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

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

January 28, 2023

Abstract

Detecting the number and composition of multiple sources in groundwaters from hydrochemical data has remained highly challenging. 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 multidimensional space of hydrochemical parameters. The key assumption of this approach is to consider 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 multidimensional nature of the data and specific physical rules. These rules induce a source configuration able to explain the observations. This distribution is achieved 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 geothermal and ore-forming 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. The main sources of surface and sub-surface waters which contribute to the composition of groundwaters in hydrothermal systems through mixing processes are shown in the Figure 1 (a)). Similar water mixing processes also occur in surface and shallow subsurface environments, potentially involving other water sources. 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").

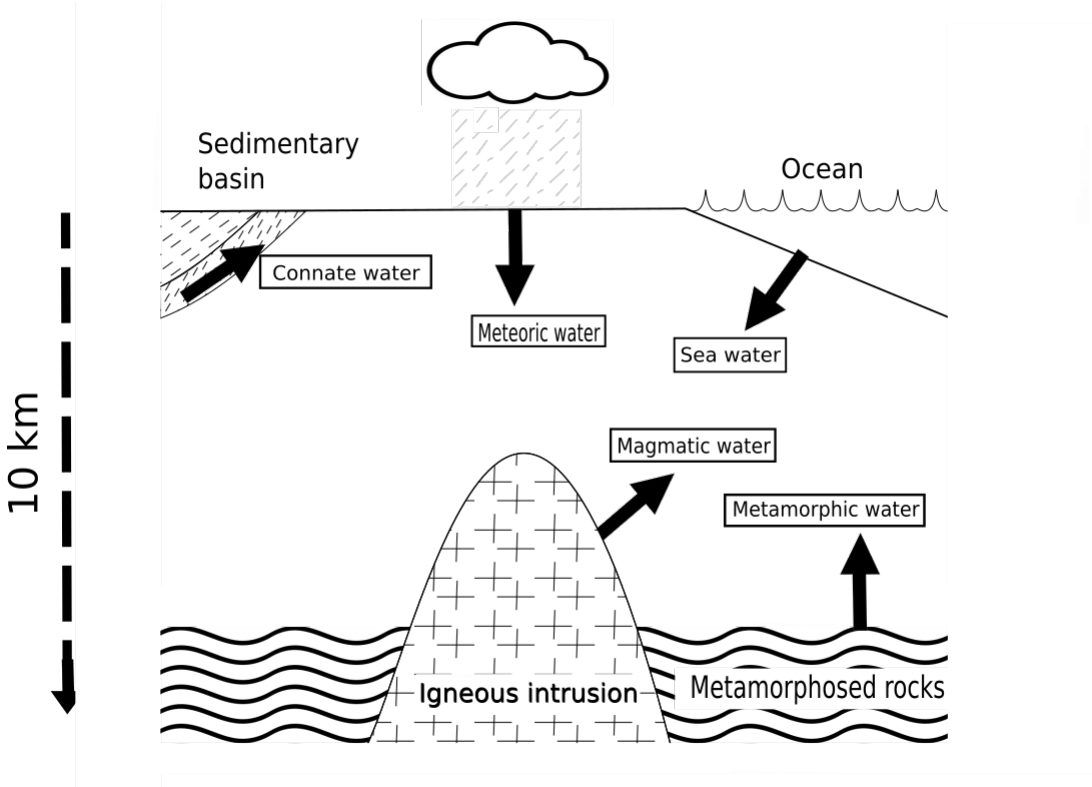


Figure 1: Conceptual cross-section of the Earth's continental crust showing the main sources of surface and deep waters (modified from [Robb, 2005]).

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 boreholes 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, mentions 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, whereas mentions of dimension refer to composition in a hydrochemical parameter. In the context of fluid mixing, a sample is considered as a barycentre of the sources: a data point d_j is a result of a mixing between the pattern of sources $\mathbf{s} = \{s_1, \ldots, s_n\}$ 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. The Figure 1 (b) is an example of a ternary diagram showing typical mixing scenario in the space of hydrochemical parameters (solute1, solute2, solute3). Blue symbols represent the sources/end-members. Black dots represent the samples/mixing terms.

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].

The already existing solutions, proposed to solve the problem of sources detection, try to tackle three challenges: the multidimensional nature 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 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 graphically 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 terms of area) that contains the data in the case of three sources in a bidimensional 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 the 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 distribution of the sources in the data space. The set of sources is estimated by the configuration of points 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 multidimensional aspect of the problem.

