HUG model: an interaction point process for Bayesian detection of multiple sources in groundwaters from hydrochemical data

Christophe Reype, Radu S. Stoica, Antonin Richard, Madalina Deaconu

HUG model: an interaction point process for Bayesian detection of multiple sources in groundwaters from hydrochemical data

C. Reype[1], R. S. Stoica[1], A. Richard[2], and M. Deaconu[1]

[1]Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France
[2]Université de Lorraine, CNRS, GeoRessources, F-54000 Nancy, France

July 29, 2022

Abstract

This paper presents a new interaction point process that integrates geological knowledge for the purpose of automatic sources detection of multiple sources in groundwaters from hydrochemical data. The observations are considered as spatial data, that is a point cloud in a multi-dimensional space of hydrogeochemical parameters. The key hypothesis of this approach is to assume the unknown sources to be the realisation of a point process. The probability density describing the sources distribution is built in order to take into account the multi-dimensional character of the data and specific physical rules. These rules induce a source configuration able to explain the observations. This distribution is completed with prior knowledge regarding the model parameters distributions. The composition of the sources is estimated by the configuration maximising the joint proposed probability density. The method was first calibrated on synthetic data and then tested on real data from hydrothermal systems.

1 Introduction

The analysis of hydrochemical data can be used to build conceptual and quantitative models of fluid and mass transfer in the sub-surface and the Earth's crust [Faure, 1997,Yardley and Bodnar, 2014,Ingebritsen et al., 2006]. The composition of many groundwaters is controlled by mixing of two or more water sources (Figure 1). In such cases, the analysis of hydrochemical data includes detecting the sources involved with mixing (*i.e.* number and composition) and estimating their contribution to the data (respectively "inverse analysis" and "forward analysis").

The sources (also refer to as end-members) are mixed together at variable proportions to result in the samples also called mixing terms. The values for each hydrochemical parameter are determined from direct sampling (from bore holes or springs) or from fluid inclusions. Hydrochemical parameters considered are the concentration of ions or molecules, isotopic composition or ratios of hydrochemical parameters. The data are seen as spatial data: a sample is represented by a point in the data space with coordinates being the value of each hydrochemical parameter. Hence mention of position and distance in this paper will not refer to sample location but composition (*e.g.* measurement of all hydrochemical parameters) and difference in composition respectively. In the context of fluid mixing, a sample is considered as a barycenter of the sources: a data point d_j is a result of a mixing between the pattern of sources $\mathbf{s} = \{s_1, \dots, s_n\}$ if

$$d_j = \sum_{i=1}^{n} \gamma_{(j);i} s_i, \tag{1}$$

with $0 \leq \gamma_{(j);i} \leq 1$ the contribution of the source i in the point j.

If the sources are known, the contribution of the sources can be estimated by using either Bayesian mixing models [Longman et al., 2018, Arendt et al., 2015, Carrera et al., 2004, Skuce et al., 2015, Parnell et al., 2010,Phillips and Gregg, 2001,Lajaunie et al., 2020,Tipton et al., 2022] or a likelihood uncertainty estimation that relies on End Member Mixing Analysis (EMMA) [Delsman et al., 2013].

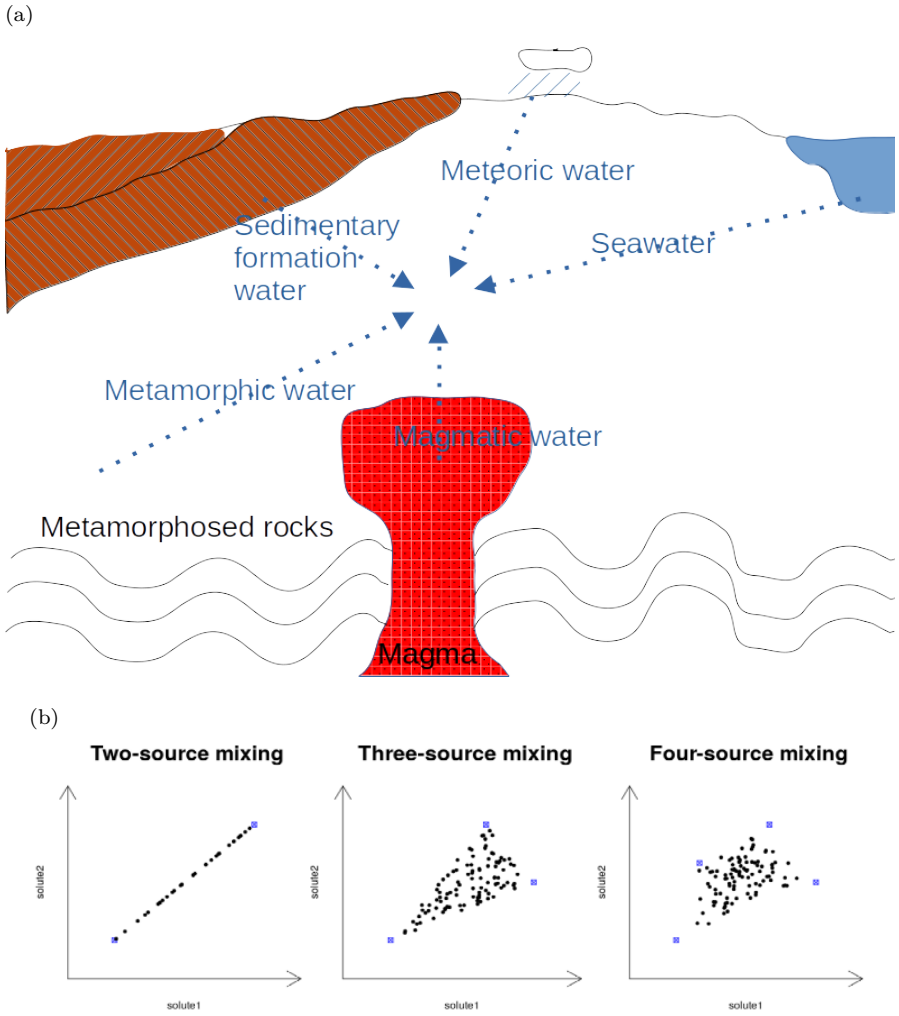


Figure 1: **Schematic view of fluid mixing scenarios in the Earth's crust** (a) Conceptual cross section of the Earth's upper continental crust showing the main sources of surface and sub-surface waters which contribute to the composition of groundwaters in hydrothermal systems through mixing processes [modified from Robb, 2005]. Similar water mixing processes also occur in surface and shallow subsurface environments, potentially involving other water sources. Figures (b) are binary diagrams showing typical mixing scenarios in the space of hydrochemical parameters (solute1, solute2). Blue squares represent the sources/end-members. Black dots represent the samples/mixing terms.

The already existing solutions, proposed to solve the sources detection problem, try to tackle three challenges: the multi-dimensional character of the data, the unknown number of sources and physical constraints. The first constraint is to minimise the number of sources. The second constraint is to select sources that explain the data (*i.e.* the convex hull of the sources tends to enclose the data). The third constraint is to consider that the data are representative of the mixing system (*i.e.* the convex hull of the sources is outlined by the data). The last constraint is to consider that the composition of the sources are is significantly different from another.

To the best of our knowledge, the existing methods presented in the literature do not take into account all these aspects together. Source detection is done either manually or in a supervised statistical analysis. A principal component analysis (PCA) is sometimes used to guide the choice of the number of sources [Christophersen and Hooper, 1992]. Based on this, the composition of the sources can be estimated by an "end-member mixing analysis" (EMMA) [Weltje, 1997]. When the number of sources is known, a more geometrical method can be used: the sources can be estimated by the vertex of the smallest triangle (in term of area) that contains the data in the case of three sources in a two dimensional data space [Pinti et al., 2020].

This paper develops a new Bayesian method of sources detection in hydrochemical data. This procedure is inspired by pattern detection methodologies used in image analysis, animal epidemiology and astronomy [Stoica et al., 2004,Stoica et al., 2007a,Stoica et al., 2005b,Stoica et al., 2007b]. It has the advantages to be unsupervised and to take into account simultaneously the previously mentioned physical constraints. Furthermore, the model considers pattern of sources with no condition on the maximum number of sources. Conditionally to the parameters of the model, the probabilistic source model considered is a Gibbs point process that controls the sources distribution in the data space. The set of sources is estimated by the points configuration that maximises the joint probability density controlling the sources and the parameters distributions. The model presented in this paper is called HUG model in reference to the way that the sources enclose the data. The optimisation procedure is implemented via a simulated annealing procedure, hence avoiding local maximum. The sampling Markov chain Monte Carlo (MCMC) algorithm at the basis of the simulated annealing procedure is achieved by a Metropolis-Hastings within Gibbs sampler. This allows to deal with the multi-dimensional aspect of the problem.

The conditions for using the HUG model are as follows: the data sets are the results of a conservative mixing (*i.e.* no chemical reaction affects the considered hydrochemical parameters during the mixing process). The composition of the sources are supposed the same or at least not significantly different for each data points. It is noteworthy that the hydrochemical parameters considered should not induce any curvature in the mixing trends projected on binary plots [Langmuir et al., 1978]. The model does not consider the incertitude of the data. Moreover the method presented in the following depends on the quality of the data: outliers should be avoided.

The structure of the paper is as it follows. Fundamental notions on point processes, their properties and simulation algorithms are presented in Section 2. Section 3 is dedicated to the description of the HUG model. The proposed solutions exhibit two main components. The first component is represented by a Gibbs point process controlling the source distribution. The second component is Metropolis within Gibbs sampler that allows to sample from the model while taking into account the multi-dimensional character of the data. Inference procedures are also presented. Section 4 presents the application of the method on synthetic data. The application on the synthetic data allows to tune the parameters priors. The application on real data from hydrothermal systems permits to test and to analyse the method's performances.

2 Point processes : definition, properties and simulation

2.1 Point processes

Let $(S, \mathcal{B}, \nu)$ be a measure space, where S is a compact subset of $\mathbb{R}^d$ of strictly positive Lebesgue measure $0 < \nu(S) < \infty$ and $\mathcal{B}$ the associated Borel σ-algebra of subsets of S. For $n \in \mathbb{N}$ let S_n be the set of all unordered configurations $\mathbf{s} = \{s_1, s_2, \ldots, s_n\}$ of n not necessarily distinct points $s_i \in S$. Let us consider

the configuration space $\Omega = \cup_{n=0}^{\infty} S_n$ equipped with the σ-algebra $\mathcal{F}$ generated by the mappings

$$\{s_1, s_2, \ldots, s_n\} \longmapsto \sum_{i=1}^{n} \mathbf{1}\{s_i \in B\}$$

counting the number of points in Borel sets $B \in \mathcal{B}$. A point process on S is a measurable map from a probability space into $(\Omega, \mathcal{F})$. For introductory material on point processes we refer the reader to the textbooks by [van Lieshout, 2000, Møller and Waagepetersen, 2003].

Maybe the most known point process is the homogeneous Poisson point process constructed as it follows. First, the number of points n in a configuration is chosen according to a Poisson law of parameter $\rho\nu(S)$ with $\rho > 0$ a positive constant named intensity. Then, the n points are spread uniformly independently in S. We write $X \sim Poisson(S, \rho)$.

The process $Poisson(S, 1)$, or more specifically its measure, is often considered as a reference measure, μ *i.e.* $(\forall B \in \mathcal{B} : \mu(B) = \rho\nu(B))$, for more elaborate models. For instance, the inhomogeneous Poisson point process driven by the intensity function $\rho : S \to \mathbb{R}^+$ has the probability measure

$$\forall F \in \mathcal{F} : \mathbb{P}(X \in F) = \sum_{n=0}^{\infty} \frac{\exp[-\nu(S)]}{n!} \int_S \cdots \int_S \mathbf{1}_F\{s_1, \cdots, s_n\} \Big(\prod_{i=1}^{n} \rho(s_i) \Big) \, d\nu(s_1) \cdots d\nu(s_n).$$

In this case, the process is spread in S independently according to the probability density $\rho(\cdot)/\int_S \rho(s) d\nu(s)$. Clearly, the probability density of this process with respect to the unit intensity stationary Poisson process is given by

$$p(\mathbf{s}) = \zeta \prod_{i=1}^{n(\mathbf{s})} \rho(s_i),$$

with $\zeta = \exp\left[\nu(S) - \int_S \rho(s) d\nu(s)\right]$ the normalising constant and $n(\mathbf{s})$ the cardinal of $\mathbf{s}$. The fact that their distribution is entirely known, makes Poisson processes extremely interesting candidates for numerous modelling approaches. Nevertheless, the independence assumption implies that no interactions of points are considered.

