Analytica Chimica Acta 794 (2013) 20–28

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

# Extension and application of multivariate curve resolution-alternating least squares to four-way quadrilinear data-obtained in the investigation of pollution patterns on Yamuna River, India—A case study

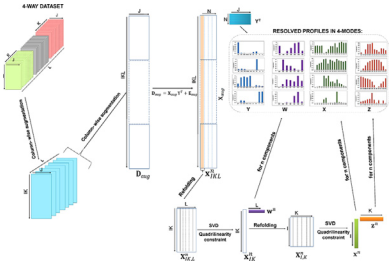
Amrita Malik*, Roma Tauler

*Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), Jordi Girona 18-26, 08034 Barcelona, Catalunya, Spain*

## HIGHLIGHTS

- This study presents a new development of the MCR-ALS method introducing a quadrilinear constraint.
- A long term four-way environmental dataset is presented as a case of study.
- MCR-ALS resolved dominant pollution patterns for the Yamuna River (India) during the years (1999–2005).
- The MCR-ALS proves to be a powerful tool to summarize and resolve large multi-dimensional datasets.

## GRAPHICAL ABSTRACT



## ARTICLE INFO

*Article history:*
Received 22 April 2013
Received in revised form 10 July 2013
Accepted 18 July 2013
Available online 29 July 2013

*Keywords:*
Chemometrics
Multivariate Curve Resolution-Alternating Least Squares
PARAFAC
River pollution
Environmental monitoring
Factor analysis

## ABSTRACT

This study focuses on the development and extension of Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) to the analysis of four-way datasets. The proposed extension of the MCR-ALS method with non-negativity and the newly developed quadrilinear constraints can be exploited to summarize and manage huge multidimensional datasets and resolve their four way component profiles. In this study, its application is demonstrated by analyzing a four-way data set obtained in a long term environmental monitoring study (15 sampling sites × 9 variables × 12 months × 7 years) belonging to the Yamuna River, one of the most polluted rivers of India and the largest tributary of the Ganges river. MCR-ALS resolved pollution profiles described appropriately the major observed changes on pH, organic pollution, bacteriological pollution and temperature, along with their spatial and temporal distribution patterns for the studied stretch of Yamuna River. Results obtained by MCR-ALS have also been compared with those obtained by another multi-way method, PARAFAC. The methodology used in this study is completely general and it can be applied to other multi-way datasets.

## 1. Introduction

With the advent of industrialization and increasing population, the requirements have increased for higher quality environment stressing for the advance research in all areas concerned to human health and its environment, and continuous monitoring of available natural and man-made resources for their efficient management and control. Modern sophisticated and sensitive analytical technologies and instrumentation are providing huge amounts of experimental observations in all research areas related to life sciences (e.g. chemical kinetic studies, pharmaceuticals, medicinal and clinical studies, proteomics, metabolomics, genomics etc.) which, in nature, are multivariate, multi-set and multi-way (three

* Corresponding author. Tel.: +34 93 400 61 40.
*E-mail addresses:* ambqam@cid.csic.es, amritamalik.b@gmail.com (A. Malik), Roma.Tauler@idaea.csic.es, rtaqam@cid.csic.es (R. Tauler).

0003-2670/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.aca.2013.07.047

or higher number of ways) data structures. For the natural environmental studies, this kind of datasets are generated, regularly, under the environmental programs run by government authorities, research institutes and Non Governmental Organizations to monitor the condition and quality of natural resources over time and space all over the world. Often, outcome of these studies are concentration information on multiple chemical compounds collected at different sampling periods from different sampling sites arranged in large tables, data matrices, or in more complex multi-way data arrays (three or more directions or modes) [1]. These data sets are frequently rather cumbersome to interpret by their simple direct observation, emphasizing, therefore the need of having appropriate data analysis tools to extract relevant environmental information from these huge monitoring data sets. The processing and interpretation of such multi-way data sets require the development and application of appropriate data analysis tools to extract reliable information about the investigated analytical systems. Multi-way data analysis methods can better summarize these huge environmental monitoring and other research data sets providing a more in-depth interpretation of the relevant information contained in them. Three-way or three-mode data (data arranged in three directions) can be analyzed either using eigenvalue/eigenvecor data decompositions, or by trilinear and non-trilinear alternating least-squares (ALS) [2]. In practice, ALS is considered to be an efficient method for multi-way data decompositions and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) has emerged as a powerful tool to analyze multiple data arrays through matrix-augmentation fulfilling different model complexities [3–6]. The main advantage of MCR-ALS [7], in this context, is that it is easily adapted to data sets of different complexity and structure, bilinear, trilinear or multilinear, providing optimal least squares solutions. MCR-ALS has been used to analyze various datasets-that can be described by a bilinear model-related to many kind of processes and mixtures such as chemical reactions, industrial processes, chromatographic analysis, spectroscopic measurements, environmental data, monitored by diverse multivariate responses [8]. The success and generalized use of MCR-ALS is related to the possibility to work with multi-way and multi-set data structures, i.e., analyzing several data tables simultaneously [4–6,9].

