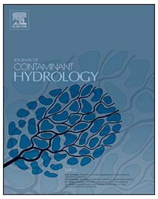
Contents lists available at ScienceDirect

Journal of Contaminant Hydrology

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



# Contaminant source identification using semi-supervised machine learning

Velimir V. Vesselinov[a,*], Boian S. Alexandrov[b], Daniel O'Malley[a]

[a] Computational Earth Science Group, Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, USA
[b] Physics and Chemistry of Materials Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, USA

ARTICLE INFO

*Keywords:*
Non-negative Matrix Factorization
Feature extraction
Blind Source Separation
Robustness analysis
Semi-supervised learning
Groundwater contamination
Source identification
Advection-diffusion transport
Geochemical signatures

ABSTRACT

Identification of the original groundwater types present in geochemical mixtures observed in an aquifer is a challenging but very important task. Frequently, some of the groundwater types are related to different infiltration and/or contamination sources associated with various geochemical signatures and origins. The characterization of groundwater mixing processes typically requires solving complex inverse models representing groundwater flow and geochemical transport in the aquifer, where the inverse analysis accounts for available site data. Usually, the model is calibrated against the available data characterizing the spatial and temporal distribution of the observed geochemical types. Numerous different geochemical constituents and processes may need to be simulated in these models which further complicates the analyses. In this paper, we propose a new contaminant source identification approach that performs decomposition of the observation mixtures based on Non-negative Matrix Factorization (NMF) method for Blind Source Separation (BSS), coupled with a custom semi-supervised clustering algorithm. Our methodology, called NMF*k*, is capable of identifying (a) the unknown number of groundwater types and (b) the original geochemical concentration of the contaminant sources from measured geochemical mixtures with unknown mixing ratios without any additional site information. NMF*k* is tested on synthetic and real-world site data. The NMF*k* algorithm works with geochemical data represented in the form of concentrations, ratios (of two constituents; for example, isotope ratios), and delta notations (standard normalized stable isotope ratios).

## 1. Introduction

For several decades, one of the most important research and real-world applications in the hydrogeological sciences has been related to aquifer contamination (Fetter and Fetter, 1999; Gelhar, 1993; Vengosh et al., 2014). The work has been driven by substantial scientific and engineering challenges associated with prediction and remediation of contaminant plumes in natural environment. Most of these challenges are due to uncertainties associated with contaminant sources. For example, the number of contaminant sources, their location and geochemical signatures are frequently unknown.

Typically, water in an aquifer is a mixture of different groundwater types with different origins and geochemical signatures (Deutsch and Siegel, 1997). For example, groundwater might be originating from different recharge sources with contrasting geochemical properties. Also groundwater may have been flowing through different rock types which may have altered the composition by means of geochemical reactions and ion exchanges. Furthermore, some of the groundwater recharge sources might be associated with contamination sources with different geochemical signatures. Data about the groundwater mixtures are typically collected at multiple sampling locations over time where the measurement data are also associated with uncertainties and measurement errors. Identification of the original groundwater types representing geochemical mixtures observed in a aquifer is a challenging but very important task (Wagner, 1992; Böhlke and Denver, 1995; Lapworth et al., 2012). This task is typically performed using complex inverse models representing groundwater mixing processes in the aquifer, where the model is calibrated against the available observation data characterizing the spatial and temporal distribution of the observed geochemical data (Wagner, 1992; Neupauer et al., 2000; Atmadja and Bagtzoglou, 2001; Michalak and Kitanidis, 2004; Guan et al., 2006; Mamonov and Tsai, 2013; Hamdi and Mahfoudhi, 2013; Murray-Bruce and Dragotti, 2014; Borukhov and Zayats, 2015). Numerous different geochemical constituents may need to be simulated in these models which further complicates the analyses.

Contemporary analyses of groundwater contamination sources are also performed implementing various multivariate statistical and machine learning techniques (Chan and Huang, 2003; Rasekh and Brumbelow, 2012). Thus, variations in chemical compositions and evolution of groundwater composition have been studied by methods used to describe variability among correlated variables (Knudson et al., 1977; Helena et al., 2000) (such as, Factor Analysis, Harman, 1976; and

[*] Corresponding author.

https://doi.org/10.1016/j.jconhyd.2017.11.002
Received 19 April 2017; Received in revised form 28 October 2017; Accepted 3 November 2017
0169-7722/ Published by Elsevier B.V.

Principle Component Analysis, Jolliffe, 2002), as well as by unsupervised machine learning methods used to characterize or separate two or more classes of objects (Shrestha and Kazama, 2007; Tariq et al., 2008) (such as, Discriminant Analysis, Scholkopft and Mullert, 1999; and Clustering Analysis, Diday and Simon, 1980). Determination of the average regional concentrations of heavy metals (based on surveys of soil contamination) has been also investigated by Principle Component Analysis and Cluster Analysis (Facchinelli et al., 2001), combined with a geostatistical method (Chiles and Delfiner, 2009) used to construct regional distribution maps for comparison with regional databases. Various supervised machine learning techniques, such as, Artificial Neural Networks, Yegnanarayana (2009), Support Vector Machines, Drucker et al. (1999), Locally Weighted Projection Regression, Vijayakumar and Schaal (2000), and Relevance Vector Machines, Tipping (2001), etc., have been also utilized to build surrogates models (based on observational data) for substitution of the much more complex and time-consuming physical models used to simulate the contamination levels. Such surrogate models have been used to predict contaminant levels in regional groundwater sites (Khalil et al., 2005). The relationship among different chemical pollutants retrieved from *in situ* measurements of underground and surface water have been investigated by an algorithm for quasi-optimal learning, Cervone et al. (2010) that explores a methodology for symbolic machine learning classification. By this method it has been shown, for example, that if one type of contaminant is dissolved in the water table, it has to be expected that other chemicals are also present (Manca and Cervone, 2013).

In this paper, we utilize a new hybrid approach, which we call NMF*k*, for identification of contaminant sources in an aquifer. NMF*k* utilizes a Blind Source Separation (BSS) technique (Belouchrani et al., 1997), based on Non-Negative Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999), combined with a custom made semi-supervised *k*-means clustering algorithm (Alexandrov and Vesselinov, 2014), to unmix the geochemical signatures in the observations and identify the contaminant sources.

Using synthetic and real-world site data, we demonstrate that NMF*k* is capable of accurately determining the unknown number of contaminant sources from observation samples of their mixtures, without any additional information. The NMF*k* methodology is coded in Julia (Bezanson et al., 2012) and the code is available upon request. The NMF*k* algorithm works with geochemical data represented in the form of concentrations, ratios (of two constituents, for example, isotope ratios), and delta notations (standard normalized stable isotope ratios).

Frequently, at the contamination sites, the groundwater in the aquifer is a mixture of waters with different origins (sources) that are commingled in the aquifer; several of these groundwater recharge sources might include contaminants. Typically, all these sources will have different geochemical signatures due to differences in their origins and flowpaths through the subsurface before infiltrating in the aquifer. The identification of the contamination/infiltration sources causing the observed geochemical concentrations in the aquifer can be very challenging at sites where complex physical and chemical processes occur.

Source identification can be complicated because (1) some of these sources may have similar geochemical signatures, (2) some of the sources may geochemically interfere with each other, and (3) groundwater transport through the subsurface (from the entry point at the ground surface to the observation point in the aquifer) may be impacted by various physical and chemical processes (e.g., diffusion, dispersion, sorption, retardation, precipitation, etc. and precipitation). To address all these issues, the source characterization is often carried out by calibrating a numerical model that simulates these complexities against the observed geochemical data. Here, we apply an alternative approach based on a novel model-free machine learning algorithm for Blind Source Separation (BSS).

