Chemometrics and Intelligent Laboratory Systems 137 (2014) 120–127

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

# Parallel factor analysis and multivariate curve resolution as data fusion tools to supervise a stream

Alejandro G. García-Reiriz

*Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario, S2002LRK, Argentina*

ARTICLE INFO

*Article history:*
Received 20 January 2014
Received in revised form 24 June 2014
Accepted 25 June 2014
Available online 2 July 2014

*Keywords:*
Chemometrics
Multivariate curve resolution
Parallel factor analysis
Environmental monitoring

ABSTRACT

In this work, a new method is proposed to monitor the distribution, evolution and correlation of dissolved organic matter on the superficial water of a stream with respect to physicochemical variables that characterize the basin and season sampling of each campaign. The method is based on measuring fluorescence emission–excitation matrices and some physicochemical parameters of water samples through both time and space. In a first phase, parallel factor analysis (PARAFAC) or multivariate curve resolution with alternating least-squares (MCR-ALS) were applied to extract the information on the relative proportions of each fluorophore on each sample. Then, MCR-ALS was applied again to the entire database, in order to study the spatial and time distribution. This methodology was used to study the behavior of a basin stream that is significantly modified by anthropic activities.

## 1. Introduction

Any global monitoring plan should include measurements of concentrations of various analytes of environmental interest for the system under study and some physicochemical variables that characterize it. Dissolved organic matter (DOM) is a simple parameter, indirectly measured by fluorescence excitation–emission matrices (EEM), which provides valuable information of anthropogenic activities in the watershed, because the DOM has different fluorescent properties depending on its origin [1].

The aim of this paper is to propose a new strategy to combine the information that can be extracted by chemometric methods from fluorescence data matrices with specific measurements of physicochemical variables and/or analytes, in order to study their possible relationships and their distribution in space and time. The final objective is to detect possible contamination sources.

The system under study is the Ludueña stream. It is located in the Santa Fe Province of Argentina, in the Rosario Department. Its basin is about 800 km$^2$. Before its confluence with the Parana River, it flows inside a tube for along 1.5 km. In the higher areas, it has an earthen dam that helps to slow the water runoff during the rainy season, and also contributes to collect water from two channels: the Ibarlucea and the Salvat channels (Fig. 1).

The Ludueña stream watershed is currently in constant modification by human activities. This is because big cities exist in its margins that contribute to sealing large areas of soils; for this reason, its caudal increases dramatically during periods of rainfall. Currently, several private and open neighborhoods are being developed in its vicinity. Also, dense and irregular settlements exist in its margins, generating clandestine channels which provide both stormwater and sewage effluents.

In natural aquatic environments, DOM is composed of a great variety of organic substances, mainly arising from two different origins: 1) autochthonous, stemming from the chemical and biological activity of microorganisms, and 2) allochthonous, due to anthropogenic activities, such as industrial wastewaters or sewage discharges. The first group of compounds comprises some humic-likes substances; their structure and composition allow one to characterize the water quality in which they are dissolved. The second group, in contrast, may be composed of proteic substances, i.e., amino acids arising from dissolved peptides and proteins, when they stem from sewage discharges [2–4]. However, other compounds of allochthonous origin may be present in natural water samples, such as pesticides, hydrocarbons, or the so-called emergent contaminants (human or veterinary pharmaceuticals), etc.

Fluorescence spectroscopy has allowed to characterize DOM in samples of different origins, to monitor the contamination level of polluted areas, and to distinguish anthropically impacted regions from less affected ones [5–7]. Natural waters usually contain a mixture of fluorophors which makes their identification difficult by means of unidimensional fluorescence spectra [1]. An excellent analytical alternative is to measure fluorescence EEMs, which allow one to obtain much richer information related to the presence and type of dissolved fluorophors. EEMs began to be studied in the decade of 1990, with the distinction of humic and non-humic-like compounds in natural waters [1,8,9].

In order to extract the chemical component information from the registered EEM data, several methodologies can be applied. One of the most popular ones is to build a three-way signal array from the recorded

---

*E-mail address:* garciareiriz@iquir-conicet.gov.ar.