Gibbs points processes are models that take into account interactions of points by means of probability density with respect to the reference measure μ. The general form of this probability density is

$$p(\mathbf{s}) = \zeta \exp[-U(\mathbf{s})], \tag{2}$$

with $U(\mathbf{s})$ the energy function specifying the points interactions in a configuration. Still, in this case the normalising constant $\zeta^{-1} = \int_\Omega \exp[-U(\mathbf{s})] d\mu(\mathbf{s})$ is no more available in analytical closed form.

There is a lot of freedom in specifying energy functions provided the resulting probability density integrates to 1. This is ensured if the model is locally stable, that is there exists $\Lambda \in \mathbb{R}^+$ such that

$$\frac{p(\mathbf{s} \cup \{\eta\})}{p(\mathbf{s})} \leq \Lambda, \quad \forall \mathbf{s} \in \Omega, \eta \in S. \tag{3}$$

There exist less restrictive conditions that ensure the integrability of point process [Ruelle, 1999]. The preference for locally stable models (3) is due to the good convergence properties induced to the corresponding simulation algorithms [van Lieshout, 2000, Møller and Waagepetersen, 2003].

2.2 Simulation

Sampling from Gibbs point process densities (2) is not trivial. This is due to the fact that the normalising constant ζ is not available in analytical closed form. The adopted solutions within this context are given by Markov chain Monte Carlo (MCMC) strategies. Among them let us mention : spatial birth-and-death processes, perfect sampling methods, Metropolis-Hastings algorithms, etc. The interested reader

may refer to [Baddeley et al., 2016, van Lieshout, 2000, Møller and Waagepetersen, 2003, van Lieshout, 2000, Geyer, 1999] for details and thorough mathematical presentations.

The principle behind the MCMC methods is to simulate a Markov chain that has as equilibrium distribution, the probability distribution of interest. In our case, this is

$$\pi(A) = \int_A p(\mathbf{s})\mu(d\mathbf{s}), A \in \mathcal{F}. \tag{4}$$

The Metropolis-Hastings (MH) algorithm for point processes, implements a Markov chain whose transition kernel is built using three types of moves or transitions: birth or adding a point to the current configuration, death or deleting a point from the current configuration and change or changing a point from the current configuration into a new point. Let $p_b, p_d, p_c \in [0, 1]$, with $p_b + p_d + p_c \leq 1$, be the probability of respectively birth, death and change. Moreover let $b(\mathbf{s}, \eta)$ and $c(\mathbf{s}, \eta, \zeta)$ be distributions of points and $d(\mathbf{s}, \eta)$ and $q(\mathbf{s}, \eta)$ be probabilities of choosing a point η in the current configuration $\mathbf{s}$. With probability p_b the event birth is selected: a new point η is generated according to $b(\mathbf{s}, \eta)$. With probability p_d the event death is selected: a point η selected in the configuration according to $d(\mathbf{s}, \eta)$ is deleted. With probability p_c the event change is selected: a point η selected in the configuration according to $q(\mathbf{s}, \eta)$ is changed into a point ζ generated according to $c(\mathbf{s}, \eta, \zeta)$. This transition kernel is synthesised within an Update procedure that is to be iterated in order to obtain the desired samples.

Algorithm MH : $\mathbf{y}$ = Update ($\mathbf{s}$)

1) Choose a transition type according to p_b, p_d and p_c, such that $p_b + p_d + p_c \leq 1$.
2) If a "birth" is chosen, generate a new point η according with $b(\mathbf{s}, \eta)$. Accept the new configuration $\mathbf{y} = \mathbf{s} \cup \{\eta\}$ with probability

$$\alpha(\mathbf{s}, \mathbf{s} \cup \{\eta\}) = \min\{1, r(\mathbf{s}, \eta)\},$$

with

$$r(\mathbf{s}, \eta) = \frac{p_d d(\mathbf{s} \cup \{\eta\}, \eta)}{p_b b(\mathbf{s}, \eta)} \frac{p(\mathbf{s} \cup \{\eta\})}{p(\mathbf{s})}. \tag{5}$$

3) If a "death" is chosen, select a candidate η to be deleted from $\mathbf{s}$ according with $d(\mathbf{s}, \eta)$. Accept the new configuration $\mathbf{y} = \mathbf{s} \setminus \{\eta\}$ with probability

$$\alpha(\mathbf{s}, \mathbf{s} \setminus \{\eta\}) = \min\{1, 1/r(\mathbf{s} \setminus \{\eta\}, \eta)\}.$$

4) If a "change" is chosen, select a candidate η from $\mathbf{s}$ according to $q(\mathbf{s}, \eta)$ and change it into a new candidate ζ according to $c(\mathbf{s}, \eta, \zeta)$. Accept the new configuration $\mathbf{y} = \mathbf{s} \setminus \{\eta\} \cup \{\zeta\}$ with probability

$$\alpha(\mathbf{s}, \mathbf{s} \setminus \{\eta\} \cup \{\zeta\}) = \min\{1, \min\{1, r(\mathbf{s}, \eta, \zeta)\}\},$$

with

$$r(\mathbf{s}, \eta, \zeta) = \frac{q(\mathbf{s} \setminus \{\eta\} \cup \{\zeta\}, \eta)}{q(\mathbf{s}, \eta)} \frac{c(\mathbf{s} \setminus \{\eta\} \cup \{\zeta\}, \zeta, \eta)}{c(\mathbf{s}, \eta, \zeta)} \frac{p(\mathbf{s} \setminus \{\eta\} \cup \{\zeta\})}{p(\mathbf{s})}. \tag{6}$$

In order to prevent the Markov chain of pathological cases, the acceptance ratio in (5) should be non-zero. Hence, $\alpha(\mathbf{s}, \mathbf{s} \cup \{\eta\}) = 0$ whenever $p(\mathbf{s} \cup \{\eta\}) = 0$. Maybe the most adopted birth and death proposals are the uniform choices

$$b(\mathbf{s}, \eta) = \frac{\mathbf{1}\{\eta \in S\}}{\nu(S)}$$

and

$$d(\mathbf{s}, \eta) = \frac{\mathbf{1}\{\eta \in \mathbf{s}\}}{n(\mathbf{s})},$$

respectively. The uniform choice is also adopted for the event change. Hence $q(\mathbf{s}, \eta) = d(\mathbf{s}, \eta)$ and the new point is generated in the ball centred in η and of radius $r_c \in \mathbb{R}^+$ noted $B(\eta, r_c)$:

$$c(\mathbf{s}, \eta, \zeta) = \frac{\mathbf{1}\{\zeta \in B(\eta, r_c)\}}{B(\eta, r_c)}.$$

These choices together with the local stability (3) guarantee the geometric ergodicity, Harris recurrence and $\phi-$irreducibility of the Markov chain simulated using this Metropolis-Hastings algorithm [Geyer, 1999, van Lieshout, 2000, Møller and Waagepetersen, 2003].

2.3 Inference

In the following, we assume that we are in the possession of a well defined source model $p(\mathbf{s})$ which is a point process density and of an appropriate sampling algorithm able to sample from it.

Under these circumstances, maximisation of the probability density can be achieved via a simulated annealing algorithm based on the previously described MH dynamics. This algorithm iteratively draws samples at a temperature $T \in \mathbb{R}^+$ from $p(\mathbf{s})^{1/T}$ while $T \to 0$. A logarithmic cooling schedule for the temperature T guarantees the convergence of the simulated annealing towards the uniform distribution of the configurations sub-space that maximises $p(\mathbf{s})$ [Stoica et al., 2005a].

The solution to the optimisation problem is not guaranteed by a unique configuration of points. Hence, in order to get more robust results, averaging may be useful. This can be achieved through the computation of level sets. Let us consider a random set, such as a point process X, and the probabilities

$$p_{r_l}(s) = \mathbb{P}(X \cap B(s, r_l) \neq \emptyset)$$

where $B(s, r_l)$ is the ball centred in $s \in S$ with given radius r_l. A level is defined by all the points in S such that

$$l(\lambda) = \{s \in S : p_{r_l}(s) \geq \lambda\}.$$

Clearly, the sets $l(\lambda)$ are quantiles of the random set X. If the random set X is the sources configuration governed by the model $p(\mathbf{s})$ the estimators of the level sets may indicate the regions in S that are visited by the model with a probability higher than λ. The derivation and the properties of these level sets estimators are given in [Heinrich et al., 2012].

3 The HUG model

The data considered are the measurements of K hydrochemical parameters of m samples denoted $\mathbf{d} = \{d_1, d_2, \ldots, d_m\}$. The data point numbered j $(1 \leq j \leq m)$ is placed in the data space with coordinates $d_j = (d_{(j);1}, \ldots, d_{(j);K})$, where $d_{(j);k} \in \mathbb{R}$ is the measurement of the hydrochemical parameter numbered k $(1 \leq k \leq K)$. Hence, the data set is a point cloud made of m points (or samples) in a K-dimensional space of finite volume.

In this space, the pattern of sources is unknown. Still, it is related to the set of data points. The different aspects of the relation between the data and the unknown sources can be synthesised by the following hypotheses:

(a) the data points originating from a mixture of sources should be rather close to the sources

(b) the data points are enclosed within the convex hull given by the positions of the sources

(c) the number of sources is not known but it should be controlled or minimised in a certain sense

(d) the composition of the sources should be significantly different from each other.

By hypothesis (a) the composition of the sources are in a bounded space: without it, every data set can be explained by three sources placed in infinity. The hypothesis (b) is a physical consequence of mixing with mass conservation [Faure, 1997]. Hypothesis (c) and (d) tend to make the problem easier by reducing the possible sources. In practice, the number of sources involved in a mixing is supposed to be less than 10 [Faure, 1997, Yardley and Bodnar, 2014].

The key idea of our work is to build a Gibbs point processes that governs the sources distribution n the data space. The energy function of the process integrates the previous hypotheses. Since, these assumptions specify relations between the data and the unknown sources, but also interactions among

sources only (*i.e.* interaction between points of a pattern of points), the probability density of the pattern of sources $\mathbf{s}$ can be written as follows

$$p_{\mathbf{d}}(\mathbf{s}|\theta) = \frac{\exp\left[-U(\mathbf{s}|\theta)\right]}{Z(\theta)} = \frac{\exp\left[-U_{\mathbf{d}}(\mathbf{s}|\theta) - U_i(\mathbf{s}|\theta)\right]}{Z(\theta)}, \tag{7}$$

where $U_{\mathbf{d}}(\mathbf{s}|\theta)$ is the data term that locates the sources in the data space - hypotheses (a) and (b), $U_i(\mathbf{s}|\theta)$ is the interaction term that manages the sources interaction - hypotheses (c) and (d) and $Z(\theta)$ the normalising constant. The sum of the data and interaction terms gives the total energy function U.

In the following, we specify the model (7). For the sake of simplicity, here we assume that $K = 2$, that is the hydrochemical space is a finite surface. The generalisation to a multidimensional volume will be considered afterwards.

3.1 Data energy function

The data term $U_{\mathbf{d}}(\mathbf{s}|\theta)$ controls the positioning of the sources with respect to the observed data points. This term allows the model to detect source patterns while taking into account hypotheses (a) and (b).

Having in mind (a), let us consider the ratio between the area of the convex hull of the sources $g(\mathbf{s})$ and the area of the convex hull of the data $g(\mathbf{d})$. More specifically we consider the statistic $g(\mathbf{s}, \mathbf{d})$:

$$g(\mathbf{s}, \mathbf{d}) = \left| \frac{g(\mathbf{s})}{g(\mathbf{d})} - 1 \right| \tag{8}$$

with $|\cdot|$ the absolute value function.

If the data and the sources convex hulls tend to have equal surfaces, the statistics value should be close to 0. Furthermore, $g(\mathbf{s}, \mathbf{d})$ is bounded, since the observation domain is bounded. The numerical computation of (8) can be performed via the Andrew's monotone chain convex hull algorithm [Andrew, 1979].