## 2. Blind Source Separation (BSS)

The main goal of the paper is to present a novel application of the BSS methodology based on NMF*k* algorithm presented in Alexandrov and Vesselinov (2014). Additionally, substantial changes and extensions of the original NMF*k* algorithm are reported in order for NMF*k* to be applicable for contaminant source identification based on geochemical observations as discussed below.

We assume that the geochemical observations are taken at several discrete detectors (sampling points; typically monitoring wells) dispersed in space. The algorithm does not require the data to account for transients. In the case of transient data, NMF*k* can be applied consecutively to representative time snapshots, which will account for changes and evolution of the mixing ratios of the detected groundwater types.

When there are multiple contamination sources in the aquifer each detector registers a mixture of contamination fields originating from different sources (release locations). Our objective is to identify the unknown number of original contamination sources, which necessitates decomposing the recorded mixtures to their original components.

Consistent with the BSS methodology (Belouchrani et al., 1997), the contaminant source identification problem addressed here can be formulated as following:

$$V = W \times H + \varepsilon, \tag{1}$$

where $V$ is a matrix ($V \in M_{n,m}(R)$) of the known observation data representing $m$ geochemical constituents detected at a set of $n$ detectors (monitoring wells). The $V$ matrix does not need to be a full matrix and there can be empty entries, where not all the geochemical constituents are observed at all the wells. $W$ is an unknown source mixing matrix ($W \in M_{n,k}(R)$) representing the mixing coefficients of $k$ unknown original groundwater types at each of the $n$ observation points. Note that in this formulation, the sum of the mixing coefficients for each observation point should add to one $\left(\sum_{j=1}^{k} W_{i,j} = 1, \text{ for each well } i\right)$ and all the mixing coefficient should be between 0 and 1 (i.e., $0 < \; = W_{i,j} < \; = 1$). These requirements come from the problem setup; the groundwater concentrations at each well are expected to be defined by mixing of all the sources, and there are no expectations to have negative source contributions. $H$ is the unknown source matrix ($H \in M_{d,m}(R)$) representing the $m$ geochemical concentrations for each $k$ unknown original sources. The matrix elements of $V$, $W$ and $H$ are expected to be positive (which is consistent with the analyzed problem; concentrations cannot be negative). $\epsilon$ denotes presence of unknown noise or unbiased errors in the measurements ($\epsilon \in V_m(\mathbb{R})$).

If there are transients in the observed data, the BSS problem formulated above can be solved for a sequence of temporally discretized snapshots. In this case, the matrices $V$, $W$ and $H$ will be time dependent. In the discrete case, the BSS analyses will solve for $k$, $H_t$ and $W_t$ based on a series of inputs $V_t$, where $t = 1,\ldots,T$, and $T$ is number discretized moments in time, at which the signals are recorded at the detectors. This is also consistent with the data acquisition strategies typically used at an actual contamination site where the geochemical data are collected on annual or quarterly basis.

The joint analyses of the transients will increase the dimensionality of the data requiring factorization of tensors not matrices as presented in Eq. (2) (cf. Cichocki et al., 2009). However, currently, there are no tensor-based methods that can be applied to solve the multi-dimensional geochemical mixing problems.

Since both factors $H$ and $W$ are unknown (the size $k$ of these matrices is also unknown, because we do not know how many sources have been mixed in each detector record), the main difficulty in solving a BSS problem is that it is under-determined.

There are two widely-used approaches to resolve this BSS under-determination: Independent Component Analysis (ICA), Amari et al. (1996), Herault and Jutten (1986), and Non-negative Matrix

Factorization (NMF), Lee and Seung (1999), Paatero and Tapper (1994). ICA presupposes a statistical independence of the original signals and thus aims to maximize the non-Gaussian characteristics of the estimated sources in $H$. The other approach, NMF, is an unsupervised learning method, created for parts-based representation, Fischler and Elschlager (1973) in the field of image recognition, Lee and Seung (1999), Paatero and Tapper (1994) that was successfully leveraged for decomposition of mixtures formed by various types of signals (Cichocki et al., 2009). In contrast to ICA, NMF does not seek statistical independence or constrain any other statistical properties (i.e., NMF allows the original sources to be correlated); instead, NMF enforces a non-negativity constraint on the original signals in $H$ and their mixing components in $W$. NMF can successfully decompose large sets of nonnegative observations, $V$, by leveraging the multiplicative update algorithm (Lee and Seung, 1999); however, NMF requires *a priori* knowledge of the number of the original sources.

Recently, we reported a methodology, called NMF*k* (Alexandrov and Vesselinov, 2014), where the coupling of the original multiplicative algorithm with a custom semi-supervised clustering enables it to identify the number of the unknown sources based on the robustness of the solutions. Here, by imposing additional (to non-negativity) constraints to the elements of the mixing matrix and applying a nonconvex nonconvex nonlinear minimization algorithm, we extend NMF*k* to be applicable for the contaminant source identification based on geochemical signatures in groundwater samples representative of contamination sites. The extended NMF*k* methodology is also applicable for any other situation where the contributions of the original signals in the observed mixtures have to add to 1 due to additional physical constraints.

## 3. Methodology

### 3.1. NMF algorithm

In a typical NMF problem, the observational data, $V$, is formed by a linear mixing of $k$ unknown original signals, $H$, blended by an also unknown mixing matrix, $W$, i.e.,

$$V_{n,m} = \sum_{i=1}^{k} W_{n,k} H_{k,m} + \epsilon, \tag{2}$$

subject to the following constraints:

$$W_{n,d} > 0, H_{d,m} > 0; \quad \forall \quad n, d, m. \tag{3}$$

Here, $\epsilon$ is a vector and denotes presence of possible noise or unbiased errors in the measurements (also unknown). If the problem is solved in a temporally discretized framework, the goal of the BSS algorithm is to retrieve the $k$ original signals, $H$, that have produced $n$ observational mixtures of these signals, $V$, recorded at a set of observation points (sensors). Here, $n$ is the number of the sensors, $k$ is the number of the unknown signals (sources) observed in the collected data ($V$), and $m$ is the number of observed geochemical constituents associated at the observation points. The algorithm returns the decomposition through the mixing matrix $W_{n,d}$ and source matrix $H_{d,m}$ with $\epsilon$ being the residual noise. The rows in $V$ correspond to the number of sensors while the rows of $H$ correspond to the number of sources. In the NMFusually, usually, the number of sensors has to be greater than the number of sources. For NMF to work, the problem must be amenable to a non-negativity constraint on the sources $H$ and mixing matrix $W$. This constrain leads to reconstruction of the observations (the rows of matrix $V$) as linear combinations of the elements of $H$ and $W$ that cannot cancel mutually.

The NMF algorithm starts with a random guess for $H$ and $W$, and proceeds by minimizing the cost (objective) function, $O$, which in our case is the Frobenius norm,

$$O = \frac{1}{2} \; \|V - W * H\|_F^2 = \frac{1}{2} \sum_{n,m} \left( V_{n,m} - \sum_{k=1}^{d} W_{n,k} H_{k,m} \right)^2 \tag{4}$$

during each iteration. Minimizing the Frobenius norm (Eq. (4)) with non-negativity constraints (Eq. (3)) is equivalent to representing the discrepancies between the observations, $V$, and the reconstruction, $W * H$, as white noise.

