Journal of Geochemical Exploration 144 (2014) 144–153

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Modeling geochemical datasets for source apportionment: Comparison of least square regression and inversion approaches

Gyana Ranjan Tripathy [a,c,*], Anirban Das [b]

[a] AIRIE Program, Department of Geosciences, Colorado State University, Fort Collins, USA
[b] Pandit Deendayal Petroleum University, Gandhinagar, India
[c] CSIR-National Institute of Oceanography, Dona Paula, Goa 403004, India

ARTICLE INFO

Article history:
Received 2 December 2013
Revised 2 March 2014
Accepted 6 March 2014
Available online 18 March 2014

Keywords:
Least square regression
Inverse modeling
Source apportionment
Geochemistry

ABSTRACT

Mathematical modeling of geochemical datasets finds frequent applications in Earth Sciences, particularly in areas of source apportionment and provenance studies. In this work, source apportionment modeling have been considered based on two commonly used methods, the Least Square Regression (LSR) and Inverse Modeling (IM), to determine the contributions of (i) solutes from different sources to *global river* water, and (ii) various rocks to a glacial till. The purpose of this exercise is to compare the results from the two mathematical methods, infer their merits and drawbacks and indicate approaches to enhance their reliability.

The application of the LSR and IM approaches to determine the source contributions to *global river* water using the same *a-priori* end member compositions yielded divergent results; the LSR analysis giving impossibly negative values of Na contribution from one of the sources (evaporites), in contrast to the IM approach which yield reliable estimates of source contributions, and a set of *a-posteriori* source compositions and associated uncertainties. Interestingly, the use of the *a-posteriori* composition derived from the IM approach in the LSR analysis as an input for end-member composition gave source contributions that were consistent with those derived from IM. Calculations based on the IM show that 46 ± 8% of Na in *global river* is derived from silicate weathering, consistent with some of the earlier reported estimates.

In case of the glacial till, the source contributions based on both the approaches were similar, however even in this case better agreement between the two approaches is obtained when the *a-posteriori* composition data of end members derived from the IM is used as input in the LSR model. These comparisons demonstrate that the IM is better suited for source apportionment studies among the two models, as it requires only rough estimates of end member composition, unlike the LSR that needs source composition to be better constrained. In addition, the IM also provides uncertainties in the source contributions and best estimates of their composition.

1. Introduction

Computational simulation methods have become important tools over the last few decades in providing quantitative information for many Earth science problems (Zhao et al., 2009 and the references therein). For instance, the computational simulation method has been used to solve not only a wide range of ore forming problems within the upper crust of the Earth (Gow et al., 2002; Hobbs et al., 2000; Ju et al., 2011; Lin et al., 2006; Liu et al., 2005, 2008, 2011; Ord et al., 2002; Schaubs and Zhao, 2002; Schmidt Mumm et al., 2010; Sorjonen-Ward and Zhang, 2002; Zhang et al., 2003, 2008), but also a broad range of other types of Earth science problems (Lin et al., 2003, 2008, 2009; Xing and Makinouchi, 2008; Yan et al., 2003; Zhang et al., 2011; Zhao et al., 2008a, 2010). Since three basic models, *i.e.* geological, mathematical and numerical simulation models, are involved in the computational simulation method, mathematical modeling plays an indispensable role in the computational simulation of Earth science problems (Zhao, 2009; Zhao et al., 2008b and references therein).

Geochemical approaches, as an import part of the computational geoscience discipline (Zhao et al., 2009 and references therein), have found extensive applications in the field of Earth sciences to study and infer about various geological processes. Chemical, mineralogical and isotopic compositions of geological samples hold clues to the sources contributing to them and their mixing proportions and thus, provide useful insights on the processes and factors responsible for their mobilization and sequestration. Source-identification and source-apportionment of elements in different earth system components

* Corresponding author at: AIRIE Program, Department of Geosciences, CO State University, Fort Collins, USA.
E-mail address: grtripathy@gmail.com (G.R. Tripathy).