The hypothesis (b) is considered by the following statistic

$$n_e(\mathbf{s}, \mathbf{d}) = 1 - \frac{n_{expl}(\mathbf{s}, \mathbf{d})}{m} \tag{9}$$

where $n_{expl}(\mathbf{s}, \mathbf{d})$ is the number of points explained by the pattern of sources, (*i.e.* the number of points of $\mathbf{d}$ inside the convex hull of $\mathbf{s}$) and m the total number of points. Whenever the sources tend to explain all the points, the statistic (9) is close to 0.

Hence the data energy function is:

$$U_{\mathbf{d}}(\mathbf{s}|\theta) = \theta_1 g(\mathbf{s}, \mathbf{d}) + \theta_2 n_e(\mathbf{s}, \mathbf{d}), \tag{10}$$

with $\theta_1, \theta_2 \in \mathbb{R}^+$ the model parameters controlling the strength of each statistic and so the weight of hypothesis (a) and (b) respectively.

3.2 Interaction energy function

The term $U_i(\mathbf{s}|\theta)$ controls the sources interactions and it does not depend on the data $\mathbf{d}$. This term allows to take into account the hypotheses (c) and (d).

The number of sources $n(\mathbf{s})$ in a configuration controls the (c) hypothesis, while the proximity of sources required by the hypothesis (d) is controlled by $n_r(\mathbf{s})$. This last statistics represents the number of pairs of sources situated within a pre-fixed distance r from each other.

Within this context the interaction energy function is:

$$U_i(\mathbf{s}|\theta) = \theta_3 n(\mathbf{s}) + \theta_4 n_r(\mathbf{s}), \tag{11}$$

with $\theta_3, \theta_4 \in \mathbb{R}^+$ the model parameters controlling the weight of hypothesis (c) and (d) respectively.

3.3 Source estimator

The HUG model is the Gibbs point process defined by the energy functions (10) and (11).

Assuming knowledge related to parameters is available through the prior $p(\theta)$, the joint distribution is written as

$$p_{\mathbf{d}}(\mathbf{s}, \theta) = p_{\mathbf{d}}(\mathbf{s}|\theta)p(\theta).$$

Within this context, the unknown source pattern together with its parameters are estimated by maximising (3.3):

$$\widehat{(\mathbf{s}, \theta)} = \arg\max_{\Omega\times\Theta} p_{\mathbf{d}}(\mathbf{s}, \theta) = \arg\max_{\Omega\times\Theta} p_{\mathbf{d}}(\mathbf{s}|\theta)p(\theta) \tag{12}$$

with the configuration space Ω and the parameter space Θ a compact region in $\mathbb{R}^4$. The computation of (12) can be achieved via a simulated annealing algorithm.

3.4 General case: $K \geq 2$

The observed data sets contain a number of hydrochemical parameters greater than two. The construction of a solution in dimension K by the generalisation of (12) is mathematically possible. Still, this straightforward approach may lead to rather heavy computations, under hypotheses difficult to test and to control.

Here, an alternative solution is preferred. Assuming the data set contains K hydrochemical parameters, let us consider all the planes obtained by taking pairs of hydrochemical parameters. The total number of different planes is $L = K(K-1)/2$ (considering hydrochemical parameter i_1 and i_2 is the same as considering i_2 and i_1). The solution we propose is to complete the distribution $p_{\mathbf{d}}(\mathbf{s}, \theta)$ with an auxiliary variable that selects such a plane, hence allowing operations only in spaces of dimension two. This auxiliary variable is discrete and finite taking values in $V = \{1, 2, \ldots, L\}$. It is governed by the prior distribution $p(v)$.

Let us consider the following model

$$p_{\mathbf{d}}(\mathbf{s}, \theta, v) = p_{\mathbf{d}}(\mathbf{s}, \theta|v)p(v) = p_{\mathbf{d}}(\mathbf{s}|\theta, v)p(\theta)p(v).$$

The conditional distribution is given by

$$p_{\mathbf{d}}(\mathbf{s}|\theta, v) \propto \exp[-U(\mathbf{s}|\theta, v)], \tag{13}$$

with energy function

$$U(\mathbf{s}|\theta, v) = U_{\mathbf{d}}(\mathbf{s}|\theta, v) + U_i(\mathbf{s}|\theta, v).$$

The data energy expression writes

$$U_{\mathbf{d}}(\mathbf{s}|\theta, v) = \sum_{l=1}^{L} U_{\mathbf{d}}^{(l)}(x|\theta)\mathbf{1}\{v = l\}$$

where each element in the sum is

$$U_{\mathbf{d}}^{(l)}(x|\theta) = \theta_1 g^{(l)}(\mathbf{s}, \mathbf{d}) + \theta_2 n_e^{(l)}(\mathbf{s}, \mathbf{d}), \quad l = 1, \ldots, L, \tag{14}$$

with $g^{(l)}(\mathbf{s}, \mathbf{d})$ and $n_e^{(l)}(\mathbf{s}, \mathbf{d})$ the data energy statistics corresponding to the parametric plane numbered l.

The energy interaction is developed in analogous manner:

$$U_i(\mathbf{s}|\theta, v) = \sum_{l=1}^{L} U_i^{(l)}(x|\theta)\mathbf{1}\{v = l\}$$

where each element in the sum is

$$U_i^{(l)}(x|\theta) = \theta_3 n(\mathbf{s}) + \theta_4 n_r^{(l)}(\mathbf{s}), \quad l = 1, \ldots, L, \tag{15}$$

with $n_r^{(l)}(\mathbf{s}, \mathbf{d})$ the interacting pairs of sources corresponding to the plane numbered l.

This framework allows to propose as joint estimator for the source pattern, model parameters and planes selector

$$\widehat{(\mathbf{s}, \theta, v)} = \arg \max_{\Omega \times \Theta \times V} p_{\mathbf{d}}(\mathbf{s}, \theta, v) = \arg \max_{\Omega \times \Theta \times V} p_{\mathbf{d}}(\mathbf{s}|\theta, v) p(\theta) p(v). \tag{16}$$

The proposed construction is a mixture of two-dimensional processes. This is not the only construction possible. It was preferred here because a sampling algorithm needed for solving (16) may be naturally proposed.

3.5 Simulation of the HUG model and implementation of the simulated annealing algorithm

The HUG model is locally stable [van Lieshout, 2000, Møller and Waagepetersen, 2003]. Hence it can be sampled using the MH algorithm presented in Section 2. There is no information available regarding the convexity of $p_{\mathbf{d}}(\mathbf{s}, \theta, v)$. The computation of (16) requires a global optimisation procedure. For this purpose, a simulated annealing algorithm may be implemented.

This algorithm requires sampling from $p(\mathbf{s}, \theta, v)$ given by (13). This can be done via a MH within Gibbs dynamics. The sampling of $p(\theta)$ and $p(v)$ does not exhibit any particular difficulty if these priors are chosen with care. Once a parameter and a hydrochemical plane are chosen by their priors, respectively, sampling from $p(\mathbf{s}|\theta, v)$ is just the simulation of two dimensional HUG model. Therefore this step can be performed using the MH algorithm in Section 2.

Regarding the cooling schedule, the authors in [Stoica et al., 2005a] proved that a logarithmic scheme guarantees the convergence of the simulated annealing based on MH algorithms for point processes. For speeding up the computation time, here preference is given to the polynomial scheme

$$T_{n+1} = cT_n, \quad c \in]0, 1[.$$

The general algorithm is presented below

Algorithm SA : Fix $p(\theta)$ and $p(v)$. Choose the initial configuration $\mathbf{s}_0$. The initial temperature is T_1, the total number of iterations is N, G the number of applications of the Gibbs dynamics and M the number of applications of the MH algorithm.

1. For k=1 to N do
   - $\theta_k \sim [p(\theta)]^{1/T_k}$
   - for g=1 to G do
     (a) $v_k \sim [p(v)]^{1/T_k}$
     (b) $\mathbf{s}_k \sim [p(\mathbf{s}|\theta_k, v_k)]^{1/T_k}$. This step is achieved by calling MH$(\mathbf{s}_{k-1}, T_k)$ successively M times.
   - $T_{k+1} = cT_k$
2. Return $(\mathbf{s}_N, \theta_N, v_N)$.

4 Application

This section demonstrates the proposed method application. First, the normalisation of the data-set and the choice of the variable needed for each of the presented algorithms are explained. Then the model is applied on synthetic data sets. The first synthetic data set allows us to evaluate the results of the model when the real sources are known and visible on each projection plane. The second synthetic data set exhibit a specificity: 4 different sources are mixed in 3 dimensions but these sources are positioned such as only 3 sources are visible on each plane. Finally, the HUG model is then applied on real data sets.

4.1 Parameters set-up

For all the data sets, a normalisation procedure is built such that the data and the simulated sources are in the unit hyper-cube $W = [0.0, 1.0]^K$. The normalisation is made dimension by dimension. More precisely, for each dimension we define the window of the range of values that a source can take and, by a linear transformation, convert it into $[0.0, 1.0]$. This range is set for each dimension $k \in [1, \ldots, K]$ to $(\min_j(d_{(j),k}) - \delta_k, \max_j(d_{(j),k}) + \delta_k)$ where δ_k is a threshold set by the user. Here we take $\delta_k = \max_j(d_{(j),k}) - \min_j(d_{(j),k})$. Regarding the interaction radius needed by the HUG model, the value $r = 0.01$ is chosen for each projection plane. If available, the Bayesian framework allows to integrate prior knowledge regarding the threshold values δ_k and the radius r.

The SA algorithm is run for $N = 3.5 * 10^6$ iterations. The number of iterations for the Gibbs dynamics was set to $G = L$. Each time the MH is called, it will perform $M = 200$ steps. The initial temperature is $T_1 = 2 * 10^4$ and the cooling coefficient is $c = 0.99999$. The temperature is cooled until $T_{min} = 10^{-6}$. The last 10^6 iterations will be run at constant temperature. At this very low temperature, these last outputs of the algorithms may be considered closed enough to the desired solution (16). Furthermore, they are identically distributed hence allowing the computation of level sets and of robust statistics.

The probabilities selecting the possible transitions allowed by the MH kernel were fixed as follows: $p_b = 0.2$ for "birth", $p_d = 0.2$ for "death" and $p_c = 0.6$ for "change". The support of the uniform proposal in "change" is given by $r_c = 0.3$.

The Table 1 gives a synthetic presentation of the previously mentioned variables.

Variable	Description	Value
L	number of planes	$K(K-1)/2$
δ_k	threshold of observation	$\max_j(d_{(j),k}) - \min_j(d_{(j),k})$
r	interaction radius	0.01
N	SA iterations	$3.5 * 10^6$
G	Gibbs dynamics iterations	L
M	MH iterations	200
T_1	initial temperature	10^4
c	cooling coefficient	0.99999
$p_b; p_d; p_c$	probabilities of birth;death;change	$0.2; 0.2; 0.6$
r_c	support of the uniform "change" proposal	0.3

Table 1: Data normalisation, model interaction radius and algorithms parameters.

The initial configuration of sources is given by distributing 4 points uniformly in the unit hyper-cube. The algorithm outputs, that is the pattern of sources, its statistics and the parameters θ are saved every 1000 iterations.

For the prior $p(v)$ the uniform distribution was adopted. The prior $p(\theta)$ was chosen after several trial and error and visual confirmation. The HUG model with different pre-fixed θ was applied on several synthetic data sets with known sources. The parameters providing sources close to the real sources were kept. Hence, the prior $p(\theta)$ was set as a Gaussian distribution. Its parameters were chosen according to the empirical mean and variance of the kept θ parameters. Table 2 shows these prior parameters.

	θ_1	θ_2	θ_3	θ_4
μ_{θ_i}	11.25	250.0	0.25	1.0
$\sigma^2_{\theta_i}$	1.0	10.0	0.01	0.01

Table 2: Parameter of the Gaussian prior of θ.

4.2 Synthetic data sets

In synthetic data sets, the detected sources can be compared to the real known ones. Synthetic data sets are made by first setting the number of dimensions K, the number of sources n, their compositions $\mathbf{s}^*$ and the number of samples m. The composition of each sample is created by generating a vector of sources contributions. Here, we assume that a sample or a data point is generated uniformly in the convex hull of the sources. The Dirichlet distribution with parameters $1_n = (1, \ldots, 1) \in \mathbb{R}^n$ generates uniformly in $[0, 1]^n$ vectors such as the sum of their coordinates equals 1.