Furthermore, to find the contaminant sources (geochemical types) represented in the observed geochemical mixtures, here we have to minimize $O$ with the additional constraints,

$$\sum_{k=1}^{d} W_{n,k} = 1; \quad \forall \quad n. \tag{5}$$

It is important to note that because of the constraints in Eq. (5), the classical multiplicative NMF optimization algorithm (Lee and Seung, 1999) is not applicable. Instead, a non-convex nonlinear optimization algorithm is needed, and for this purpose we utilized the nonlinear minimization procedure provided by Julia packages JuMP.jl and Ipopt.jl. JuMP.jl is a modeling language for mathematical optimization embedded in Julia (Dunning et al., 2015). It supports a number of open-source and commercial solvers for a variety of optimization problems. Here, JuMP.jl is applied for nonlinear programming using Ipopt.jl. Ipopt (Interior Point OPTimizer) is an open-software package for large-scale nonlinear optimization (Wächter, 2002; Wächter and Biegler, 2005, 2006). Here, Ipopt is applied to perform non-convex constrained second-order minimization.

### 3.2. NMFk algorithm

If we knew the number of sources, the first step described in the previous section would be all that is needed: from the best solution of the minimization procedure (with known $k$) we would extract the desired estimates of the physical parameters, and thus solve the inverse problem. Unfortunately, the true number of sources is typically unknown, and thus the number of the sources is an unknown parameter which we have to identified from the observations. Further, the solutions of Eq. (2), is based on random initial conditions. A naive approach would be to explore all of the possible solutions applying the nonlinear minimization described in the previous section for a range of possible number of sources. Then the solution with the smallest norm will identify the number of sources, $k_s$. However, this is obviously flawed approach – — the over-fitting will certainly lead to an over-estimation of the number of sources: more free parameters will generally lead to a better fit, irrespective of how close the estimated number of sources is to the real one.

The classical NMF also requires *a priory priori* knowledge of the number of the original sources. Previously, by coupling the NMF with a custom semi-supervised clustering, we have demonstrated that the number of the original sources can be estimated based on their robustness/reproducibility (Alexandrov and Vesselinov, 2014). This approach was introduced to decompose the largest available dataset of human cancer genomes (Alexandrov et al., 2013), and then extended for decomposition of physical signals/transients (Alexandrov and Vesselinov, 2014).

Specifically, our methodology, called NMF*k*, explores consecutively all possible numbers of original sources $k$ ranging from 1 to $d$ ($k = 1{,}2, \ldots{,}d$), and then estimates the accuracy and robustness of large set of solutions with different number of sources.

In NMF*k*, the maximum number of explored sources $d$ is user defined and it is not expected to exceed the number of observed geochemical components or the number of observation points (although, theoretically, the used here minimization algorithm can be applied for any $k > 1$).

Thus, NMF*k* performs $M$ sets of simulations, called NMF runs, where each run is using different number of sources, $k = 1{,}2{,}\ldots{,}d$, with random initial conditions. At the end of each NMF run, we get a set of $M$ solutions, $U_k$, where each solution contains two matrices, $H_k^j$, $W_k^j$, (for $k$ original sources, and $j = 1,2,\ldots,M$),

$$U_k = ([H_k^1; W_k^1], [H_k^2; W_k^2], \ldots, [H_k^M; W_k^M]). \tag{6}$$

After that, NMF$k$ leverages a custom semi-supervised clustering to assign each of these $M$ solutions in a given set, $U_d$, to one of $k$ specific clusters. This custom semi-supervised method is based on $k$-means clustering that keeps the number of solutions in each cluster equal to the number of NMF runs. For example, for the case with $k = 2$, after the execution of $M = 1000$ NMF runs (performed with random initial guesses for the $W$ and $H$ matrix elements), each of the two clusters will contain 1,000 1000 solutions. Note that we have to enforce the condition that the clusters are with equal number of solutions, since each NMF simulation contributes equal number of solutions for each source. During the clustering, the similarity between sources $H_{i1}$ and $H_{i2}$ is measured using the cosine distance (also known as cosine similarity) (Pang-Ning et al., 2006; Alexandrov and Vesselinov, 2014).

The main idea for estimating the unknown number of sources in NMF$k$ is to use the separation between the clusters as a measure of how good a particular choice of $k$ is as an accurate estimate of the number of unknown sources. We estimate the degree of clustering for different number of sources, and plot it as a function of $k$, we expect a sharp drop after we cross the $k_s$ value (Alexandrov and Vesselinov, 2014).

To quantify this behavior, after the clustering, we compute the average silhouette width (Rousseeuw, 1987), $S(k)$, which is a measure of how well the solutions are clustered for given number of original sources, $k$. The average silhouette width of the clusters for the NMF$k$ solutions for different $S(k)$ values can be applied to evaluate the optimal number of contaminant sources, $k_s$. In general, $S(k)$ declines as $k$ increases. Theoretically, $S(k)$ varies between 1 and $-1$. For $k = 1$, $S(1) = 1$ since there is only one solution. Typically, $S(k)$ declines sharply after the optimal number of contaminant sources, $k_s$, is reached.

In NMF$k$, in addition to the robustness, the average reconstruction error (Eq. (4)) is used to evaluate the accuracy with which the derived average (cluster) solutions $[H_k^a; W_k^a]$ reproduce the observations $V$. In general, the solution accuracy increases (while the solution robustness decreases) with the increase of the number of unknown sources. Hence, the average silhouette width and Frobenius norm for each of the $k$ cluster solutions can be used to define the optimal number of contaminant sources, $k_s$. Specifically, $k_s$ can be select to be equal to the minimum number of sources that accurately reconstruct the observations (i.e., the Frobenius norm is less than a given value or hit plateau) and the clusters of solutions are sufficiently robust (or stable, i.e., the average silhouette width $S$ is bigger than 0.8).

When some of the source geochemical compositions are very close to each other or do not demonstrate clear features, it is more useful to formulate another criteria for the NMF$k$ solution robustness, which is based on the Akaike Information criterion ($AIC$) (Akaike, 2011). Specifically, to compare the NMF models with different number of sources we calculate for each of them the $AIC$ value. To calculate $AIC$, we take from each of the sets of solutions with different number of sources, $U_k$, the best NMF solution, and use the corresponding Frobenius norm, $O^{(k)}$, in the $AIC$ formula:

$$AIC = 2N - 2\ln(L) = 2(k(n+m) - n) + nm\ln\left(\frac{O^{(k)}}{nm}\right). \tag{7}$$

Here, the number of adjustable NMF$k$ parameters, $N$, is equal to the number of components in the $W$ and $H$ matrices minus the number of observations points, because we impose the constraint $\sum_{j=1}^{k} W_{n,j} = 1$ for each observation points, which reduces the number of adjustable parameters. Thus, we have, $N = nk + mk - n = k(n + m) - n$, where $k$ is the number of sources, $m$ is the number of wells and $n$ is the number of the observation points. $L$ is the likelihood functions of the NMF solution with given $k$, and we define it using the reconstruction error $O^{(k)}$ of the NMF solutions: $\ln(L) = -(nm/2)\ln(O^{(k)}/nm)$ ($nm$ is the total number of observational data points; the product of the number of detectors by the number of time slices).

The $AIC$ is a standard measure of the relative quality of statistical models, which takes into account both the likelihood function (in our case determined by the reconstruction error) and the independent degrees of freedom needed to achieve this level of likelihood (the elements of the matrices $W$ and $H$). Choosing the model that minimizes $AIC$ helps avoid over-fitting. In general, $AIC$ decreases with increasing the number of estimated sources $k$. Typically, $AIC$ substantially drops when $k = k_s$. For $k > k_s$, the $AIC$ values commonly plateau and do not exhibit substantial changes.

In general, both the average silhouette width $S$ and $AIC$ should estimate the same number of sources $k_s$. If there is discrepancy, typically, $S$-based estimate is smaller than the $AIC$-based estimate (this type of situation is discussed in the results section below). In general, $S$-based estimate of $k_s$ should be preferred because the solutions for $k > k_s$ are potentially over fitting the data.

## 4. Results

### 4.1. Synthetic Analysis analysis

#### 4.1.1. Example with two sources and three geochemical constituents

To illustrate our method, we apply the NMF$k$ algorithm described above to identify the source concentrations from a series of synthetic data sets representing realistic scenarios generally consistent with real world conditions.