The conditions for using the Hug model are as follows: the datasets 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 structure of the paper is as 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 multidimensional nature 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 tuning 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 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 for all $F \in \mathcal{F}$:

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

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 cardinality 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 analytically 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: adding a point to the current configuration

(birth), deleting a point from the current configuration (death) and changing a point from the current configuration into a new position (change). Let $p_b, p_d, p_c \in [0,1]$, with $p_b+p_d+p_c \leq 1$, be the probability of birth, death and change, respectively. With probability p_b the move birth is selected: a new point η is generated according to a distribution $b(\mathbf{s}, \eta)$. With probability p_d the move death is selected: a point η chosen in the configuration according to a distribution $d(\mathbf{s}, \eta)$ is deleted. With probability p_c the move change is selected: a point η chosen in the configuration according to a distribution $q(\mathbf{s}, \eta)$ is changed into a location ζ generated according to a distribution $c(\mathbf{s}, \eta, \zeta)$. As indicated in the frame below, this transition kernel is embedded 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 to $b(\mathbf{s}, \eta)$. 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 to $d(\mathbf{s}, \eta)$. 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

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

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]. For our situation, this implies that the simulated Markov chain with this transition kernel converges towards the distribution of the proposed model from any initial condition, with a geometric speed.

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 subspace 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 X be a point process and $S = \bigcup_{i=1}^{m} \tilde{s_i}$ a decomposition of the domain S in a finite number of cells m, such that $\nu(\tilde{s_i}) = \text{ct.}$, for all i. Let $\tilde{S}$ be the set of all the cells. The contact probability between the point process and a grid cell is

$$p(s) = \mathbb{P}(s \cap X \neq \emptyset).$$

For $\lambda \in [0, 1]$, let us consider the level set given by

$$l_\lambda = \{s \in \tilde{S} \ : p(s) > \lambda\}.$$

The practical computation of $p(s)$ is done via Monte-Carlo methods, the estimator

$$p_n(s) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{s \cap S_i \neq \emptyset\}}.$$

with $S_1, S_2, \ldots, S_n$ i.i.d realisations of $p(\mathbf{s}|\theta)$. Hence, the estimator of the level set is

$$l_{n,\lambda} = \{w \in \tilde{S} \ : p_n(s) > \lambda\}. \tag{7}$$

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 $\mathbf{d} = \{d_1, d_2, \ldots, d_m\}$. For instance such a hydrochemical parameter may be the concentration of a given chemical element. The sample or data point numbered j $(1 \leq j \leq m)$ is placed in the data space with the coordinates $d_j = (d_{(j);1}, \ldots, d_{(j);K})$. Here $d_{(j);k} \in \mathbb{R}$ represents the measurement of the hydrochemical parameter numbered k $(1 \leq k \leq K)$ associate to sample j. Hence, the dataset 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 the data points. The different aspects of the relationship between the data and the unknown sources can be synthesised by the following assumptions:

(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 source: this is due to the fact that the data points are barycentres 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, that is their location in the data space should be significantly different, from one source to another.

By hypothesis (a) the position of the sources are in a bounded space: without it, every dataset 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) take into account the fact that 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 in 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

$$p_{\mathbf{d}}(\mathbf{s}|\theta) = \frac{\exp\left[-U(\mathbf{s}|\theta)\right]}{Z(\theta)} = \frac{\exp\left[-U_{\mathbf{d}}(\mathbf{s}|\theta) - U_i(\mathbf{s}|\theta)\right]}{Z(\theta)}. \tag{8}$$

Here $U_{\mathbf{d}}(\mathbf{s}|\theta)$ is the data term. It locates the sources in the data space while taking into account the hypotheses (a) and (b). It depends on the observed data $\mathbf{d}$. The term $U_i(\mathbf{s}|\theta)$ is the interaction energy that manages the sources relative position by taking into consideration the hypotheses (c) and (d). This term does not depend on $\mathbf{d}$. Finally, $Z(\theta)$ represents the normalising constant. The sum of the data and interaction terms gives the total energy function U.

In the following, we specify the model (8). First, the model is presented when $K = 2$, that is the hydrochemical space is a finite surface. Then the model is generalised for $K \geq 2$.