The first synthetic data set contains $m = 200$ points resulting from the mixing system described in the Table 3. The data space is made by 3 hydrochemical parameters named "solute1", "solute2" and "solute3". Because the data are normalised, these hydrochemical parameters can represent for example either concentration, elemental ratio or isotopic composition, provided no curvature effect occurs. The proposed sources will be updated in $L = 3 * 2/2 = 3$ planes.

Sources	solute1	solute2	solute3
1	0.3	0.78	0.8
2	0.8	0.13	0.8
3	0.7	0.7	0.1
4	0.2	0.2	0.2

Table 3: Composition of the real sources ($\mathbf{s}^*$) for the first synthetic data set.

For each plane, the HUG model statistics for the known sources are given in Table 4.

$g(\mathbf{s}^*, \mathbf{d})$	$n_e(\mathbf{s}^*, \mathbf{d})$	$n(\mathbf{s}^*)$	$n_r(\mathbf{s}^*)$	plane
0.358501	0	4	0	1
0.294945	0	4	0	2
0.299012	0	4	0	3

Table 4: HUG model statistics of the known sources in each plane for the first synthetic data set.

The source configurations are estimated using the last 500 iterations. The Figure 2 shows the cumulative means of the statistics series. It can be observed that they clearly tend to approach the true statistics obtained from the known sources.

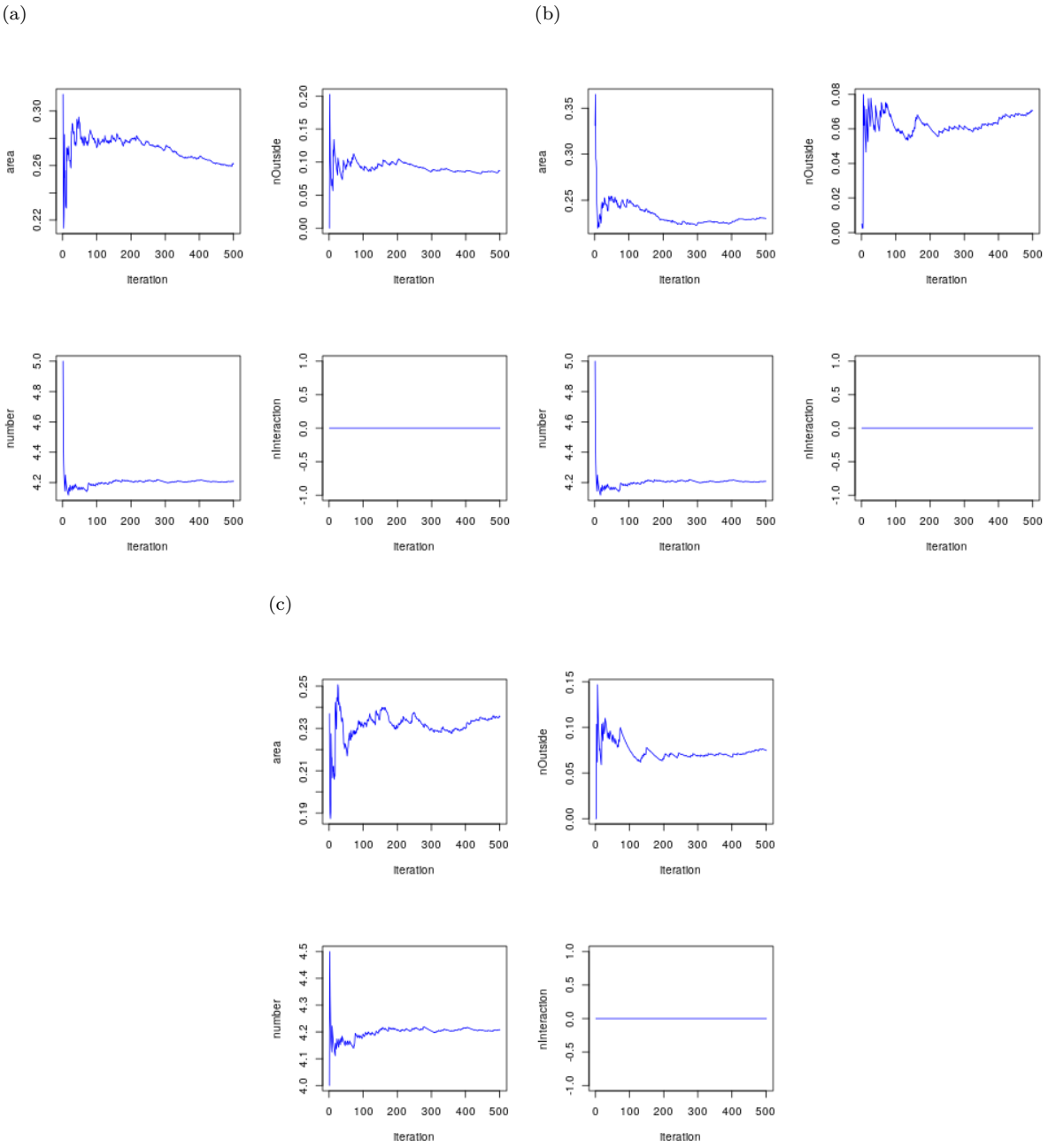


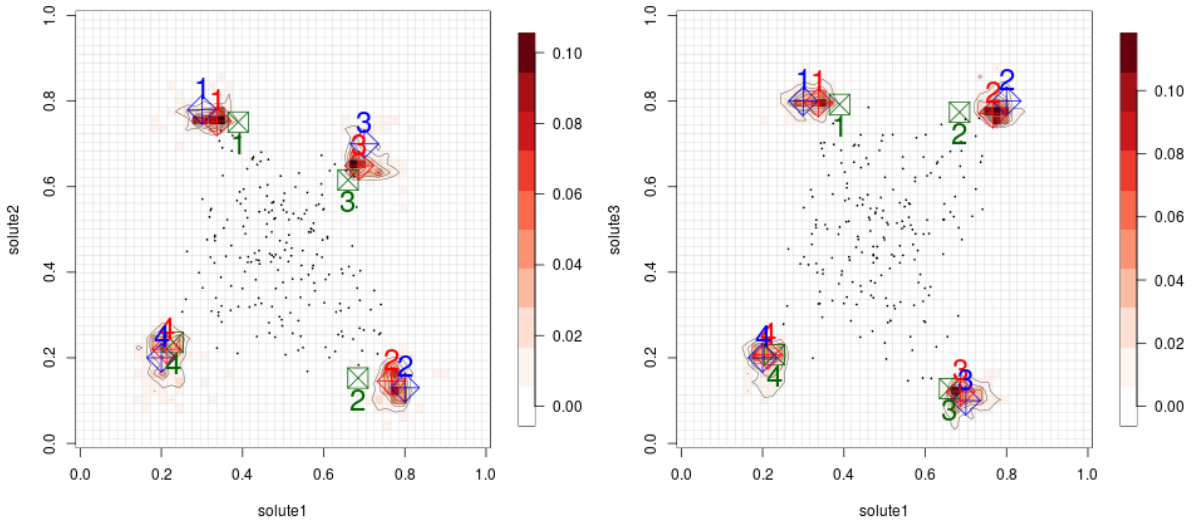
Figure 2: **Cumulative means of the statistics for the first synthetic data set:** first plane (a), second plane (b) and third plane (c). The first statistic is represented in the top left ($g(s,d)$), the second statistic is in the top right ($n_e(s,d)$), the third statistic in the bottom left ($n(s)$) and the fourth statistic in the bottom right ($n_r(s)$).

The 500 patterns of simulated sources are projected on every plane. Following [Heinrich et al., 2012], level sets are estimated in order obtain more reliable results. The computation was done for a regular grid with cells of length 0.02.

The Figure 3 presents the obtained results. The known sources are represented by blue symbols. The model finds in each plane 4 areas exhibiting high probability values of being touched by the estimated sources. The centres of each of these areas may be estimated by a clustering $k-$means algorithm. Hence, an approximation of the known sources is obtained by clustering the simulated sources on the whole

space. The number of cluster is set to 4 according to the results the Figure 2. The cluster centres are shown in green in Figure 3. An other way to estimate the sources is to consider the median point of each cluster (represented in red): each coordinate of this point is greater than 50% of the coordinates of the points in the cluster. The proposed pattern is the pattern of median points specified in the Table 5. This choice is made because the mean is impacted by the extreme value in each cluster, whereas the effect of an extremely high value is compensated by an extremely low value in the median.

(a) (b)



(c)

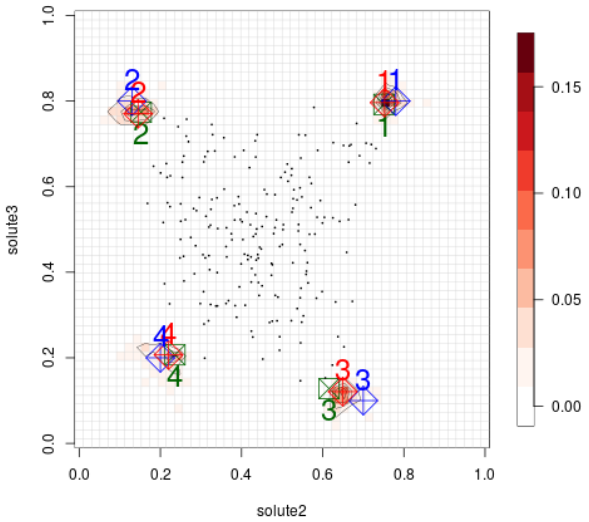


Figure 3: **Level sets computed the first synthetic data set:** first plane (a), second plane (b) and third plane (c). The probability of having a simulated sources in a cell is indicated by the coloured scale. This probability is estimated by checking the number of detected sources in a cell and dividing by the number of considered simulations. The blue symbols represent the known sources and the green symbols are the centre of the clusters obtained by the k-means algorithm. The proposed pattern, made by the median points, is represented by the red symbols.

Sources	solute1	solute2	solute3
1	0.34	0.75	0.80
2	0.77	0.15	0.77
3	0.69	0.65	0.12
4	0.21	0.22	0.21

Table 5: Proposed pattern for the first synthetic data set made by the median points of clusters computed using a k-means with 4 classes on the whole space.

To improve the results, a k-means can be run in each projection plane. The matching between the sources detected in the whole space and in the projection planes can be done manually, here. Still, recovering the source position from these projections it is not a trivial task since it is data dependent. In the next section, we propose to solve this problem by means of a sequential $k-$means algorithm.

The distance between the proposed sources i and the real sources j is calculated on the dimension k by the formula $\frac{s_{(i);k}-s^*_{(j);k}}{s^*_{(i);k}} \times 100$ which is the relative error. For each dimension the average error is given in percentage by "Mean Error Dimension" and the average error for each source is given in percentage by "Mean Error Source" in the Table 6. The proposed sources are close to the real sources and seem not systematically biased to have higher or lower value than the real sources.

Sources	solute1	solute2	solute3	Mean Error Source
1	13.3	-3.8	0.00	3.2
2	-3.8	15.4	-3.8	2.6
3	-1.4	-7.1	20.0	3.8
4	5.00	10.0	5.0	6.7
Mean Error Dimension	3.3	3.6	5.3	4.1

Table 6: Relative error in percentage for the proposed sources with respect to the known ones, and the mean error for each sources and each dimension, for the first synthetic data set.

The quality of the results relies on several elements: the model construction, the model and the algorithms parameters and the data itself. Clearly, whenever the source detection is tackled all these elements should be taken into account together in order to formulate reliable answers to the formulated questions. The data structure and its multi-dimensional character require maybe the most of the efforts whenever attempting formulating solutions to the problem of source detection.

4.2.1 Estimation of the source composition

Due to projection effects that depend on the data structure, detecting the number of sources is not a trivial task. The projected real sources are not always located in the convex hull of the real sources in every plane. The second synthetic data set is such an example. The data set contains 100 points obtained by a mixing system with 4 sources in a 3 dimensional space (see Table 7). The special feature of this data set is that only 3 are visible on each plane.

Sources	solute1	solute2	solute3
1	0.29	0.32	0.33
2	0.67	0.32	0.33
3	0.67	0.67	0.33
4	0.67	0.67	0.76

Table 7: Sources positions for the second synthetic data set.

A method that estimate the sources composition from the detected sources is developed for this situation.