http://dx.doi.org/10.1016/j.gexplo.2014.03.004
0375-6742/

are, however, not straight forward; they rely largely on the mathematical modeling of compositional data. Many mathematical models are available to apportion source contributions in a mixture. These include the principal component analysis (PCA), positive matrix factorization (PMF), least square regression (LSR), forward and inverse modeling (IM) (Allegre et al., 1983; Beus and Grigorian, 1977; Bickle et al., 2005; Bryan et al., 1969; Huang and Conte, 2009; Krishnaswami et al., 1999; Makinen and Gustavsson, 1999; Negrel et al., 1993; Paatero, 1997; Sohn, 2005; Tripathy et al., 2014; Weltje, 1997; Wu et al., 2005). Among these, the PCA and factor analysis are often used to quantify source contribution to mixtures such as river water, sediments and aerosols. The merit of these two approaches is that knowledge on the number of sources and their compositions is not a prerequisite for their application to derive information on the source contribution to a mixture. The applications of both PCA and factor analysis, however, are limited as at times these approaches can provide non-positive results for source contributions (Larsen and Baker, 2003; Sofowote et al., 2008). This problem of the PCA and factor analysis methods is addressed in the PMF model, which has recently gained attention in aerosol studies (Kim et al., 2004; Paatero, 1997; Paatero and Tapper, 1993; Sudheer and Rengarajan, 2012). Among the other mathematical models, least square regression has been widely used in geochemical problems (Bryan et al., 1969; Le Maitre, 1979; Wright and Doherty, 1970). For example, Makinen and Gustavsson (1999) used Chebyshev's series solution of linear programming to decipher the relative contribution of amphibolites, granitoids, quartzite and sand to a till (mixture). Quantification of source contribution to mixtures (river water) has also been carried out either through the forward model (Das et al., 2005; Krishnaswami et al., 1999; Moon et al., 2007; Tripathy et al., 2010) using a suite of mass balance equations and pre-assigned source composition or through inverse modeling, based on the composition of mixture and approximate information on the composition of their possible sources. Allegre et al. (1983) successfully used the inverse model to derive the contribution from various reservoirs to basalts to explain their chemical and isotopic composition. Subsequently, the inverse model has found extensive application to apportion the chemical and isotopic composition of rivers among its sources that include weathering of major lithologies present in their basins and atmospheric deposition (Millot et al., 2003; Moon et al., 2007; Negrel et al., 1993; Tripathy and Singh, 2010; Wu et al., 2005). More recently, the IM has also been extensively used to quantify various oceanic processes (Rahaman and Singh, 2012; Singh et al., 2012).

These models although used widely for source apportionment and quantification, only a few studies have attempted to intercompare the results yielded from various models. Such exercises are needed to enhance the confidence in the application of these models and interpret their quantitative results to address source apportionment and their geochemical significance. Morandi et al. (1991), based on inter-comparison of results from two different models (*viz.* modified version of factor analysis/multiple regression and regression on absolute principal components), observed that although the results showed an overall consistency, they also had some discernible differences. This led the authors to suggest the need for using more than one model to accurately quantify source contribution from datasets. More recently, the results of Tripathy and Singh (2010) on the inter-comparison of source contributions derived using the forward and inverse models of chemical and Sr isotopic composition of the Ganga headwaters showed that the results exhibit statistically significant differences. These observations highlight the importance of and need for inter-comparison of results from different mathematical approaches to obtain robust data on source contribution and inferences on associated geochemical processes. The present study is an attempt in this direction. It inter-compares the results on source apportionment obtained using two commonly employed mathematical techniques: the least square regression and the inverse model. The study has been conducted on two mixtures: the *global* river water and a glacial till. The results have led to better understanding of the inherent merits and limitations of these models, as discussed in the paper.

2. Methods

2.1. Least square approximation using QR decomposition

The source-apportionment models/programs are based on the mass balance approach that involves the formulation of suitable equations for the budget of each element in the mixture and solving the set of equations to determine the contribution from various sources. In the LSR method, the known parameters are the precisely constrained elemental abundances (or their ratios) in the mixture (*e.g.* river water or sediments) and in their various possible sources (end members). The unknown parameters, which need to be quantified, are the relative contribution from each source to the mixture.

A set of linear equations totaling the number of unknowns can provide a unique solution. However, geochemical studies focusing on source apportionment often have more number of equations (*i.e.*, number of geochemical parameters) than the number of unknowns (*i.e.*, relative contribution from different sources). Eq. (1) presents such an over-determined linear system, where the measured chemical data ($b_i = 1$ to m) of the system are related to the chemical composition of its possible sources/end-members (a_{ij}; $i =$ 1 to n; $j = 1$ to m) and their relative contributions to the mixture (x_i; $i = 1$ to m). In this case, the number of equations (n + 1) is more compared to the number of unknowns (m).

$$\left|\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots. + a_{1m}x_m = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots. + a_{2m}x_m = b_2 \\ \ldots\ldots\ldots \\ \ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots. + a_{nm}x_m = b_n \\ x_1 + x_2 + \ldots. + x_m = 1 \end{array}\right|. \quad (1)$$

These over-determined set of mass balance equations does not have a unique solution. However, the use of least square regression approach can provide a 'best estimate' solution for the contribution from each end member to the mixture (x_i; i = 1, m of Eq. (1)) which can satisfy Eq. (1) with the least residual.

As a first step to find the best-possible solution using the LSR, Eq. (1) can be rewritten as an equation of matrices, *i.e.*,

$$AX = B \quad (2)$$

where B and A are matrices containing chemical composition (n elements) of the mixture and all its possible sources (m), respectively and X lists the relative contribution from each of the sources. In order to solve Eq. (2) using the LSR approach, the matrices A and B need to be known accurately. To achieve this, least square approximation using the QR factorization of Eq. (2) was adopted in this study to find the solution of X. The QR decomposition of the matrix A is its factorization into an orthogonal matrix (Q) and a triangular matrix (R), *i.e.*

$$A = QR \quad (3)$$

where, Q is an orthogonal matrix, *i.e.* $Q^TQ = \mathrm{I}$. I is the identity matrix and superscript 'T' stands for the transpose of the matrix. The QR factorization transforms a linear over-determined least square problem into a well-defined triangular system.