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 $g(\mathbf{s}, \mathbf{d})$:

$$g(\mathbf{s}, \mathbf{d}) = \left| \frac{g(\mathbf{s})}{g(\mathbf{d})} - 1 \right| \tag{9}$$

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

If the data and the convex hull determined by the sources 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 (9) can be performed via the Andrew's monotone chain convex hull algorithm [Andrew, 1979].

A simple solution to prevent pathological cases is to require a number of minimum three sources. As it will be shown later, the impact of this supplementary condition is attenuated by the use of level sets and the sequential $k-$means algorithm presented in Section 4.2.1.

The hypothesis (b) is considered by the following statistic

$$n_e(\mathbf{s}, \mathbf{d}) = 1 - \frac{n_{expl}(\mathbf{s}, \mathbf{d})}{m} \tag{10}$$

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 samples. Whenever the sources tend to explain all the points, the statistic (10) is close to 0.

The data energy function is:

$$U_{\mathbf{d}}(\mathbf{s}|\theta) = \theta_1 g(\mathbf{s}, \mathbf{d}) + \theta_2 n_e(\mathbf{s}, \mathbf{d}), \tag{11}$$

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 taking 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 statistic represents the number of pairs of sources situated within a pre-fixed distance r from each other:

$$n_r(\mathbf{s}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbf{1}\{d(s_i, s_j) \leq r\}, \tag{12}$$

here $d(s_i, s_j)$ is the Euclidean distance between s_i and s_j.

The interaction energy function is:

$$U_i(\mathbf{s}|\theta) = \theta_3 n(\mathbf{s}) + \theta_4 n_r(\mathbf{s}), \tag{13}$$

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 a Gibbs point process defined by the energy functions (11) and (13).

Assuming knowledge related to the model 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):

$$\widehat{(\mathbf{s}, \theta)} = \arg\max_{\Omega \times \Theta} p_{\mathbf{d}}(\mathbf{s}, \theta) = \arg\max_{\Omega \times \Theta} p_{\mathbf{d}}(\mathbf{s}|\theta)p(\theta) \tag{14}$$

with the configuration space Ω and the parameter space Θ a compact region in $\mathbb{R}^4$. The computation of (14) can be achieved via a simulated annealing algorithm.

3.4 General case: $K \geq 2$

The observed datasets contain a number of hydrochemical parameters greater than two. The construction of a solution in dimension K by the generalisation of (14) is mathematically possible. Still, this straightforward approach may lead to extremely heavy computations.

Here, an alternative solution is preferred. Assuming the dataset 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 achieve 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

$$p_{\mathbf{d}}(\mathbf{s}|\theta, v) \propto \exp[-U(\mathbf{s}|\theta, v)], \tag{15}$$

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

$$U_{\mathbf{d}}^{(l)}(x|\theta) = \theta_1 g^{(l)}(\mathbf{s}, \mathbf{d}) + \theta_2 n_e^{(l)}(\mathbf{s}, \mathbf{d}), \quad l = 1, \ldots, L, \tag{16}$$

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

$$U_i^{(l)}(x|\theta) = \theta_3 n(\mathbf{s}) + \theta_4 n_r^{(l)}(\mathbf{s}), \quad l = 1, \ldots, L, \tag{17}$$

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

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

$$\widehat{(\mathbf{s}, \theta, v)} = \arg \max_{\Omega \times \Theta \times V} p_{\mathbf{d}}(\mathbf{s}, \theta, v) = \arg \max_{\Omega \times \Theta \times V} p_{\mathbf{d}}(\mathbf{s}|\theta, v)p(\theta)p(v). \tag{18}$$

The proposed construction is a mixture of bidimensional processes, as in the graphical detection made by hydrogeologist. Furthermore, as it will be shown just below, the sampling of $p_{\mathbf{d}}(s, \theta, v)$ can be done using a Metropolis-Hasting within Gibbs algorithm, allowing the computation of (18) based on a simulated annealing dynamics.

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

Theorem 3.1. *Let us assume for all $l \in [1, L]$ that $g^{(l)}(\mathbf{d}) > 0$. The HUG model with $K = 2$ is integrable for all $\theta_1, \theta_2, \theta_3, \theta_4 > 0$.*

The proof of this theorem is presented in the Appendix A.

The previous result allows to sample the HUG model with $K = 2$ with 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 (18) requires a global optimisation procedure. For this purpose, a simulated annealing algorithm may be implemented.