First, the HUG model is applied. The model and dynamics set-up was the same as for the previous data set. Similarly as before, the cumulative means for the sufficient statistics are computed from the last 500 saved outputs with the same initialisation as for the previous data set. The results are shown in Figure 4. It can be noticed that the average number of sources is greater than 3 in each projected plane.

The considered results are given by the last 500 saved source configurations.

(a) (b)

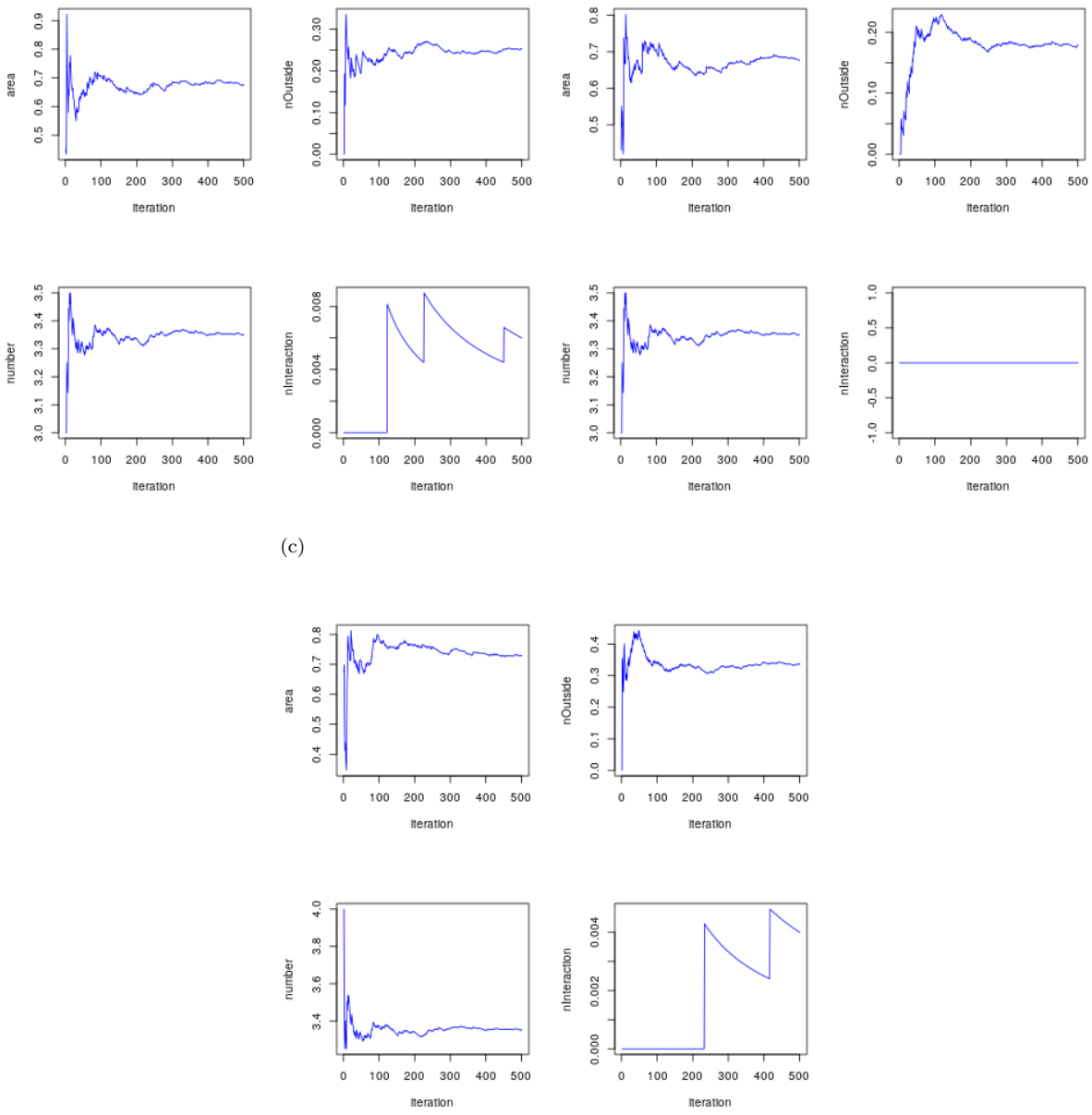
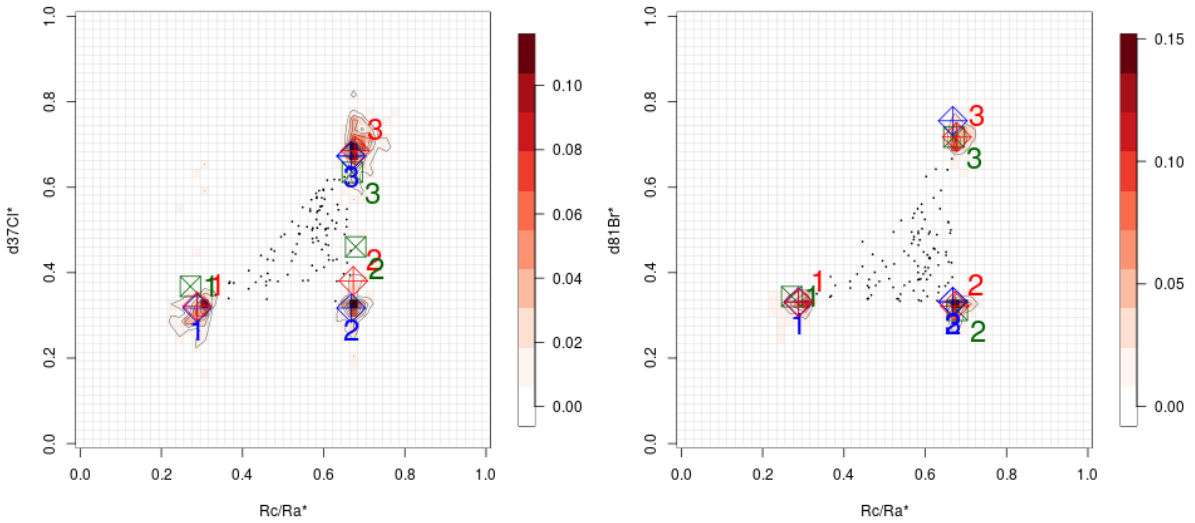


Figure 4: **Cumulative means of the statistics for the second synthetic data set:** first plane (a), second plane (b) and third plane (c). The first statistic is represented in the top left ($g(s,d)$), the second statistic is in the top right ($n_e(s,d)$), the third statistic in the bottom left ($n(s)$) and the fourth statistic in the bottom right ($n_r(s)$).

Figure 5 shows the source projections in each plane. The estimated level sets indicate three major

regions in each projection plane. The evolution of the third statistics (the number of sources), and more specifically the mean value of sources, does not always allow to conclude on the exact number of sources and by extension their position.

(a) (b)



(c)

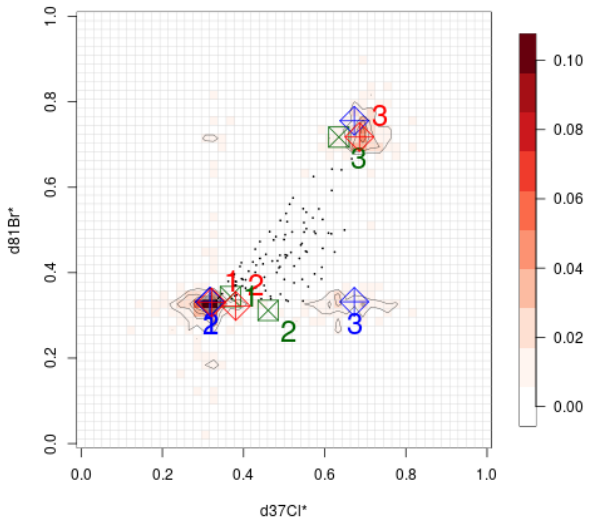


Figure 5: **Level sets computed for the second synthetic data set:** first plane (a), second plane (b) and third plane (c). The probability of having a simulated sources in a cell is given by the coloured scale. This probability is estimated by checking the number of simulated sources and dividing by the number of considered simulations. The blue symbols represent the real sources and the green symbols are the centres of the clusters obtained by a k-means algorithm with 3 sources. The red symbols are the median value of the clusters.

The number of sources cannot be known by only considering these statistics. In order to remediate this

difficulty a sequential $k-$means is proposed.

First a projection plane is chosen, the number of classes is chosen depending on the observed value of $n(\mathbf{s})$. Next a $k-$means is performed in the projection plane. Finally, the coordinates of the projected sources are replaced with the coordinates of the detected cluster centres. The algorithm is iterated till convergence, by choosing another remaining projection plane.

Algorithm Sequential k-means : Let $V = 1, \ldots, L$ be the set of all the plane and for each plane v, fix the number of clusters k_v, the simulated sources $S = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$. And note c_k^v the position of the centres obtain by a k-means with k clusters on the plane v.

1. Till $V = \emptyset$
   - Choose v uniformly without replacement in the set of projection planes
   - Compute c_k^v
   - Replace the coordinate of each simulated sources by the coordinate of the centre of its cluster
   - Remove v in V
2. Return S.

By applying the sequential k-means algorithm with 3 clusters in each plane (Figure 4), we obtain a pattern with 4 sources, rather close to the real sources, described in Table 8.

Sources	solute1	solute2	solute3
1	0.27	0.35	0.33
2	0.67	0.35	0.33
3	0.67	0.64	0.33
4	0.67	0.64	0.72

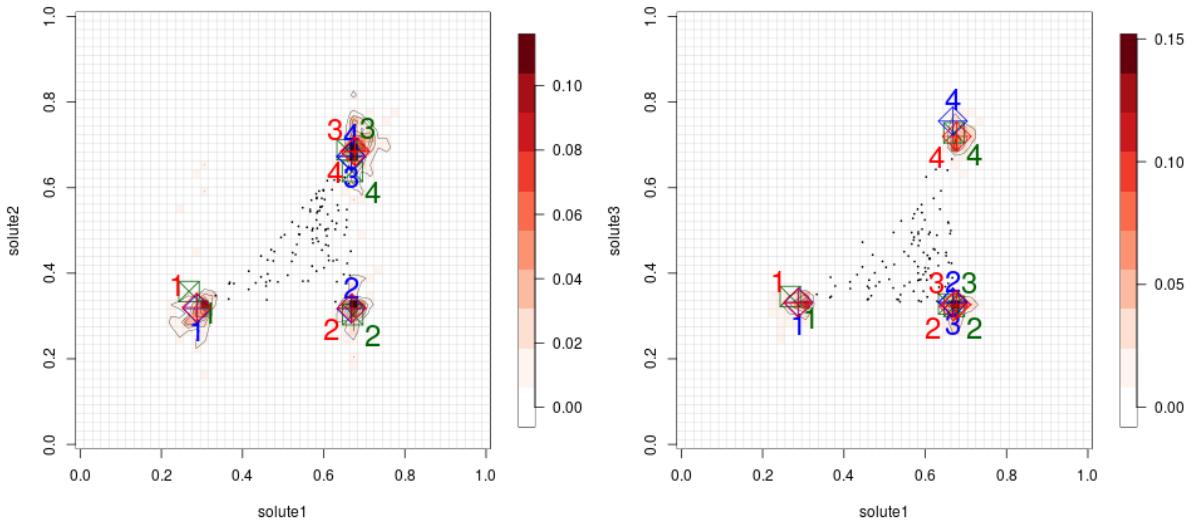
Table 8: Position of the sources obtained from the sequential k-means clustering.

A verification is done by performing a k-means with 4 clusters on the saved pattern in the whole data space. As previously, the proposed pattern is made by the median point of each cluster and is described in the Table 9 and represented in Figure 6. This pattern appears to be satisfactory: the proposed sources, as numerous as the real sources, are close to the real sources.

Sources	solute1	solute2	solute3
1	0.29	0.32	0.33
2	0.67	0.32	0.32
3	0.68	0.68	0.33
4	0.68	0.68	0.72

Table 9: Proposed pattern for the second synthetic data set made by the median points of clusters computed using a k-means with 4 classes on the whole space.