First, we consider an example generated to represent two unknown synthetic sources (groundwater types). The "true" unknown concentrations of three geochemical constituents (A, B & C) representing the two synthetic sources are presented in Table 1. These sources are mixed at each well using "true" unknown mixing coefficient show shown in Table 2. These are the "true" unknown matrices $W$ and $H$, respectively, as presented in Eq. (2). These matrices are unknown; the number of sources are also unknown. They are presented here just to demonstrate the applicability of the method. The "true" matrices $W$ and $H$ in Tables 1 and 2 are multiplied to estimate the "true" known concentrations $V$ (Table 3) of three geochemical constituents (A, B & C) at the five monitoring wells. Here, the measurement errors are assumed to be zero.

Here and in the examples presented below, the source concentrations and well mixing coefficients (Tables 1 and 2) are generated using standard pseudo random pseudo-random number generation capabilities provided in Julia; the random numbers have uniform distribution between 0 and 1. For convenience and without lost of generality, the source concentrations are scaled so that the maximum concentration at the sources for each species is 1. The random mixing coefficients are also scaled so that each row in Table 2 adds up to 1. As discussed above, this requirement comes from the problem setup; the groundwater concentrations at each well are expected to be defined by mixing of all the sources.

We applied $V$ in NMF$k$ to estimate the number of sources and reconstruct the unknown source concentrations and mixing coefficients. Based on Table 4, the number of source is two. This is estimated by the behavior of the robustness and $AIC$ criteria. The robustness is close to 1

**Table 1**
True and estimated concentrations of three geochemical constituents (A, B & C) representing two synthetic sources (S1 & S2).

| Source | True | | | Estimated | | |
|---|---|---|---|---|---|---|
| | A | B | C | A | B | C |
| S1 | 0.932661 | 0.793833 | 1.0 | 0.927047 | 0.776642 | 1.07732 |
| S2 | 1.0 | 1.0 | 0.0727242 | 0.996028 | 0.987838 | 0.127424 |

**Table 2**
True and estimated mixing coefficients of the two sources at five monitoring wells.

| Well | True | | Estimated | |
|---|---|---|---|---|
| | S1 | S2 | S1 | S2 |
| W1 | 0.901005 | 0.0989955 | 0.821967 | 0.178033 |
| W2 | 0.734414 | 0.265586 | 0.659343 | 0.340657 |
| W3 | 0.33299 | 0.66701 | 0.267476 | 0.732524 |
| W4 | 0.466407 | 0.533593 | 0.397717 | 0.602283 |
| W5 | 0.468169 | 0.531831 | 0.399436 | 0.600564 |

**Table 3**
True and estimated concentrations of the three geochemical constituents (A, B & C) observed at five observation points.

| Well | True | | | Estimated | | |
|---|---|---|---|---|---|---|
| | A | B | C | A | B | C |
| W1 | 0.939328 | 0.814242 | 0.908204 | 0.939328 | 0.814242 | 0.908204 |
| W2 | 0.950546 | 0.848588 | 0.753729 | 0.950546 | 0.848588 | 0.753729 |
| W3 | 0.977577 | 0.931348 | 0.381497 | 0.977577 | 0.931348 | 0.381497 |
| W4 | 0.968593 | 0.903842 | 0.505212 | 0.968593 | 0.903842 | 0.505212 |
| W5 | 0.968474 | 0.903479 | 0.506846 | 0.968474 | 0.903479 | 0.506846 |

**Table 4**
NMF*k* results for the problem presented in Table 3; the reconstruction quality *O*, silhouette width *S*, and *AIC* are estimated for number of sources *k* = 1,2,3.

| *k* | *O* | *S* | *AIC* |
|---|---|---|---|
| 1 | 0.1934462 | 1 | 11.45456 |
| 2 | $8.345958 \times 10^{-16}$ | 0.9843997 | −235.0658 |
| 3 | $2.505855 \times 10^{-16}$ | 0.5936544 | −242.4891 |

for the cases of 1 and 2 sources; however, it drops substantially for 3 sources. This suggest that the solution for 3 sources is not stable and non-unique and solution for 2 sources should be preferred. Similarly, *AIC* shows a substantial drop between cases of 1 and 2 sources; this also suggests that the solution with 2 sources should be selected. The same conclusion can be also drawn here by the reconstruction quality. Clearly the solution for 2 sources produces a much better fit of the data than the solution for 1 source. The solution for 3 sources produces a better match but based on parsimony principle (also captured by *AIC*) but using much more model parameters (i.e., more degrees of freedom). In this case, the 2 source solution has 16 model parameters (5 × 2 + 2 × 3) while the 3 source solutions has 21 model parameters (5 × 3 + 3 × 3). In all cases, there are only 15 observations (5 × 3).

The estimated unknown concentrations of the three geochemical constituents (A, B & C) representing two synthetic sources are presented in Table 1. The estimated unknown mixing coefficients of the identified two sources in the five wells are shown in 2Table 2. As can be seen, the algorithm accurately estimates the number of sources. It is also capable of almost perfectly reproducing the observed concentrations (Table 3) which is not surprising considering the large number of degrees of freedom. The algorithm accurately captures the general pattern of geochemical constituent concentrations in the original sources (Table 1). It also accurately identifies the general pattern of representation (mixing) of the two original sources in the five monitoring wells (Table 2).

The same synthetic problem was rerun 1,000 1000 times with different random concentrations. In all the 1000 cases, the algorithm correctly identified the true number of sources. The same synthetic problem was also rerun 1,000 1000 times adding random noise with normal distribution (mean equal to zero and standard deviation equal to 0.01) representing measurement errors. Again, the algorithm correctly identified the true number of sources all test cases.

**Table 5**
Concentrations of six geochemical constituents observed at 30 observation points with noise.

| Well | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| W1 | 0.824159 | 0.770184 | 0.650546 | 0.672767 | 0.328136 | 0.755944 |
| W2 | 0.721702 | 0.657699 | 0.659243 | 0.519665 | 0.127078 | 0.846111 |
| W3 | 0.806203 | 0.620875 | 0.560077 | 0.658265 | 0.465434 | 0.705513 |
| W4 | 0.846015 | 0.867051 | 0.657893 | 0.690346 | 0.49869 | 0.714246 |
| W5 | 0.717991 | 0.797765 | 0.752877 | 0.484986 | 0.124759 | 0.866724 |
| W6 | 0.870077 | 0.694686 | 0.577234 | 0.757505 | 0.538225 | 0.67329 |
| W7 | 0.705147 | 0.633675 | 0.632464 | 0.526834 | 0.200167 | 0.812269 |
| W8 | 0.837582 | 0.734872 | 0.592473 | 0.707935 | 0.351192 | 0.735757 |
| W9 | 0.839538 | 0.727734 | 0.557825 | 0.759988 | 0.269487 | 0.727261 |
| W10 | 0.837105 | 0.599062 | 0.488185 | 0.755697 | 0.464891 | 0.686943 |
| W11 | 0.823122 | 0.775698 | 0.615535 | 0.711279 | 0.369447 | 0.728862 |
| W12 | 0.70415 | 0.820778 | 0.75013 | 0.427328 | 0.177106 | 0.866145 |
| W13 | 0.816797 | 0.911729 | 0.742964 | 0.614281 | 0.10748 | 0.828933 |
| W14 | 0.774421 | 0.828243 | 0.693439 | 0.619605 | 0.181521 | 0.824042 |
| W15 | 0.682291 | 0.690584 | 0.706534 | 0.451311 | 0.121616 | 0.873707 |
| W16 | 0.891054 | 0.785168 | 0.567378 | 0.847566 | 0.426966 | 0.658821 |
| W17 | 0.818681 | 0.892126 | 0.714567 | 0.662251 | 0.30273 | 0.783264 |
| W18 | 0.849113 | 0.939263 | 0.710556 | 0.689737 | 0.383933 | 0.748372 |
| W19 | 0.824445 | 0.83857 | 0.673778 | 0.664912 | 0.411412 | 0.749026 |
| W20 | 0.92851 | 0.72809 | 0.539854 | 0.829349 | 0.750985 | 0.597475 |
| W21 | 0.809177 | 0.776766 | 0.640878 | 0.669215 | 0.301763 | 0.764136 |
| W22 | 0.748145 | 0.724945 | 0.683463 | 0.506134 | 0.112246 | 0.84714 |
| W23 | 0.67255 | 0.971162 | 0.950029 | 0.301939 | 0.088057 | 0.96026 |
| W24 | 0.802038 | 0.740081 | 0.582657 | 0.742845 | 0.157864 | 0.766039 |
| W25 | 0.796033 | 0.795485 | 0.646577 | 0.640355 | 0.412495 | 0.750769 |
| W26 | 0.749082 | 0.479525 | 0.495895 | 0.630318 | 0.171125 | 0.78558 |
| W27 | 0.793644 | 0.692601 | 0.603123 | 0.68066 | 0.25344 | 0.758854 |
| W28 | 0.776949 | 0.783003 | 0.652028 | 0.627913 | 0.258182 | 0.775171 |
| W29 | 0.817962 | 0.826797 | 0.661437 | 0.65274 | 0.448026 | 0.737439 |
| W30 | 0.821272 | 0.726155 | 0.589883 | 0.684018 | 0.371474 | 0.723059 |