Here, the simulated annealing samples iteratively from $p(\mathbf{s}, \theta, v)^{1/T}$ while the temperature T goes slowly to 0.

The sampling from $p(\mathbf{s}, \theta, v)$ a MH within Gibbs dynamics. The priors $p(\theta)$ and $p(v)$ are chosen to be easy to sample. 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 bidimensional Hug model. This step is 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:

Algorithm SA : Fix $p(\theta)$ and $p(v)$. Choose a random initial configuration $\mathbf{s}_0$. The initial temperature is T_1, the cooling coefficient is c, the total number of iterations is N, G the number of applications of the Gibbs sampler 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 dataset and the parameter set-up for each of the presented algorithms are explained. Then the model is applied on synthetic datasets. The first synthetic dataset allows us to evaluate the results of the model when the real sources are known and visible on each projection plane. The second synthetic dataset accounts for a mixing of four sources and considers three hydrochemical parameters. The sources are positioned such as only 3 sources are visible on each projection plane. Finally, the Hug model is then applied on real datasets.

4.1 Data and parameters set-up

For all the datasets, 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 integrating prior knowledge regarding the threshold values δ_k and the radius r.

The algorithms parameters were chosen based on trial and error procedures. The SA algorithm was run for $N = 3.5 * 10^6$ iterations. For each iteration, the Gibbs dynamics was applied $G = L$ times. Each time the MH is called, it performs $M = 200$ 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 are performed at constant temperature. At this very low temperature, these last outputs of the algorithms may be considered closed enough to the desired solution (18). Furthermore, they tend to be identically distributed, 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	number of applications of the Gibbs sampler	L
M	number of steps in the MH algorithm	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 W. The algorithm outputs, that is the pattern of sources, its statistics and the parameters θ are saved every 1000 iterations. This gives a total of 1000 saved samples containing the detected patterns of sources and their associate sufficient statistics. Only the last 500 saved patterns are used for statistical inference.

For the prior $p(v)$ the uniform distribution was adopted. The prior $p(\theta)$ was chosen following a strategy similar to classical Approximate Bayesian Computation (ABC) principles [Blum, 2010]. The Hug model with different pre-fixed θ values was applied on several synthetic datasets 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 datasets

In synthetic datasets, the detected sources can be compared to the real known ones. Synthetic datasets are made by first setting the number of dimensions K, the number of sources n, their positions $\mathbf{s}^*$ and the number of samples m. The position 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 dataset contains $m = 200$ points resulting from the mixing system described in the Table 3. The data space is made by three 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: Position of the real sources ($\mathbf{s}^*$) for the first synthetic dataset.

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: Statistics of the real sources for the first synthetic dataset.

The Figure 2 shows the cumulative means of the statistics series. It can be observed that they clearly tend to approach constant close to the true statistics obtained from the known sources.