http://dx.doi.org/10.1016/j.chemolab.2014.06.016
0169-7439/

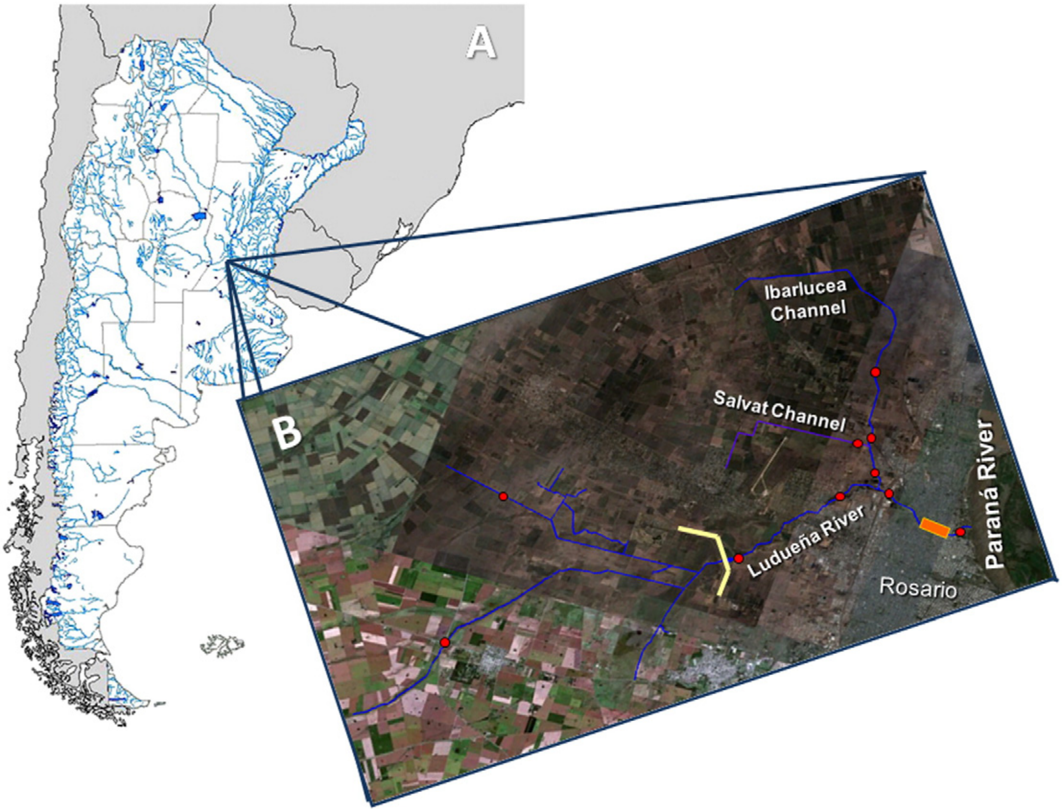

**Fig. 1.** Location of the study area; **A**, in Argentina; and **B**, inside Rosario Department.

EEMs, followed by parallel factor analysis (PARAFAC) [10]. This algorithm allowed to identify humic-like and proteic-like substances in water samples [11], to characterize the DOM present in lakes and soils [12], to identify anthropogenic contaminants and metal traces in waters [13], to study the discharge of effluents into rivers and marine waters associated with ranges of salinity and nutrients [14], to detect fulvic acids and tryptophan-rich proteins in sewage discharges [15], and to classify water samples based only on the content of humic acids [16]. A related algorithm, multivariate curve resolution coupled to alternating least-squares (MCR-ALS) was also employed for similar purposes [17]. Therefore, both methods were applied to process the EEMs and to compare their results. MCR-ALS works similarly to PARAFAC when it is applied with the trilinearity constraint. However, instead of working with a three-way array, an augmented data matrix is built by placing each EEM below the other one (see specific details below).

To study the distribution of each fluorophore obtained by PARAFAC together with other variables in time and space, a new MCR-ALS approach can be applied [18]. This allows to group the variables with similar behavior and provides spatial and time distribution. In this way, it was possible to make an interpretation about their origin and interrelation.

MCR-ALS is a powerful chemometric tool with an increasing application for the analysis of environmental monitoring data sets [19,20]. Other chemometric methods have also been applied to the investigation of environmental data, such as partial least-squares (PLS) [21,22], PARAFAC and Tucker3 models [23]. The use of multivariate factor analysis, such as those proposed in the present work, has also been discussed in several books [24,25]. Additionally, other recent examples exist proposing similar approaches for the resolution and interpretation of major contamination sources of surface waters operating in several river basins over the world [26].

Thus there are three main objectives in this work: 1) the possibility of combining data of different complexity to obtain a better characterization of the system, 2) the investigation of main long-term diffuse contamination sources of organic contaminants in the Ludueña stream basin area, and 3) the estimation of their geographical distribution, in order to contribute to the evaluation of the environmental health of the surface waters of the region under study. To achieve these three goals, multivariate data methods of analysis based on combinations of PARAFAC and MCR-ALS were applied.

In order to get useful environmental information from the data, the application of modern chemometric methods based in multivariate factor analysis tools is proposed [27]. The basic assumption of these methods when they are applied to environmental data tables is that each value of a measured variable in a particular sample is due to the sum of contributions from individual independent sources of different origin. Each of these sources is characterized by a particular chemical composition profile and is distributed among samples in a different way. As a result of the application of chemometric methods, the main point and diffuse sources of contamination in the environment and their origin may be identified and their distribution profiles among samples (geographical, temporal, among environmental compartments) are characterized.

## 2. Material and methods

### 2.1. Equipment

Fluorescence spectral measurements were performed on a fast-scanning Varian Cary Eclipse fluorescence spectrophotometer, equipped with two Czerny–Turner monochromators and a xenon flash lamp, and connected to a PC microcomputer via an IEEE 488 (GPIB) serial interface. Excitation–emission data arrays were recorded in a 10 mm quartz cell, in the following ranges: excitation, 220–481 nm each 3 nm; emission, 280–600 nm each 5 nm. Thus, the size of each data matrix was $88 \times 65 = 5720$ data points. The wavelength scanning speed used was 12,000 nm/min. The detector voltage was fixed at 700 V.