### 4.1.2. *Example with four sources and six geochemical constituents*

As a second test, we consider an example generated to represent four unknown synthetic sources (groundwater types) observed at 30 observation points. The synthetic observations *V* are presented in Table 5. The concentration data is perturbed by adding random noise with normal distribution (mean equal to zero and standard deviation equal to 0.01) representing measurement errors. We applied *V* in NMF*k* to estimate the number of sources.

Based on Table 6, the number of source is four. This is estimated by the behavior of the average silhouette width *S* and *AIC* criteria as a function of the number of sources *k*. The average silhouette width *S* is close to 1 for the cases when *k* ≤ 4. *S* drops slightly for *k* = 4 but it is still close to 1. Substantial drop for *S* occurs for *k* > 4 This suggest that the solution for more than 4 sources are non-unique and depends strongly on the random initial guesses for the unknown matrix components *W* and *H*.

*AIC* shows a substantial drop between cases of 3 and 4 sources; this suggests that the solution with 4 sources should be selected.

In this case, the same conclusion can be also drawn here by the reconstruction quality *O*. Clearly, the solution for 4 sources produces much better fit of the data than the solution for 3 sources. The solution for 5 sources also produces a good match but based on parsimony principle (also captured by *AIC*), it should be rejected because it is using

**Table 6**
NMF*k* results for the problem presented in Table 5; the reconstruction quality *O*, silhouette width *S*, and *AIC* are estimated for number of sources *k* = 1,…,6.

| *k* | *O* | *S* | *AIC* |
|---|---|---|---|
| 1 | 1.97693 | 1 | −800.0541 |
| 2 | 0.5999861 | 0.9999628 | −942.685 |
| 3 | 0.1606472 | 0.9998313 | −1107.87 |
| 4 | 0.007977784 | 0.9499047 | −1576.329 |
| 5 | 0.00354382 | −0.2686716 | −1650.391 |
| 6 | 0.001058438 | −0.5015211 | −1795.905 |

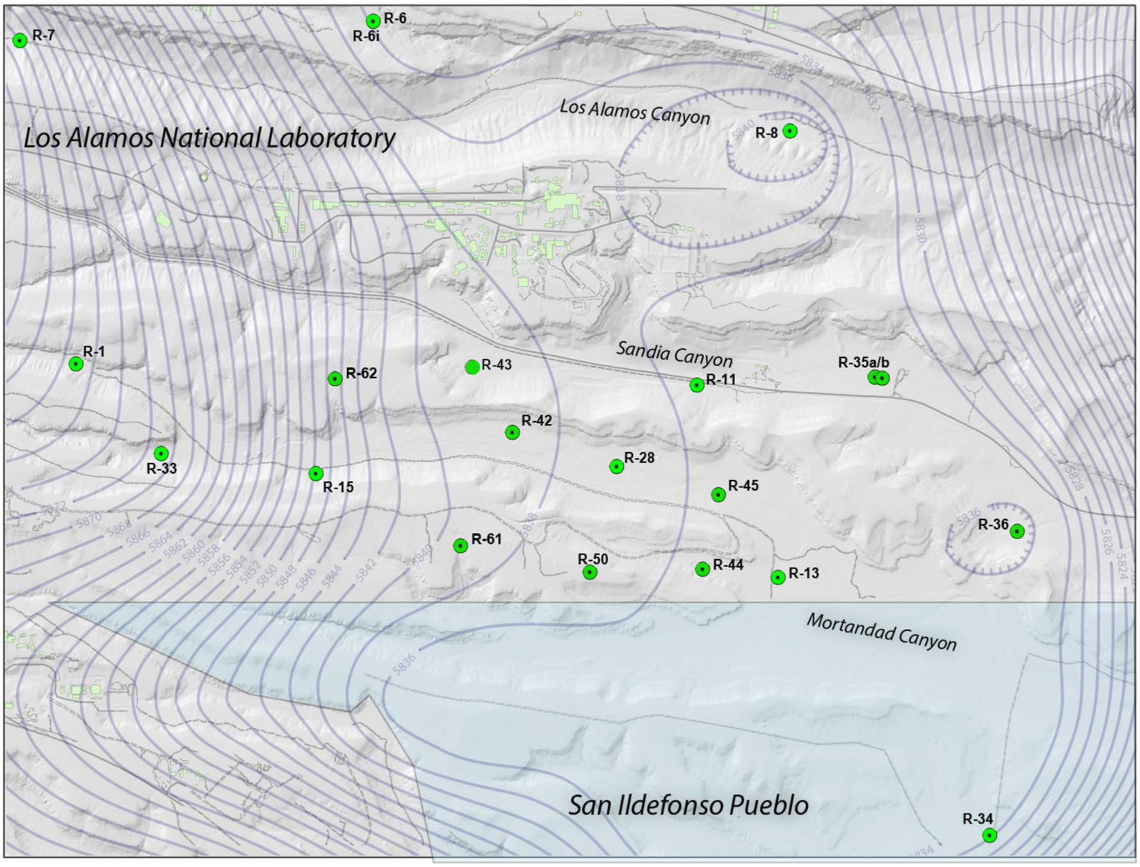


**Fig. 1.** LANL site map showing location of the monitoring wells.

much more model adjustable parameters. In this case, the 4 source NMF*k* solution has 114 adjustable parameters ($30 \times 4 + 4 \times 6 - 30$) while the 5 source solution has 150 parameters ($30 \times 5 + 5 \times 6 - 30$). In all cases, there are only 180 observations ($30 \times 6$).

This synthetic problem was rerun 1,000 1000 times with different random concentrations. All the runs are performed adding random noise with normal distribution (mean equal to zero and standard deviation equal to 0.01). In 912 cases, the algorithm correctly identified the true number of sources.