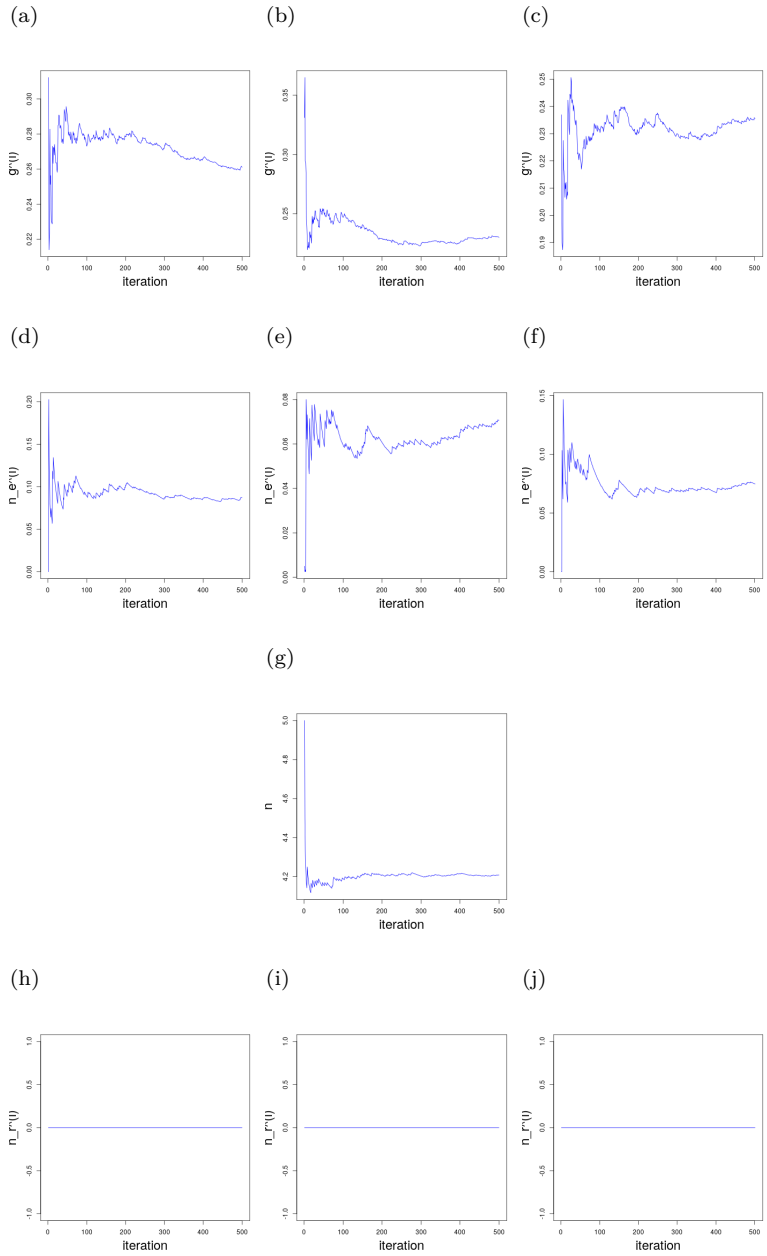


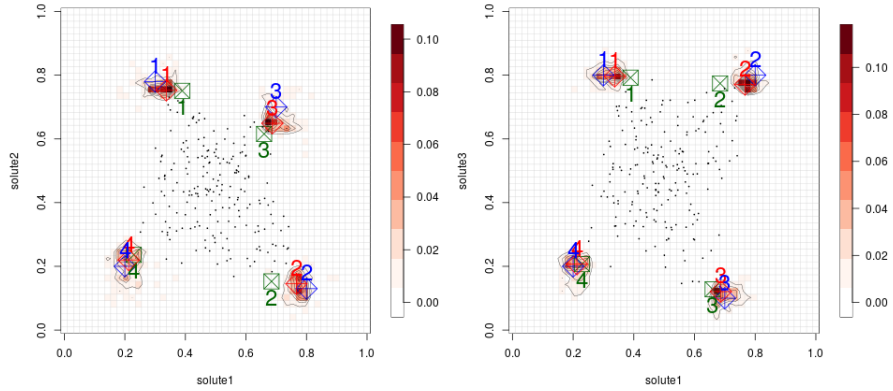
Figure 2: Cumulative means of the statistics for the first synthetic dataset: first plane (a,d,g,h), second plane (b,e,g,i) and third plane (c,f,g,j). Each row represents a statistic.

The patterns of the detected sources are projected on every plane. Following [Heinrich et al., 2012], level sets are estimated. The computation was done for a regular grid with cells of length 0.02. The probability of having a detected source in a cell is indicated by the coloured scale. This probability is estimated by (7).

The Figure 3 presents the obtained results. The model finds in each plane 4 regions associated with level sets exhibiting high probability values that indicate the presence of sources. On each plane, these regions are close to the real sources (in blue). These results match the behaviour of the statistics in Figure 2. The cumulative means of the statistics approach the corresponding values computed from the known sources. The detected sources are grouped in 4 clusters using a k-means algorithm. The position of clusters centres are shown in green in Figure 3. The median point of each cluster (represented in red) is made by considering the median coordinate of each cluster. The proposed pattern is the pattern of median points specified in the Table 5. This choice is adopted in order to diminish the impact of extreme values.

(a) first plane

(b) second plane



(c) third plane

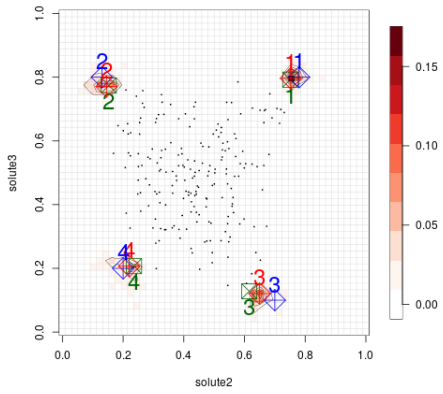


Figure 3: Level sets computed the first synthetic dataset. The blue symbols represent the known sources, the green symbols are the centres of the clusters obtained by the k-means algorithm with four clusters on the pattern of sources, and the red symbols are the median points.

