualisation (landscapes)	Principal Components Analysis
Classification (supervised)	K-Nearest Neighbour SIMCA, ALLOC, UNEQ Artificial NN*
Classification (unsupervised)	Clustering, ALLOC Kohonen networks*
Source Allocation/apportionment	Target Transformation FA Artificial NN* PLS
Time/space dependent data arrays	Evolving FA* Three-way FA
Improvement/cleaning of data	
Noise removal	Wavelets*
Outlier detection	Preprocessing (synchronisation of time arrays*) Robust techniques*
Monitoring/control	
Time dependent data arrays (forecasting and modelling)	Multivariate* time series analysis (ARIMA)
Internal checking of monitoring devices	Multivariate SPC* Kalman filtering* Artificial NN*
Thermodynamic modelling	
Receptor and chemical balance studies	Non-linear modelling Partial Least Squares Simulated Annealing* Artificial NN*
Transform of data to concentration	
Hyphenated techniques Sensors	Orthogonal Projection Analysis Curve resolution, three-way FA

The table demonstrates the importance and impact of statistical/chemometrical techniques in the field of environmental chemistry. Because of the specific problems associated with data from eco-systems, highly specialised experts are active in this field, which is known as environmetrics.

Some of the mature chemometric methods are readily available to the end-user, mainly aimed at the visualisation of the data by PCA or by another displaying technique. Software is commercially available in dedicated packages and within general purpose statistical packages such as SAS. However the massiveness of the historical data files may cause problems when using some of the older packages. When exploiting the much larger computing power of workstations, integrated software systems, which are intended to mine the data, so-called data-mining, are becoming available. They offer a choice of modern displaying techniques, factor analysis, artificial neural network, modelling and pattern recognition.

A particular problem with environmental data is the occurrence of outliers, missing data, data below the detection limit etc. This largely influences the ability to extract the information from the data and calls for caution when using 'black box' procedures canned in data mining systems.

Many of the more modern chemometric methods are still in the hands of the specialist and are not common knowledge to the end-user. We refer here to source allocation by target transformation factor analysis and three-way analysis (PARAFAC) of cubes of data.
(space*concentration*time). A nice overview of these specialised techniques can be found in a special issue of Chemometrics and Intelligent laboratory systems [9] which reports on the third international conference on environmetrics and chemometrics held in Las Vegas (1995).

In chemometrics literature not much work has been reported on the design of control/monitoring systems, which implies time series analysis, the implementation of forecast and prediction systems, for example by Kalman filtering, and the design of optimal sampling schemes. 'water' quality is inherently an array of quality parameters (constituents and physical parameters), which necessitates the design of systems based on multivariate statistical process control, unless one combines all quality parameters in a single figure, which is the usually approach.
Literature becomes even more scarce when besides objective measurements one wants to include subjective information of the operator in the control system.

3. New developments

The complexity and massiveness of environmental data is a real challenge to the chemometrician and environmetrician. From a chemometrics point of view, there are several areas which need to be further explored. Some of them have been summarized below, without pretending to give a full list :

- Exploration and introduction of data mining technologies to the end-user
- Exploration of three-way techniques and evolving FA for the factor analysis of three way data

- Development and implementation of robust multivariate SPC techniques for water quality monitoring.
- Development and implementation of robust multivariate data analysis (robust for outliers, missing values)
- Development of techniques to analyse non-linear systems and non-linear relationships with non-linear models (Artificial neural nets, etc.)
- Development of techniques for the preprocessing of multivariate data, selection and compression of variables into features, for example by the application of Wavelets.
- Development of measurement devices with internal checks for consistency, automatic fault warning and auto-calibration.

REFERENCES

1. Wenning and G.A. Erickson "Interpretation and analysis of complex environmental data using chemometric methods", Trac 13 446-457, 1994.

2. Cuesta Sanchez, B.G.M. Vandeginste, T.M. Hancewicz and D.L. Massart "Resolution of complex liquid chromatography-Fourier transform Infrared spectroscopic data", Anal. Chem., 69 1477-1484, 1997

3. Hopke, X.-H. Song "The chemical mass balance as a multivariate calibration problem" Chemom. Intell. Lab. Syst. 37 5-14, 1997.

4. Third international conference on environmetrics and chemometrics, Chemom. Intell. Lab. Syst. 37 1-27, 1997.