### 4.2. Site Analysis analysis

NMF*k* is applied to analyze and deconstruct the groundwater geochemistry observed in the regional aquifer beneath the Los Alamos National Laboratory (LANL) site for characterization of contaminant sources. LANL is a research facility operated by the U.S. Department of Energy in north-central New Mexico. LANL is currently investigating a chromium ($Cr^{6+}$) plume in the regional aquifer beneath Sandia and Mortandad Canyons (Fig. 1) to ensure contaminants do not threaten human health or the environment. The chromium contamination is caused by infiltration of liquid effluents released from an electric power plant. A comprehensive investigation of this plume has been ongoing since 2005 (Vesselinov et al., 2015, 2013). The site conceptual model describing the physical and biogeochemical processes controlling the movement of groundwater and contaminants in the environment is presented in detail in LANL (2012), and supported by multiples lines of evidence. In general, the site conceptual model that was proposed in LANL (2012) is still consistent with the recently collected data. The establishment and the ongoing testing of the current conceptual model involved a series of field, laboratory and modeling analyses (LANL, 2012; Vesselinov et al., 2015, 2013).

The contaminant source mass and the release history on the ground surface is highly uncertain. Also uncertain are the volume, transients and location of the infiltrating water carrying the contamination in the subsurface. The water and contaminants infiltrated through 300 m thick vadose zone which includes several perching horizons before reaching the regional aquifer water table. The hydraulic properties of the regional aquifer are highly heterogeneous. The shape of the regional water-table is impacted by these heterogeneities as well as by zones of infiltration from the vadose zone. Furthermore, migration of the contaminants in the subsurface is also influenced by effluent and water discharges in two neighboring canyons: Mortandad and Los Alamos (Fig. 1). Past Mortandad Canyon effluent releases contain nitrate ($NO_3^-$), and tritium ($^3H$). Past effluent releases in Los Alamos Canyon are characterized by elevated $^3H$ concentrations. Some of these Mortandad and Los Alamos Canyon tracers were collocated with contaminant released in Sandia Canyon and detected in the regional aquifer.

As a result, the geochemistry of the regional aquifer groundwater is expected to be representative of several commingled groundwater types with different geochemical signatures. The groundwater types can be related to different infiltration flowpaths as wells as background aquifer groundwater coming upgradient from the site.

A subset of the data collected at the site is applied for the NMF*k* analysis and is presented in Table 7. The NMF*k* results are presented in Table 8. The NMF*k* analysis suggest 5 original groundwater sources with different geochemical composition are mixed in the aquifer. This estimate is based on the silhouette width $S$ values. Note that $S \approx 1$ for $k \leq 5$. The *AIC* values potentially suggest existence of 7 sources. However, the solutions for 6 and 7 sources are potentially over-fitting the data.

Some of the estimated 5 sources (groundwater types) are associated with contaminant releases (Table 9). Based on Table 9, the fifth source, $s_5$, has the highest $^3H$ concentrations; the other geochemical

**Table 7**
Concentrations of six geochemical constituents observed at 18 monitoring wells (observation points) at the LANL site; the number after # defines the screen number for two-screen wells; #1 and #2 are the shallow and deep screens, respectively.

| Well | $Cr^{6+}$ | $Br^-$ | $Cl^-$ | $^3H$ | $NO_3^-$ | $SO_4^{2-}$ |
|---|---|---|---|---|---|---|
| R-14#1 | 5.72 | 0.2 | 1.64182 | 0.196983 | 0.328417 | 1.91 |
| R-1 | 5.74846 | 0.186167 | 1.87167 | 0.0901333 | 0.349923 | 2.40083 |
| R-33#1 | 5.61786 | 0.2 | 2.27143 | 0.460143 | 0.589214 | 3.25357 |
| R-33#2 | 6.25786 | 0.2 | 1.96929 | 0.23756 | 0.341214 | 2.39929 |
| R-15 | 12.1439 | 0.116743 | 4.19294 | 29.6705 | 2.20824 | 6.52529 |
| R-62 | 150.571 | 0.120293 | 7.95461 | 7.67933 | 1.19193 | 13.8543 |
| R-61#1 | 14.4147 | 0.184533 | 3.136 | 23.213 | 1.6878 | 5.15267 |
| R-43#1 | 54.2395 | 0.120368 | 6.39273 | 0.0723 | 5.44318 | 12.93 |
| R-43#2 | 5.49273 | 0.146953 | 4.00864 | 0.24958 | 1.43264 | 5.18409 |
| R-42 | 899.682 | 0.241824 | 40.44 | 234.059 | 5.54735 | 75.675 |
| R-28 | 389.684 | 0.267933 | 34.0263 | 190.348 | 3.87 | 49.2263 |
| R-50#1 | 90.3083 | 0.148145 | 7.31739 | 19.9802 | 1.49742 | 11.0457 |
| R-50#2 | 4.89545 | 0.2 | 2.14435 | 1.24693 | 0.544261 | 2.82478 |
| R-11 | 21.8478 | 0.11732 | 5.02652 | 4.70076 | 5.35304 | 11.9074 |
| R-44#1 | 15.2316 | 0.2 | 2.28842 | 1.11807 | 0.999684 | 3.78421 |
| R-44#2 | 6.07895 | 0.2 | 2.28421 | 0.21406 | 0.630342 | 3.33789 |
| R-45#1 | 22.8647 | 0.146993 | 4.03647 | 2.36947 | 2.39859 | 6.22588 |
| R-45#2 | 12.2253 | 0.19115 | 3.46412 | 1.4288 | 0.667412 | 4.59059 |

**Table 8**
NMF*k* results for the LANL site problem presented in Table 7; the reconstruction quality *O*, silhouette width *S*, and *AIC* are estimated for number of sources $k = 1,\ldots,6$.

| *k* | *O* | *S* | *AIC* |
|---|---|---|---|
| 1 | 918514 | 1 | 989.2252 |
| 2 | 9054.305 | 1 | 538.3174 |
| 3 | 202.5992 | 0.9972838 | 175.9426 |
| 4 | 25.65629 | 0.9998287 | 0.7669993 |
| 5 | 0.8233524 | 0.9990816 | −322.6622 |
| 6 | 0.009291258 | 0.7588948 | −758.9678 |
| 7 | 1.054021e−14 | −0.3827634 | −3681.497 |
| 8 | 1.312425e−14 | −0.2217047 | −3609.816 |

**Table 9**
NMF*k* estimated concentrations of the 5 groundwater types (contaminant sources) mixed at each observation well.

| Species | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $Cr^{6+}$ | 0.651768 | 4.4077 | 5.26019 | 1603.81 | 144.144 |
| $Br^-$ | 0.190179 | 0.0748653 | 0.166503 | 0.0159012 | 0.650596 |
| $Cl^-$ | 0.233575 | 1.49177 | 20.7714 | 34.2884 | 66.4602 |
| $^3H$ | 0.0241554 | 0.0268574 | 0.0693364 | 0.569992 | 723.896 |
| $NO_3^-$ | 0.0661616 | 7.87773 | 0.368709 | 2.28238 | 11.4631 |
| $SO_4^{2-}$ | 0.390592 | 10.1767 | 18.3248 | 90.7997 | 80.515 |

**Table 10**
NMF*k* estimated mixing coefficient of the 5 groundwater types (contaminant sources) for each observation well at LANL site; the values along each row add up to 1.