Sources	solute1			solute2			solute3		
	median	mean	sd	median	mean	sd	median	mean	sd
1	0.69	0.7	0.03	0.64	0.65	0.03	0.11	0.11	0.02
2	0.77	0.77	0.02	0.13	0.13	0.03	0.76	0.77	0.02
3	0.21	0.21	0.02	0.22	0.21	0.04	0.21	0.2	0.03
4	0.33	0.33	0.03	0.76	0.76	0.02	0.8	0.8	0.02

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

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 difference. For each dimension, the average error is given (in %) by "Mean Error Dimension" and the average error for each source is given (in %) by "Mean Error Source" in the Table 6.

Sources	solute1	solute2	solute3	Mean Error Source
1	13.3	3.8	0.0	5.7
2	3.8	15.4	3.8	7.7
3	1.4	7.1	20.0	9.5
4	5.0	10.0	5.0	6.7
Mean Error Dimension	5.9	9.1	7.2	7.4

Table 6: Relative difference (in %) for the proposed sources with respect to the known ones, and the mean error for each source and each dimension, for the first synthetic dataset.

4.2.1 Estimation of the source position

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 dataset is such an example. The dataset contains 100 points obtained by a mixing system with 4 sources in a 3 dimensional space (see Table 7). The special feature of this dataset 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: Position of the real sources ($\mathbf{s}^*$) for the second synthetic dataset.

A method that estimates the sources position 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 dataset. 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 dataset. 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.

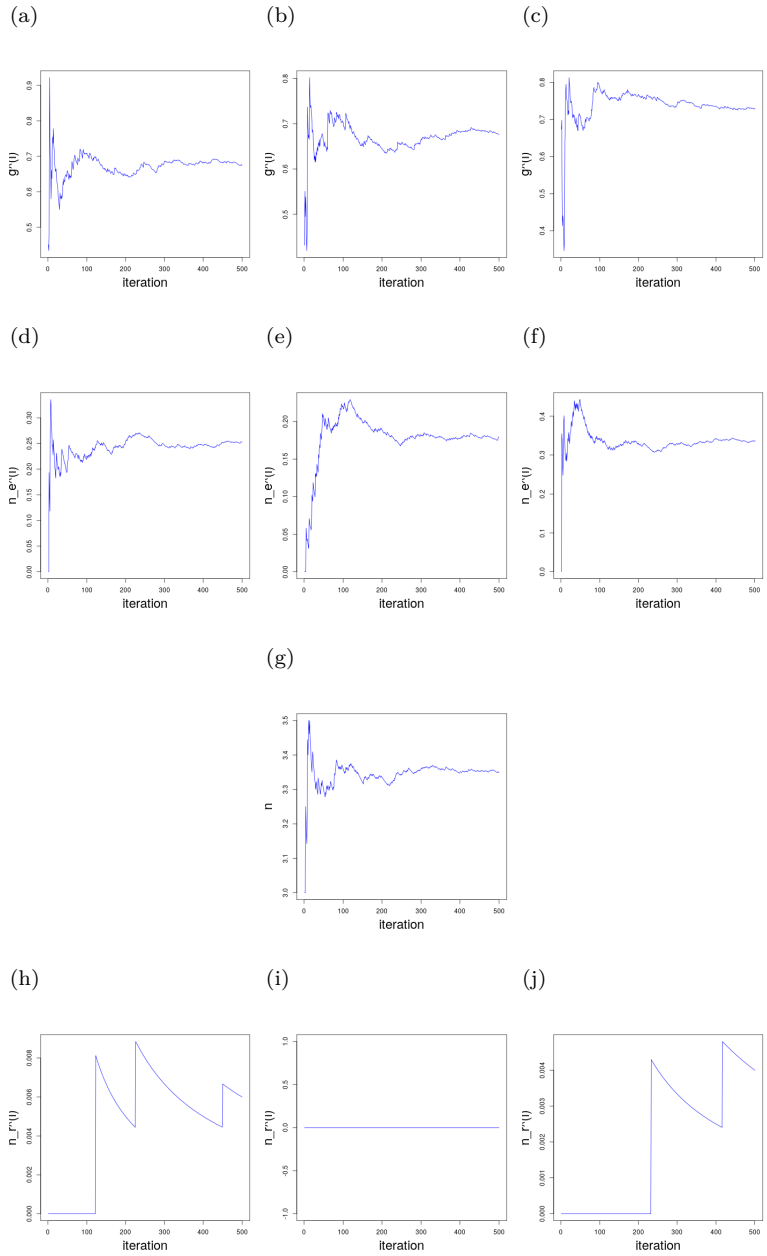
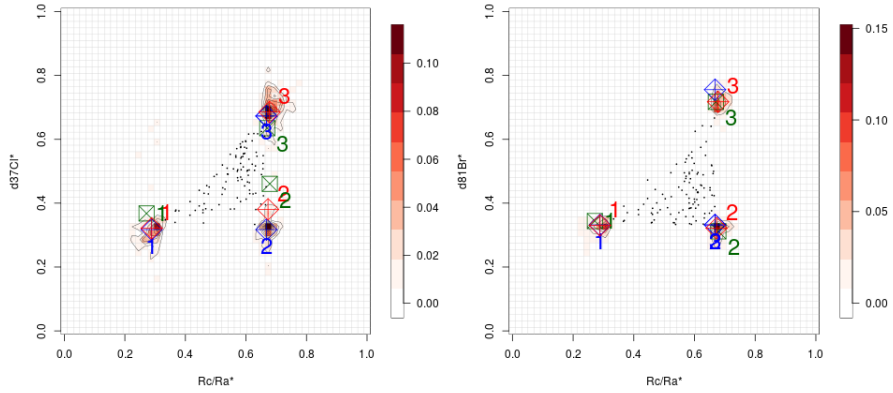