(a) (b)



(c)

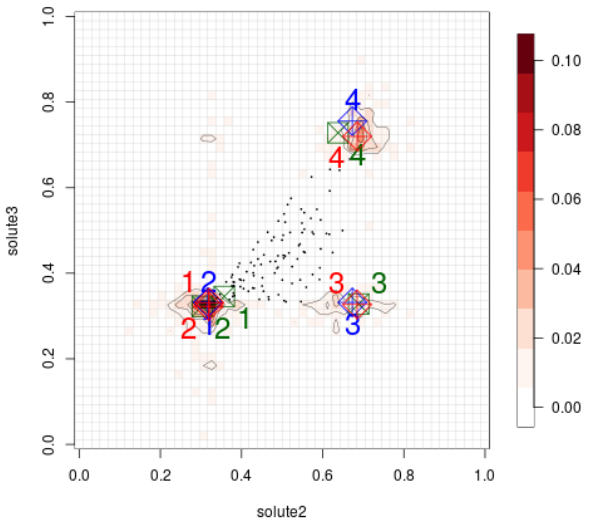


Figure 6: **Level sets computed the second synthetic data set:** first plane (a), second plane (b) and third plane (c). The probability of having a simulated sources in a cell is indicated by the coloured scale. This probability is estimated by checking the number of detected sources in a cell and dividing by the number of considered simulations. The blue symbols represent the known sources and the green symbols are the centre of the clusters obtained by the k-means algorithm. The proposed pattern, made by the median points, is represented by the red symbols.

Clearly, such a procedure is not needed if the exact number of sources is known. But this is precisely the problem to be solved. The previous application validates *a posteriori* the sequential k-means algorithm.

The error between the proposed sources and the real sources are given in the Table 10. The average percentage are all under 5%.

Sources	solute1	solute2	solute3	Mean Error Source
1	0.0	0.0	0.0	0.0
2	0.0	0.0	-3.0	-1.0
3	1.5	1.5	-0.0	1.0
4	1.5	1.5	-5.3	-0.8
Mean Error Dimension	0.8	0.8	-2.1	-0.2

Table 10: Relative error in percentage for the proposed pattern of sources with respect to the known ones, and the mean error for each sources and each dimension, for the second synthetic data set.

An alternative way of getting useful insight regarding the number of sources is the hierarchical cluster algorithm. This algorithm minimises the within-cluster variance. The result is the dendrogram of Figure 7. The algorithm starts by considering that each element is a cluster. For each step, the two clusters, that create the cluster with the smallest variance, are merged. The algorithm ends when only one cluster remains. In the figure, the level of a horizontal segment indicate the within-cluster variance of a cluster. The vertical segments are the clusters. For a given within-cluster variance, the number of cluster is given by counting the number of vertical segments.

As seen in the figure, the within-cluster variance is high if less than 4 clusters are considered. However the variance does not decrease significantly when more than 9 clusters are considered. Hence the number of sources is assumed to be between 4 and 9. The green boxes contain the 4 clusters.

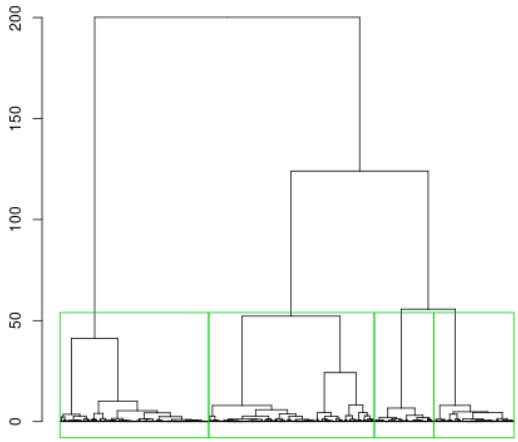


Figure 7: **Dendrogram obtain from a hierarchical clustering that minimised the within-cluster variance for the second synthetic data set**. Each green rectangle contains the simulated sources that belong to one of the four clusters.

The saved pattern of sources are now clustered in 5, 6, 7, 8 and 9 clusters. The proportion of sources in the four biggest clusters are given in the Table 11. More than 75% of simulated sources are inside 4 clusters: we can conclude that the data are made by a mixing system with 4 sources.

	5	6	7	8	9
proportion of sources in the four biggest clusters	0.92	0.84	0.82	0.75	0.77

Table 11: Proportion of simulated sources in the 4 clusters containing the most simulated sources when the k-means is applied with 5, 6, 7, 8 and 9 clusters.

4.3 Real data sets

The previously described methods are applied to two real data sets.

4.3.1 First data-set :

The first real data set on which the HUG model is applied is from [Pinti et al., 2020]. In this data set the stable isotopic composition of chlorine $\delta^{37}Cl$ ($d37Cl$ in the figures), the stable isotopic composition of bromine $\delta^{81}Br$ ($d81Br$) and $^3He/^4He$ (Rc/Ra) are measured on $m = 75$ samples from geothermal wells from Mexico and supposed results of a three-source mixing system. Note that halogens and noble gases behave conservatively during fluid mixing and that the mixing trends in the two planes considered here are not affected by curvature [Pinti et al., 2020] so that the conditions of the use of the HUG model apply here and $L = 2$.

Because the data are supposed results of a three-source mixing system, the source can be considered as the vertex of a triangle containing the data on each plane. In [Pinti et al., 2020], the sources are estimated by the vertex of the smallest triangle (in term of area) containing the data on each plane (Table 12).

sources	"Rc/Ra"	"delta37Cl" in ‰	"Rc/Ra"	"delta81Br" in ‰
1 (mantle)	7.76	0.88	8.26	0.75
2 (subduction)	6.45	-0.43	7.17	-1.03
3 (crust)	1.68	0.11	1.89	0.26

Table 12: Sources estimated in [Pinti et al., 2020] when the data are supposed bi-dimensional and resulting from a three-source mixing system. The sources are the vertex of the smallest triangle that contains the data.

To reconstruct the sources in the 3 dimensional space, these sources are merged by the coordinate that they have in common. The position in "Rc/Ra" is the middle between the two sources merged. The reconstructed sources are described in Table 13.

sources	"Rc/Ra"	"delta37Cl" in ‰	"delta81Br" in ‰
1 (mantle)	8.01	0.88	0.75
2 (subduction)	6.81	-0.43	-1.03
3 (crust)	1.78	0.11	0.26

Table 13: Position of the reconstructed sources.

The HUG model is applied with the same initialisation as in the previous section. As previously the last 500 saved pattern are projected on every normalised plane in regular grid with cells of length 0.02. The normalised dimensions are indicated by adding a $*$ to the raw dimensions. The Figure 8 presents the results. The blue symbols are the previously bi-dimensional mentioned sources and the reconstructed sources. The model finds on each plane 3 areas with high probability of containing an estimated sources. The centres of theses areas, in green in the figure, are also estimated by k-means algorithm with 3

clusters. The proposed pattern made by the median points of the clusters is represented by the red symbols in the figure.

(a) (b)

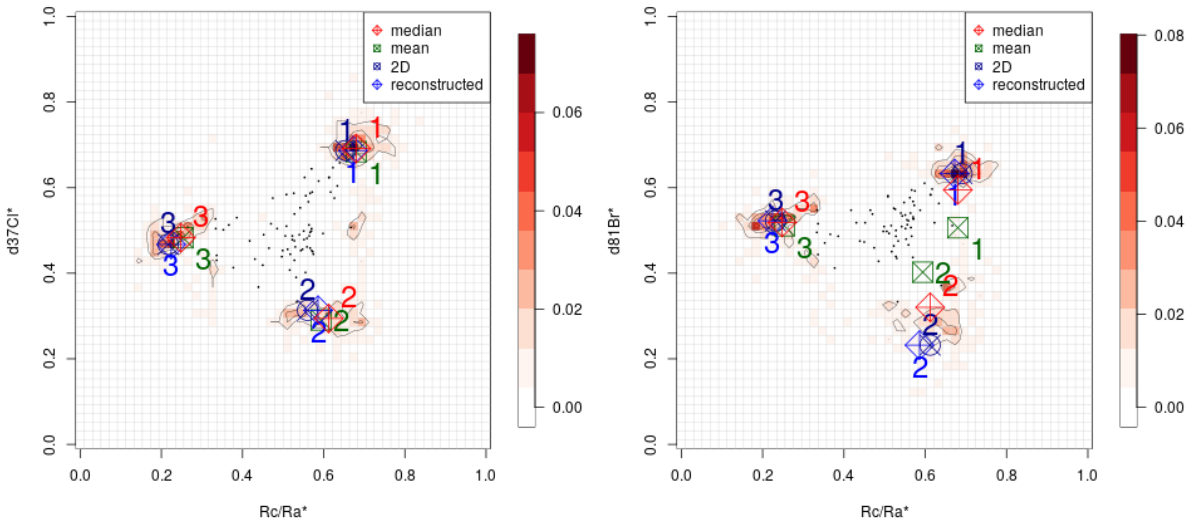


Figure 8: **Level sets computed the first real data set:** first normalised plane (a) and second normalised plane (b). The probability of having a simulated sources in a cell is indicated by the coloured scale. This probability is estimated by checking the number of detected sources in a cell and dividing by the number of considered simulations. The blue symbols represent the vertex of the smallest triangle containing the data and the reconstructed sources. The green symbols are the centre of the clusters obtained by the k-means algorithm. The proposed pattern, made by the median points, is represented by the red symbols.

The proposed pattern of sources is not useful in this state because it can not be compare to the raw data and the sources of Table 12. Hence the reverse procedure of the normalisation, presented in section 4.1, has to be applied on the detected sources and the proposed pattern in Table 14.

(a) (b)

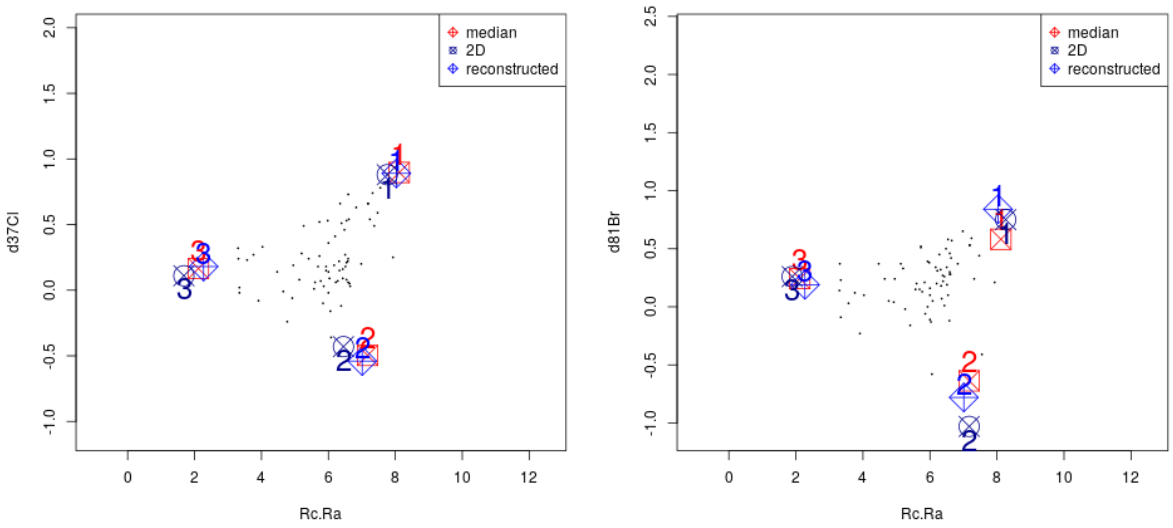


Figure 9: **Position of the proposed sources:** first plane (a) and second plane (b). The composition in d37Cl and d81Br are given in ‰. The blue symbols represent the vertex of the smallest triangle containing the data and the reconstructed sources. The proposed pattern, made by the median points, is represented by the red symbols.

sources	"Rc/Ra"	"delta37Cl" in ‰	"delta81Br" in ‰
1 (mantle)	8.12	0.90	0.58
2 (subduction)	7.17	-0.50	-0.64
3 (crust)	2.12	0.17	0.24

Table 14: Proposed pattern for the first real data set made by the median points of clusters computed using a k-means with 3 classes on the whole space and after reversing the transformation presented in section 4.1

The relative error in percentage between the reconstructed sources and the proposed sources, after reversing the transformation are given in Table 15.

Sources	"Rc/Ra"	"delta37Cl"	"delta81Br"	Mean Error Source
1	1.4	2.3	-22.7	-6.3
2	5.3	16.3	-37.9	-5.4
3	19.1	54.5	-7.7	22.0
Mean Error Dimension	8.6	24.4	-22.8	3.4

Table 15: Relative error in percentage for the detected sources with respect to the known ones, and the mean error for each sources and each dimension, for the first real data set.