| Wells | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| R-14#1 | 0.91462 | 0.0227863 | 0.0594461 | 0.00291517 | 0.000232176 |
| R-1 | 0.884697 | 0.038273 | 0.0740757 | 0.00286992 | 8.46965e−5 |
| R-33#1 | 0.82384 | 0.0795862 | 0.0933861 | 0.00259239 | 0.00059573 |
| R-33#2 | 0.887723 | 0.0321049 | 0.0767092 | 0.00317564 | 0.000287629 |
| R-15 | 0.716003 | 0.200056 | 0.0400746 | 0.00291741 | 0.0409486 |
| R-62 | 0.612985 | 0.0997541 | 0.184955 | 0.0918114 | 0.0104945 |
| R-61#1 | 0.795433 | 0.149787 | 0.0174367 | 0.00531572 | 0.0320281 |
| R-43#1 | 0.123962 | 0.644323 | 0.200355 | 0.0313342 | 2.57275e−5 |
| R-43#2 | 0.666539 | 0.162999 | 0.168036 | 0.00212731 | 0.000298321 |
| R-42 | 0.0344507 | 0.0804616 | 0.030571 | 0.53161 | 0.322907 |
| R-28 | 0.0569058 | 0.0266741 | 0.435852 | 0.217835 | 0.262733 |
| R-50#1 | 0.660187 | 0.102806 | 0.156722 | 0.0527676 | 0.0275169 |
| R-50#2 | 0.853376 | 0.0607406 | 0.0820812 | 0.00211916 | 0.00168277 |
| R-11 | 0.121542 | 0.697494 | 0.163975 | 0.0105457 | 0.00644303 |
| R-44#1 | 0.799138 | 0.11571 | 0.0751787 | 0.0084736 | 0.00150003 |
| R-44#2 | 0.816575 | 0.0858645 | 0.0944118 | 0.00289291 | 0.000255517 |
| R-45#1 | 0.609305 | 0.247331 | 0.12753 | 0.0126154 | 0.00321839 |
| R-45#2 | 0.767181 | 0.0841347 | 0.140307 | 0.00644903 | 0.00192798 |

components are elevated as well. This source might be associated with infiltration along Los Alamos Canyon. The fourth source $s_4$ has high $Cr^{6+}$ values and might be associated with infiltration along Sandia Canyon. The second and third sources (groundwater types) are associated with elevated $SO_4^{2-}$ / $NO_3^-$ and $Cl^-$ / $SO_4^{2-}$ concentrations, respectively; their origin is unknown; they might be a result of geochemical processes occurring during groundwater infiltration. The fifth source (groundwater type) represents background concentrations.

Table 10 shows the mixing coefficient of the 5 groundwater types (contaminant sources) for each observation well. Note that the mixing coefficients for each well add up to 1. The background groundwater type is predominantly detected at the upgradient wells (e.g., R-14#1, R-1, R-33) as well as in the deep screens of some of the (#2) of the two-screen wells. The fourth and fifth source sources are detected at R-42 and R-28 which are located at the centroid of the existing chromium plume in the regional aquifer. The third source (groundwater type) is predominantly detected in R-28, but, interestingly, it is not observed in R-42. The second source (groundwater type) is predominantly detected in R-43#1 and R-11 which are located in the northern portion of the site (Fig. 1).

The source concentrations estimated by NMF*k* are somewhat consistent with more complicated inverse analyses using numerical models applied to solve this problem (LANL, 2012; Vesselinov et al., 2015, 2013). In the future, the NMF*k* results will be applied as input to inverse analyses of site numerical models. In this way, instead of calibrating against all the geochemical data, the numerical models would be calibrated against the NMF*k* predicted geochemical mixtures.

## 5. Conclusions

Our analyses demonstrate the applicability of our NMF*k* approach for identification of contaminant sources based on a Non-negative Matrix Factorization (NMF) technique combined with a custom semi-supervised clustering. The NMF*k* approach was originally presented by Alexandrov and Vesselinov (2014). However, important changes and extensions were made in NMF*k* to develop an algorithm applicable for blind source identification based on geochemical data. The analyses required application of non-convex nonlinear optimization algorithm. The classical multiplicative NMF optimization algorithm (Lee and Seung, 1999) is not applicable in this case. Furthermore, additional constraints are imposed on the NMF*k* solutions. The original method relied only non-negativity constraints. Here the NMF*k* algorithm includes constraints on one of the factorized matrices (the mixing matrix in Eqs. (2) and 5) where the mixture coefficients add up to 1 for each observation point.

The inverse problem solved in the NMF*k* analysis is under-determined (ill-posed). To address this, the NMF*k* algorithm thoroughly explores the plausible inverse solutions, and seeks to narrow the set of possible solutions by estimating the optimal number of contaminant source signals needed to robustly and accurately reconstruct the observed data. This allows us to estimate the number of contaminant sources.

In the synthetic tests, we generated datasets representing unknown contaminant sources detected as a set of mixed signals (groundwater types/contamination sources) at a series monitoring wells (detectors/sensors). Using only the synthetic dataset representing the observations at the monitoring wells, we correctly identified the number of contaminant sources. We also applied NMF*k* on real-world dataset related to the LANL chromium contamination site. The results of this analysis are consistent with previous data and model analyses (LANL, 2012; Vesselinov et al., 2015, 2013).

NMF*k* allows the contaminant fields observed at a series of the detectors to be "unmixed" into a series of independent plumes with independent contamination sources. This information can be applied to guide the conceptualization of the site conditions and the design of numerical models that are set up to represent these conditions. In some cases, decoupled model analyses might be applied to independently analyze the groundwater transport of each contaminant source which can be computationally more efficient. NMF*k* results coupled with modeling analysis can yield information needed for site contaminant fate and transport predictions, hazard and risk assessments, and contaminant remediation.

It is important to note that the presented NMF analyses are following the classical BSS formulation assuming a linear mixing problem (Eq. (2)). However, since the NMF problem is solved using nonlinear minimization procedure as discussed in Section 2, the BSS problem can be expanded to account for nonlinear mixing and geochemical processes occurring in the subsurface. This will increase the number of unknowns in Eq. (2) as well as the computational complexity but as long as data are available to represent nonlinear mixing process, the BSS problem can be solved. We plan to extend our analyses to account for nonlinear mixing and geochemical processes in the future.

The presented analyses are focusing on two-dimensional (matrix) data where the data define the concentrations of a series of geochemical species as measure at a series of monitoring wells. The analyses currently ignore information about the well spatial coordinates. The incorporation of the well locations in the analyses will increase the dimensionality of the data (e.g., the problem can be five dimensional if the species concentrations depend on three spatial coordinates and time). This will require application of tensor-based factorization methods (Cichocki et al., 2009). However, currently, there are no tensor-based methods that can be applied to solve the multi-dimensional geochemical mixing problems and we are currently working to address this issue.

The possible applications of the NMF*k* approach are not limited to groundwater contamination problems. For example, NMF*k* can readily be be used to identify contaminant sources based on soil and air pollution data. NMF*k* can be applied to analyze any mixtures of ingredients. In this case, our constrained NMF*k* procedure can be applied to identify the ingredients of the sources that are mixed to produce observed mixtures.

## Acknowledgments

This research was funded by the Environmental Programs Directorate of the Los Alamos National LaboratoryLos Alamos National Laboratory. In addition, Velimir V. Vesselinov and Daniel O'Malley were supported by the DiaMonD project (An Integrated Multifaceted Approach to Mathematics at the Interfaces of Data, Models, and Decisions, U.S. Department of Energy Office of ScienceU.S. Department of Energy Office of Science, Grant #1114568711145687).

## References

Akaike, H., 2011. Akaike's information criterion. In: International Encyclopedia of Statistical Science. Springer 25–25.

Alexandrov, B.S., Vesselinov, V.V., 2014. Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization. Water Resour. Res. 50 (9), 7332–7347.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., Stratton, M.R., 2013. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 3 (1), 246–259.

Amari, S.-i., Cichocki, A., Yang, H.H., 1996. A new learning algorithm for blind signal separation. Adv. Neural Inf. Proces. Syst. 8, 757–763.

Atmadja, J., Bagtzoglou, A.C., 2001. Pollution source identification in heterogeneous porous media. Water Resour. Res. 37 (8), 2113–2125.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E., 1997. A blind source separation technique using second-order statistics. IEEE Trans. Signal Process. 45 (2), 434–444.