Figure 4: Cumulative means of the statistics for the second synthetic dataset: first plane (a,d,g,h), second plane (b,e,g,i) and third plane (c,f,g,j). Each row represents a statistic.

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 concluding on the exact number of sources and by extension their position.

(a) first plane

(b) second plane



(c) third plane

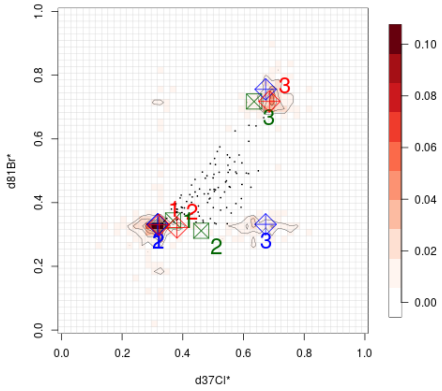


Figure 5: Level sets computed the second synthetic dataset. The blue symbols represent the known sources, the green symbols are the centres of the clusters obtained by the k-means algorithm with three clusters on the pattern of sources, and the red symbols are the median points.

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, described in the following, 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 $S = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$.

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

By applying the sequential k-means algorithm with 3 clusters in each plane, we obtain a pattern with 4

sources, rather close to the real sources, described in Table 8.

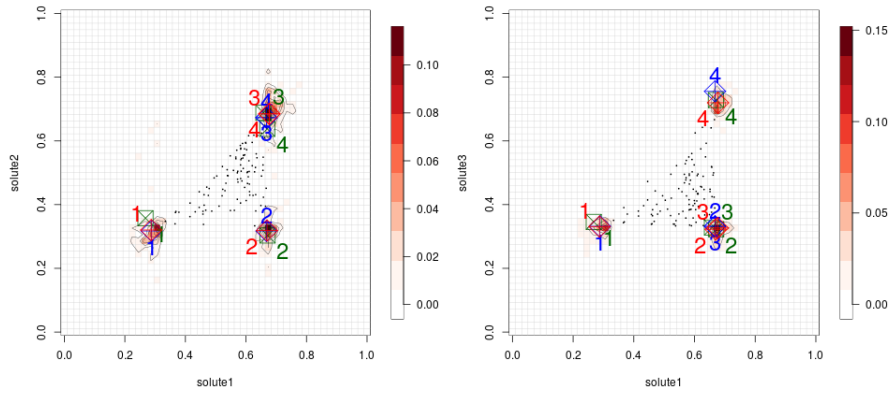
Sources	solute1	solute2	solute3
1	0.27	0.35	0.33
2	0.67	0.35	0.33
3	0.67	0.64	0.33
4	0.67	0.64	0.72

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.