Eventually, applying the HUG model provides fairly consistent results with the geometrical approach proposed by [Pinti et al., 2020] on this rather simple data set (i.e. 3 parameters considered and 3 proposed sources). The relatively small differences in composition of the sources detected by the HUG model compared to those proposed by [Pinti et al., 2020] do not imply to reconsider the geological interpretations regarding the origin of the geothermal fluids

4.3.2 Second data-set :

The last data-set considered in this paper is the Athabasca data-set made from [Richard et al., 2010], [Richard et al., 2016] and [Martz et al., 2019]. Fluid inclusions from uranium deposit of the Athabasca Basin (Canada) are studied. The concentration in chemical elements are obtained by a Laser Ablation-Inductively Coupled Plasma Mass Spetrometry (LA-ICPMS). In the articles that presented the data, only two sources were considered: NaCl-rich brine and $CaCl_2$-rich brine. These sources are defined by the projected data on the considered planes. Data with $[Na] > 80000$ are regrouped in the first cluster, and data with $[Na] < 30000$ in the second cluster. The distribution of the composition in each element of these cluster are calculated. The quantile at 25% and 75% are given in Table 16. These sources will be represented in each plane by the blue rectangles made respectively by continuous line and dotted line. The sides of these rectangles are the given quantile. This method consider bi-dimensional data. The nature of these sources are different from the sources detected by the HUG model: this is the reason why the number of proposed sources will be higher for the HUG model. These sources are inside the convex hull of the data while the sources detected by the HUG model are generally outside the convex hull of the data. Hence, the sources of the HUG model is more dependant on the quality of the data than the sources from [Richard et al., 2016]: an outlier will change the convex hull and so the proposed sources. This reliance is controlled by the parameters θ_1 and θ_2: low value will reduce the effect of outliers.

Sources	NaCl-rich brine		$CaCl_2$-rich brine	
	Q25	Q75	Q25	Q75
[Li] in ppm	900	3000	520	6000
[Na] in ppm	80000	100000	15000	22000
[Mg] in ppm	4000	9000	22000	40000
[K] in ppm	1700	5200	8000	17000
[Ca] in ppm	11000	32000	27000	60000

Table 16: **Range of the sources detected in [Richard et al., 2016].** The data are regroup in a group containing the data with $[Na] > 80000$ and a group containing data with $[Na] < 30000$. For these groups, respectively NaCl-rich brine and $CaCl_2$-rich brine, are given the quantile at 25% (Q25) and 75% (Q75).

Here, the concentration of 5 chemical elements (lithium Li, sodium Na, magnesium Mg, potassium K and calcium Ca) and so $L = 5*(5-1)/2 = 10$ planes are considered. This choice is made in order to have a relatively fast computation. Moreover, because the data set is made by 3 different studies, not every chemical element is measured for each fluid inclusion: by considering these 5 dimensions all the fluid inclusions will be considered. With the previous normalisation of the data, the sources can be negative. However concentration are positive, hence a new normalisation is made: for each dimension $k \in [1, \dots, 5]$ the window $(max\{\min_j(d_{(j),k}) - \delta_k, 0\}, \max_j(d_{(j),k}) + \delta_k)$, with $\delta_k = \max_j(d_{(j),k}) - \min_j(d_{(j),k})$, is transformed into $[0.0, 1.0]$.

As previously, the HUG model is applied on the data. The last 500 saved pattern are projected on the 10 normalised planes in regular grid with cells of length 0.02. The results are shown in the Figure 10. On each plane, the model detects 3 areas with a rather high probability, indicating the potential sources presence. The sequential k-means detects 6 sources. The two brine are represented by the blue rectangles obtained by normalisation. The centres of the 6 clusters made by a k-means are represented by the green symbols. The proposed pattern is made by the median points of these clusters and is represented by the red symbols.

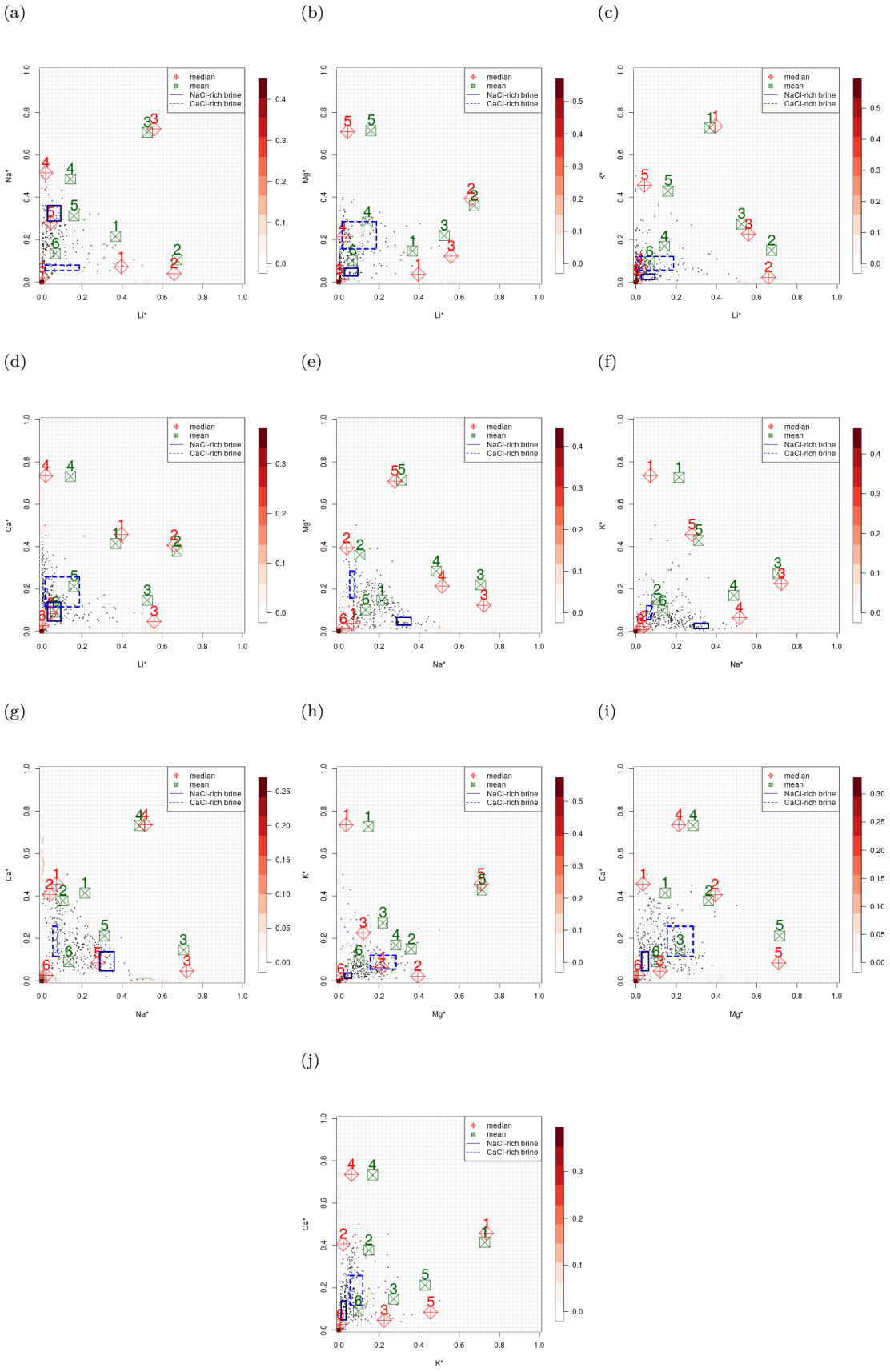


Figure 10: **Level sets computed the second real data set:** normalised plane 1 (a) to normalised plane 10 (j). The probability of having a source in a cell is indicated by the coloured scale. This probability is estimated by checking the number of detected sources in a cell and dividing by the number of considered simulations. The blue rectangles made by the continuous line and the dotted line represent respectively the NaCl-rich brine and the $CaCl_2$-rich brine presented in [Richard et al., 2016]. The green symbols are the centre of the clusters obtained by the k-means algorithm. The proposed pattern, made by the median points, is represented by the red symbols.

The hierarchical cluster algorithm is also applied to confirm the number of sources. As seen in Figure 11, the within-cluster variance is rather high when less than 6 clusters are considered. Moreover when more than 10 clusters are considered, the variance is very low. In conclusion the number of clusters should be between 6 and 9. The 6 clusters are delineated by the green boxes.

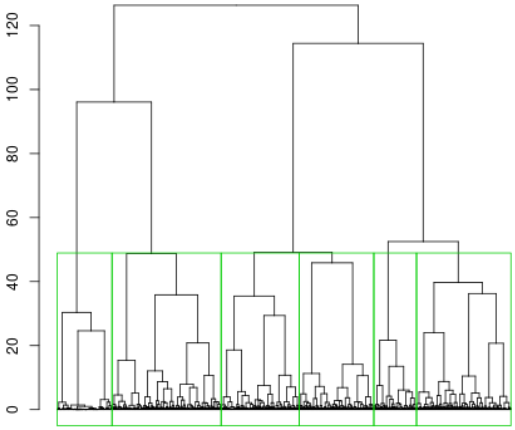


Figure 11: **Dendrogram obtain from a hierarchical clustering that minimised the within-cluster variance.** Each green rectangle contains the simulated sources that belong to one of the six clusters.

The proportion of simulated sources in the 6 biggest clusters obtained by a k-means with 7, 8 and 9 clusters, is given in Table 17. The proportion is greater than 70%, hence the hypothesis of 6 clusters can be assumed.

	7	8	9
proportion of points in the six biggest clusters	0.90	0.79	0.72

Table 17: Proportion of simulated sources in the 6 clusters containing the most simulated sources when the fuzzy k-means is applied with 7, 8 and 9 clusters.

The proposed pattern of sources is given in Table 18.

Sources	Li*	Na*	Mg*	K*	Ca*
1	0.35	0.74	0.36	0.41	0.33
2	0.71	0.43	0.50	0.32	0.31
3	0.29	0.36	0.30	0.36	0.73
4	0.37	0.34	0.34	0.69	0.31
5	0.31	0.39	0.71	0.44	0.56
6	0.73	0.39	0.41	0.44	0.73

Table 18: Proposed pattern for the second real data set made by the median points of clusters computed using a k-means with 6 classes on the whole normalised space.

The proposed pattern of sources is presented in the original data space in the Table 19 and the Figure 12.