Bezanson, J., Karpinski, S., Shah, V.B., Edelman, A., 2012. Julia: A Fast Dynamic Language for Technical Computing. (arXiv preprint arXiv:1209.5145).

Böhlke, J., Denver, J., 1995. Combined use of groundwater dating, chemical, and isotopic analyses to resolve the history and fate of nitrate contamination in two agricultural watersheds, Atlantic Coastal Plain, Maryland. Water Resour. Res. 31 (9), 2319–2339.

Borukhov, V., Zayats, G., 2015. Identification of a time-dependent source term in nonlinear hyperbolic or parabolic heat equation. Int. J. Heat Mass Transf. 91, 1106–1113.

Cervone, G., Franzese, P., Keesee, A.P., 2010. Algorithm quasi-optimal (aq) learning. Wiley Interdiscip. Rev. Comput. Stat. 2 (2), 218–236.

Chan, C.W., Huang, G.H., 2003. Artificial intelligence for management and control of pollution minimization and mitigation processes. Eng. Appl. Artif. Intell. 16 (2), 75–90.

Chiles, J.-P., Delfiner, P., 2009. Geostatistics: Modeling Spatial Uncertainty. vol. 497 John Wiley & Sons.

Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.-i., 2009. Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. John Wiley & Sons ISBN 978-0-470-74666-0.

Deutsch, W.J., Siegel, R., 1997. Groundwater Geochemistry: Fundamentals and Applications to Contamination. CRC Press.

Diday, E., Simon, J., 1980. Clustering analysis. In: Digital Pattern Recognition. Springer, pp. 47–94.

Drucker, H., Wu, D., Vapnik, V.N., 1999. Support vector machines for spam categorization. IEEE Trans. Neural Netw. 10 (5), 1048–1054.

Dunning, I., Huchette, J., Lubin, M., 2015. JuMP: A Modeling Language for Mathematical Optimization. (arXiv preprint arXiv:1508.01982).

Facchinelli, A., Sacchi, E., Mallen, L., 2001. Multivariate statistical and GIS-based approach to identify heavy metal sources in soils. Environ. Pollut. 114 (3), 313–324.

Fetter, C.W., Fetter, C., 1999. Contaminant Hydrogeology. vol. 500 Prentice Hall New Jersey.

Fischler, M.A., Elschlager, R.A., 1973. The representation and matching of pictorial structures. IEEE Trans. Comput. 100 (1), 67–92.

Gelhar, L.W., 1993. Stochastic Subsurface Hydrology. Prentice-Hall.

Guan, J., Aral, M.M., Maslia, M.L., Grayman, W.M., 2006. Identification of contaminant sources in water distribution systems using simulation-optimization method: case study. J. Water Resour. Plan. Manag. 132 (4), 252–262.

Hamdi, A., Mahfoudhi, I., 2013. Inverse source problem in a one-dimensional evolution linear transport equation with spatially varying coefficients: application to surface water pollution. Inverse Prob. Sci. Eng. 21 (6), 1007–1031.

Harman, H.H., 1976. Modern Factor Analysis. University of Chicago Press.

Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J.M., Fernandez, L., 2000. Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Res. 34 (3), 807–816.

Herault, J., Jutten, C., 1986. Space or time adaptive signal processing by neural network models. In: Neural Networks for Computing. vol. 151. American Institute of Physics Publishing, pp. 206–211.

Jolliffe, I., 2002. Principal Component Analysis. Wiley Online Library.

Khalil, A., Almasri, M.N., McKee, M., Kaluarachchi, J.J., 2005. Applicability of statistical learning algorithms in groundwater quality modeling. Water Resour. Res. 41 (5).

Knudson, E.J., Duewer, D.L., Christian, G.D., Larson, T.V., 1977. Application of factor analysis to the study of rain chemistry in the Puget Sound region. In: Chemometric: Theory and Application. ACS Symposium Series, Washington, DC, pp. 80–116.

LANL, 2012. Phase II Investigation Report for Sandia Canyon. In: Tech. Rep. LANL.

Lapworth, D., Baran, N., Stuart, M., Ward, R., 2012. Emerging organic contaminants in groundwater: a review of sources, fate and occurrence. Environ. Pollut. 163, 287–303.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–791.

Mamonov, A.V., Tsai, Y.R., 2013. Point source identification in nonlinear advection-diffusion-reaction systems. Inverse Prob. 29 (3), 035009.

Manca, G., Cervone, G., 2013. The case of arsenic contamination in the Sardinian Geopark, Italy, analyzed using symbolic machine learning. Environmetrics 24 (6), 400–406.

Michalak, A.M., Kitanidis, P.K., 2004. Estimation of historical groundwater contaminant distribution using the adjoint state method applied to geostatistical inverse modeling. Water Resour. Res. 40 (8).

Murray-Bruce, J., Dragotti, P.L., 2014. Spatio-temporal sampling and reconstruction of diffusion fields induced by point sources. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 31–35.

Neupauer, R.M., Borchers, B., Wilson, J.L., et al., 2000. Comparison of inverse methods for reconstructing the release history of a groundwater contamination source. Water Resour. Res. 36 (9), 2469–2475.

Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5 (2), 111–126.

Pang-Ning, T., Steinbach, M., Kumar, V., 2006. Introduction to Data Mining. Addison-Wesley, pp. 769 ISBN 978-0321321367.

Rasekh, A., Brumbelow, K., 2012. Machine learning approach for contamination source identification in water distribution systems. In: World Environmental and Water Resources Congress. Palm Springs, CA.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Scholkopft, B., Mullert, K.-R., 1999. Fisher discriminant analysis with kernels. Neural Netw. Signal Process. IX 1 (1), 1.

Shrestha, S., Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji River basin, Japan. Environ. Model Softw. 22 (4), 464–475.

Tariq, S.R., Shah, M.H., Shaheen, N., Jaffar, M., Khalique, A., 2008. Statistical source identification of metals in groundwater exposed to industrial contamination. Environ. Monit. Assess. 138 (1-3), 159–165.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1 (Jun), 211–244.

Vengosh, A., Jackson, R.B., Warner, N., Darrah, T.H., Kondash, A., 2014. A critical review of the risks to water resources from unconventional shale gas development and hydraulic fracturing in the United States. Environ. Sci. Technol. 48 (15), 8334–8348.

Vesselinov, V.V., 0'Malley, D., Katzman, D., 2015. Model-assisted Decision Analyses Related to a Chromium Plume at Los Alamos National Laboratory. In: WMSYM2015, Phoenix, Arizona, USA.

Vesselinov, V.V., Broxton, D., Birdsell, K., Reneau, S., Harp, D.R., Mishra, P.K., Katzman, D., Goering, T., Vaniman, D., Longmire, P., Fabryka-Martin, J., Heikoop, J., Ding, M., Hickmott, D., Jacobs, E., 2013. Data and Model-driven Decision Support for Environmental Management of a Chromium Plume at Los Alamos National Laboratory. In: WMSYM2013, Phoenix, Arizona, USA, . http://www.wmsym.org/archives/2013/papers/13264.pdf.

Vijayakumar, S., Schaal, S., 2000. Locally weighted projection regression: incremental real time learning in high dimensional space. In: Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 1079–1086.

Wächter, A., 2002. An Interior Point Algorithm for Large-scale Nonlinear Optimization with Applications in Process Engineering. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA Ph.D. thesis.

Wächter, A., Biegler, L.T., 2005. Line search filter methods for nonlinear programming: motivation and global convergence. SIAM J. Optim. 16 (1), 1–31.

Wächter, A., Biegler, L.T., 2006. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. (1436-4646) 106 (1), 25–57.

Wagner, B.J., 1992. Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling. J. Hydrol. 135 (1), 275–303.

Yegnanarayana, B., 2009. Artificial Neural Networks. PHI Learning Pvt. Ltd.