For the analysis of four-way quadrilinear data, the available methodologies are usually based on quadrilinear alternating least-squares and non-quadrilinear latent structure models. The multi-way PARAFAC method [10] and some of its complementary variants such as alternating penalty quadrilinear decomposition [11] and alternating weighted residual constraint quadrilinear decomposition [12] are available for processing data complying with the quadrilinear model condition [13]. This work is focused on the extension of MCR-ALS to handle four-way data under non-negativity and the newly developed quadrilinearity constraint. In its simplest configuration, MCR-ALS is based only in the fulfillment of the bilinear model and in having only one data mode in common. However, if the data has two or more modes in common, multilinear models, like the trilinear or the quadrilinear models can optionally be implemented inside the ALS algorithm as constraints. Using this approach MCR-ALS has already been extended and applied to analyze three-way trilinear data [14]. This study focuses on the development of MCR-ALS with a newly developed quadrilinear constraint for the analysis of four-way datasets. To demonstrate the efficiency of MCR-ALS with the newly developed quadrilinear constraint, a four-way dataset originated through the long term regular monitoring campaign (1999–2005) of the Yamuna River, in India, was used as a case of study to resolve the dominant pollution patterns and their distribution along time and geographical axis. Application of the new quadrilinear MCR-ALS algorithm is shown to be able to resolve specific pollution patterns on temporal and geographical modes. For comparison purpose the dataset was also analyzed with PCA and PARAFAC methods, which are traditionally used for this kind of studies. This study is intended to provide help to understand the use of multi-way methods to analyze the huge datasets collected and recorded in large project or environmental monitoring reports, thus helping the decision making authorities to know what are the main contamination patterns over a particular geographical area and time frame, and to conclude alterative solutions for the health and management of environmental resources.

## 2. Methods

### 2.1. Dataset

The dataset used in this study was obtained from the regular monitoring of the Yamuna River, by Central Pollution Control Board (CPCB), India [15]. Yamuna River is one of the most polluted and largest tributary of the Ganges River, India. The total length of Yamuna River from origin (31°2′12″ N and 78°26′10″ E) to its confluence with Ganges is 1376 km. The Yamuna River covers water demands of rural and urban settlements like Delhi, Mathura, Agra and Allahabad. In turn, this river receives back outfall of a number of drains carrying domestic and industrial wastes rich in organic matter. To check Yamuna River for water quality control and mitigation purposes, CPCB regularly monitors the river at selected sampling sites. For this study, monthly river quality data from seven years (1999 to 2005) has been used. Sampling sites and parameters were selected based on their availability and continuity during the study period. The final dataset consisted of 15 sampling sites (S1–S15), 9 measured variables or parameters (pH, COD, BOD, $NH_4$, TKN, DO, WT (Water Temperature), TC (Total Coliforms), FC (Fecal Coliforms)) monitored every month (12 measurements) for 7 years (1999–2005). The locations of sampling points on Yamuna River are presented in Fig. 1. The total amount of data values simultaneously analyzed are therefore $15 \times 9 \times 12 \times 7 = 11{,}340$ individual values.

### 2.2. Data organization

Environmental datasets can be arranged according to the number of ways or modes they have. For example, a data set obtained in a single monitoring campaign measuring multiple variables or parameters over a set of samples is a two-way data consisting of two ways or modes: samples (first mode) and variables or parameters (second mode) or vice versa, which can be arranged in a two-way data table or data matrix. If, instead the dataset consists of the same variables or parameters measured for different conditions (e.g. at different times) for every sample, then one data matrix is collected per sample and a three-way data set forming a cube is obtained, i.e. three ways or modes are: samples (first mode), variables (second mode) and conditions (third mode). The term ‘ways’ and ‘modes’ are analogous and used interchangeably in this work. The dataset having a three-way or three-mode data cube per sample will give a four-way data structure. This is the case for instance when, as in this work, two types of time conditions (months and years) are measured per sample and variable. As described earlier, the dataset used under this case of study was obtained by the regular monitoring of 9 water quality parameters at 15 different (but fixed) sites of the Yamuna river in every month of the year and for seven consecutive years. Thus, this dataset has four modes: first mode (samples), second mode (variables), third mode (months) and fourth mode (years).

The whole data set was initially arranged into 7 matrices, belonging to each year, having all of them 180 rows (15 sampling sites × 12 months) and 9 columns (variables). These 7 matrices can be arranged into different ways for their simultaneous analysis.