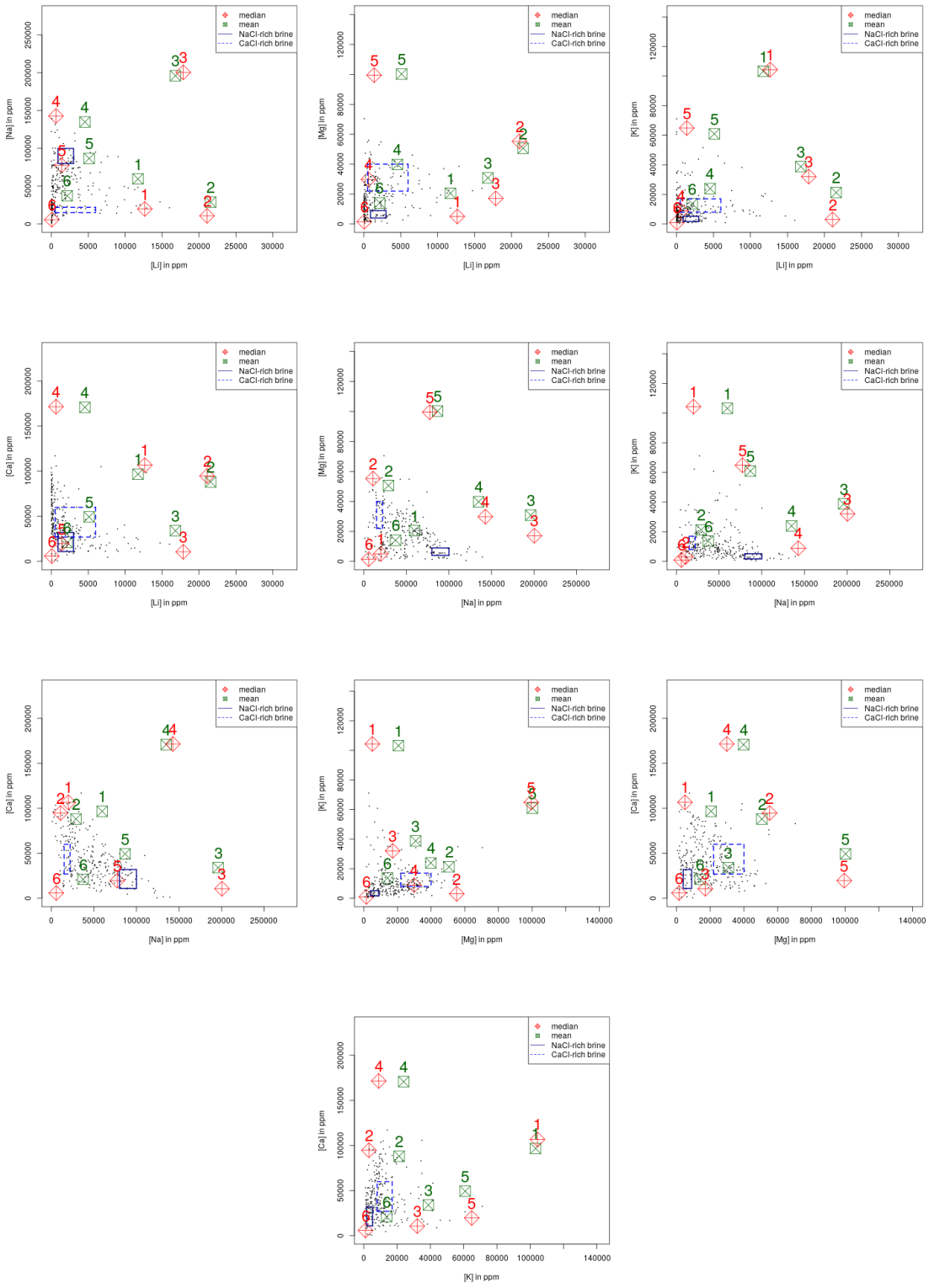


Figure 12: **Position of the centres of the clusters obtained from a k-mean clustering with 6 clusters for the second real data.** The blue rectangle and the red rectangle represent respectively the NaCl-rich brine and the $Cacl_2$-rich brine presented in [Richard et al., 2016].

On each plane there are three areas: a first area with several sources near the origin of the plane (low abscissa values, low ordinate values), a second area at the top of the data (low abscissa values, high ordinate values) and a last area at the right of the data (high abscissa values, low ordinate values).

The geological meaning of the proposed sources is beyond the scope of this article.

Sources	[Li] in ppm	[Na] in ppm	[Mg] in ppm	[K] in ppm	[Ca] in ppm
1	12658	19934	5070	104325	106728
2	21084	10813	55239	3051	94757
3	17870	200441	17131	32081	10663
4	624	142790	29854	8944	171537
5	1393	77471	99486	64833	19675
6	57	5744	1524	992	5900

Table 19: Position of the centres of the clusters obtained from from a fuzzy k-mean clustering with 6 clusters after the reverse transformation.

5 Conclusions and perspectives

This paper presents a new interaction point process that integrates geological knowledge for the purpose of automatic sources detection. The construction of the model takes into account the high dimensional character of the data. A Metropolis Hastings within Gibbs simulation dynamics was built for the model in order to manage the multidimensional aspect of the problem. The source pattern is estimate by the point process configuration that maximise the probability density describing the model. Based on the proposed Metropolis-Hastings dynamics, a simulated annealing algorithm was made, in order to avoid local minima. Level sets estimation is used in order to provide more reliable results and to reduce uncertainties. The adopted strategy to cope with the multidimensional of the problem, was to perform inference on projection planes. The synthesis of the obtained results was done by constructing a new sequential $k-$means algorithm.

The model parameters set-up was done by using synthetic data where the sources were known. This allowed the construction of parametric priors $p(\theta)$. Detection errors are provided for the considered synthetic data-sets. Numerical experiences done using known real data sets already show that the results obtained with our automatic method match the ones presented in the literature.

Clearly, the prior choice is a crucial point that influences the general performances of the method. Currently, new procedures for inferring the model parameters are studied in order to improve the quality of the results furnished by the proposed algorithms.

The HUG model is a flexible tool that detects sources patterns in multidimensional data, without knowledge on the number of sources. At our best knowledge, automatic result validation is still an open problem. For the moment, the results should always be confirmed and confronted by human expert knowledge.

If it is to list challenges regarding the present approach, we would like to mention the consideration of chemical reactions and curvature effects. Moreover in this paper the proposed sources are supposed to contribute to every data point. Hence it may be interesting to introduce consideration of time by searching for sub-systems of mixing or adding the geographic position of the data points. Last but not least, the data are considered by the presented model without considering uncertainty: a consideration of uncertainty may be interesting.

Acknowledgements

This work was performed in the frame of the DEEPSURF project (http://lue.univ-lorraine.fr/fr/impact-deepsurf) at Université de Lorraine. This work was supported partly by the french PIA project Lorraine Université d'Excellence, reference ANR-15-IDEX-04-LUE.

References

[Andrew, 1979] Andrew, A. M. (1979). Another efficient algorithm for convex hulls in two dimensions. *Information Processing Letters*, 9(5):216–219.

[Arendt et al., 2015] Arendt, C., Aciego, S., and Hetland, E. (2015). An open source Bayesian Monte Carlo isotope mixing model with applications in Earth surface processes. *Geochemistry, Geophysics, Geosystems*, 16(5):1274–1292.

[Baddeley et al., 2016] Baddeley, A. J., Rubak, E., and Turner, R. (2016). *Spatial Point Patterns: Methodology and Applications with R.* Chapman and Hall/CRC Press, London.

[Carrera et al., 2004] Carrera, J., Vázquez-Suñé, E., Castillo, O., and Sánchez-Vila, X. (2004). A methodology to compute mixing ratios with uncertain end-members. *Water resources research*, 40(12).

[Christophersen and Hooper, 1992] Christophersen, N. and Hooper, R. P. (1992). Multivariate analysis of stream water chemical data: The use of principal components analysis for the end-member mixing problem. *Water Resources Research*, 28(1):99–107.

[Delsman et al., 2013] Delsman, J. R., Essink, G. H. O., Beven, K. J., and Stuyfzand, P. J. (2013). Uncertainty estimation of end-member mixing using generalized likelihood uncertainty estimation (GLUE), applied in a lowland catchment. *Water Resources Research*, 49(8):4792–4806.

[Faure, 1997] Faure, G. (1997). *Principles and applications of geochemistry*, volume 625. Prentice Hall New Jersey, United States,.

[Geyer, 1999] Geyer, C. J. (1999). Likelihood inference for spatial point processes. In Barndorff-Nielsen, O., Kendall, W., and van Lieshout, M., editors, *Stochastic Geometry, Likelihood and Computation.* CRC Press/Chapman and Hall, Boca Raton.

[Heinrich et al., 2012] Heinrich, P., Stoica, R. S., and Tran, V. C. (2012). Level sets estimation and Vorob'ev expectation of random compact sets. *Spatial Statistics*, 2:47–61.

[Ingebritsen et al., 2006] Ingebritsen, S. E., Sanford, W. E., and Neuzil, C. E. (2006). *Groundwater in geologic processes.* Cambridge University Press.

[Lajaunie et al., 2020] Lajaunie, C., Renard, D., Quentin, A., Le Guen, V., and Caffari, Y. (2020). A non-homogeneous model for kriging dosimetric data. *Mathematical Geosciences*, 52(7):847–863.

[Langmuir et al., 1978] Langmuir, C. H., Vocke Jr, R. D., Hanson, G. N., and Hart, S. R. (1978). A general mixing equation with applications to icelandic basalts. *Earth and Planetary Science Letters*, 37(3):380–392.

[Longman et al., 2018] Longman, J., Veres, D., Ersek, V., Phillips, D. L., Chauvel, C., and Tamas, C. G. (2018). Quantitative assessment of Pb sources in isotopic mixtures using a Bayesian mixing model. *Scientific reports*, 8(1):6154.

[Martz et al., 2019] Martz, P., Mercadier, J., Cathelineau, M., Boiron, M.-C., Quirt, D., Doney, A., Gerbeaud, O., De Wally, E., and Ledru, P. (2019). Formation of U-rich mineralizing fluids through basinal brine migration within basement-hosted shear zones: A large-scale study of the fluid chemistry around the unconformity-related Cigar Lake U deposit (Saskatchewan, Canada). *Chemical Geology*, 508:116–143.

[Møller and Waagepetersen, 2003] Møller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes.* Chapman and Hall/CRC.

[Parnell et al., 2010] Parnell, A. C., Inger, R., Bearhop, S., and Jackson, A. L. (2010). Source partitioning using stable isotopes: coping with too much variation. *PloS one*, 5(3):e9672.

[Phillips and Gregg, 2001] Phillips, D. L. and Gregg, J. W. (2001). Uncertainty in source partitioning using stable isotopes. *Oecologia*, 127(2):171–179.

[Pinti et al., 2020] Pinti, D. L., Shouakar-Stash, O., Castro, M. C., Lopez-Hernández, A., Hall, C. M., Rocher, O., Shibata, T., and Ramírez-Montes, M. (2020). The bromine and chlorine isotopic composition of the mantle as revealed by deep geothermal fluids. *Geochimica et Cosmochimica Acta*.

[Richard et al., 2016] Richard, A., Cathelineau, M., Boiron, M.-C., Mercadier, J., Banks, D. A., and Cuney, M. (2016). Metal-rich fluid inclusions provide new insights into unconformity-related U deposits (Athabasca basin and basement, Canada). *Mineralium Deposita*, 51(2):249–270.

[Richard et al., 2010] Richard, A., Pettke, T., Cathelineau, M., Boiron, M.-C., Mercadier, J., Cuney, M., and Derome, D. (2010). Brine–rock interaction in the Athabasca basement (McArthur river U deposit, Canada): consequences for fluid chemistry and uranium uptake. *Terra Nova*, 22(4):303–308.

[Robb, 2005] Robb, R. (2005). Introduction to ore-forming processes, book.

[Ruelle, 1999] Ruelle, D. (1999). *Statistical Mechanics : Rigorous Results*. Imperial College Press, World Scientific Publishing,.

[Skuce et al., 2015] Skuce, M., Longstaffe, F., Carter, T., and Potter, J. (2015). Isotopic fingerprinting of groundwaters in southwestern Ontario: Applications to abandoned well remediation. *Applied Geochemistry*, 58:1–13.

[Stoica et al., 2004] Stoica, R., Descombes, X., and Zerubia, J. (2004). A Gibbs point process for road extraction from remotely sensed images. *International Journal of Computer Vision*, 57(2):121–136.

[Stoica et al., 2007a] Stoica, R., Gay, E., and Kretzschmar, A. (2007a). Cluster pattern detection in spatial data based on Monte Carlo inference. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 49(4):505–519.

[Stoica et al., 2005a] Stoica, R., Gregori, P., and J. Mateu, J. (2005a). Simulated annealing and object point processes : tools for analysis of spatial patterns. *Stochastic Processes and their Applications*, 115:1860–1882.

[Stoica et al., 2005b] Stoica, R., Martinez, V. J., Mateu, J., and Saar, E. (2005b). Detection of cosmic filaments using the Candy model. *Astronomy & Astrophysics*, 434(2):423–432.

[Stoica et al., 2007b] Stoica, R., Martínez, V. J., and Saar, E. (2007b). A three-dimensional object point process for detection of cosmic filaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(4):459–477.

[Tipton et al., 2022] Tipton, J. R., Sharman, G. R., and Johnstone, S. A. (2022). A bayesian nonparametric approach to unmixing detrital geochronologic data. *Mathematical Geosciences*, 54(1):151–176.

[van Lieshout, 2000] van Lieshout, M. N. M. (2000). *Markov Point Processes and their Applications*. Imperial College Press, London.

[Weltje, 1997] Weltje, G. J. (1997). End-member modeling of compositional data: Numerical-statistical algorithms for solving the explicit mixing problem. *Mathematical Geology*, 29(4):503–549.

[Yardley and Bodnar, 2014] Yardley, B. W. and Bodnar, R. J. (2014). Fluids in the continental crust. *Geochemical Perspectives*, 3(1):1–2.