(a) first plane

(b) second plane



(c) third plane

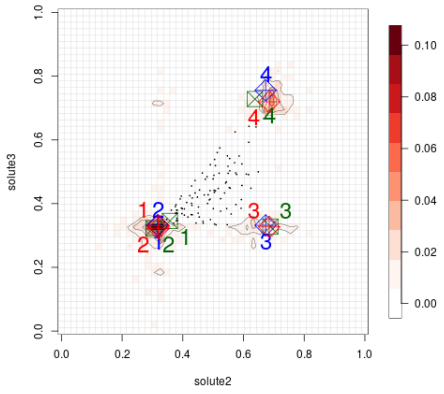


Figure 6: Level sets computed the second synthetic dataset. The blue symbols represent the known sources, the green symbols are the centres of the clusters obtained by the k-means algorithm with four clusters, and the red symbols are the median points.

Sources	solute1			solute2			solute3		
	median	mean	sd	median	mean	sd	median	mean	sd
1	0.68	0.66	0.09	0.68	0.64	0.13	0.71	0.71	0.07
2	0.67	0.65	0.06	0.32	0.32	0.07	0.33	0.36	0.11
3	0.3	0.29	0.06	0.33	0.37	0.12	0.33	0.35	0.1
4	0.66	0.63	0.1	0.64	0.63	0.08	0.33	0.35	0.08

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

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	2.8
Mean Error Dimension	0.8	0.8	2.1	1.2

Table 10: Relative difference (in %) for the proposed pattern of sources with respect to the known ones, and the mean error for each source and each dimension, for the second synthetic dataset.

The number of sources may also be deduced from the hierarchical clustering algorithm. This clustering is iterative and minimises in each step the within-cluster variance. At each step the cluster, made by merging two existing clusters, with the smallest variance is created. The algorithm ends when only one cluster remains. The within-cluster variance at each step is represented in the dendrogram of Figure 7.

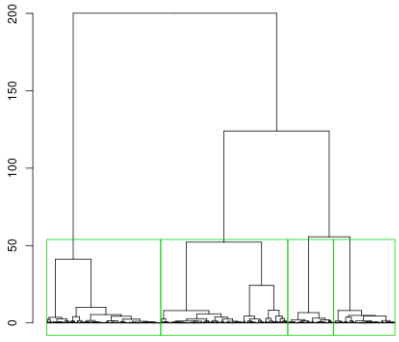


Figure 7: Dendrogram obtain from a hierarchical clustering that minimised the within-cluster variance for the second synthetic dataset. Each green rectangle contains the simulated sources that belong to one of the four clusters.

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.

The saved configurations 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. In each case, the four biggest clusters contain at least 75% of the sources: the sources can be clustered into four clusters.

	5	6	7	8	9
proportion	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 datasets

The previously described methods are applied to two real datasets.

4.3.1 First dataset :

The first real dataset on which the Hug model is applied is from [Pinti et al., 2020]. In this dataset the stable isotopic composition of chlorine δ^{37}Cl, the stable isotopic composition of bromine δ^{81}Br and the stable isotopic composition of helium Rc/Ra (^{3}He/^{4}He normalized to that of the Atmosphere and corrected for the air component) are measured on $m = 75$ samples from geothermal wells from Mexico and are supposed to be the result of a three-source mixing system (mantle, subduction and crust). 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 terms of area) containing the data on each plane (Table 12).

sources	Rc/Ra	δ^{37}Cl in ‰	Rc/Ra	δ^{81}Br 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: their value is the means between the two previous values. The reconstructed sources are described in Table 13.

sources	Rc/Ra	δ^{37}Cl in ‰	δ^{81}Br 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: Composition of the sources estimated in [Pinti et al., 2020] in the complete data space. The reconstruction is made by merging by the coordinate Rc/Ra. The value of this coordinate is the means between this coordinate on each plane.

The Hug model is applied with the same initialisation as in the previous section. As previously, the last 500 saved configurations 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. Figure 8 and Table

14 present 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 source. By applying the sequential k-means algorithm with 3 clusters on each plane, 3 sources are estimated. These sources are given by the k-means algorithm with 3 clusters.

(a) first normalised plane (b) second normalised plane

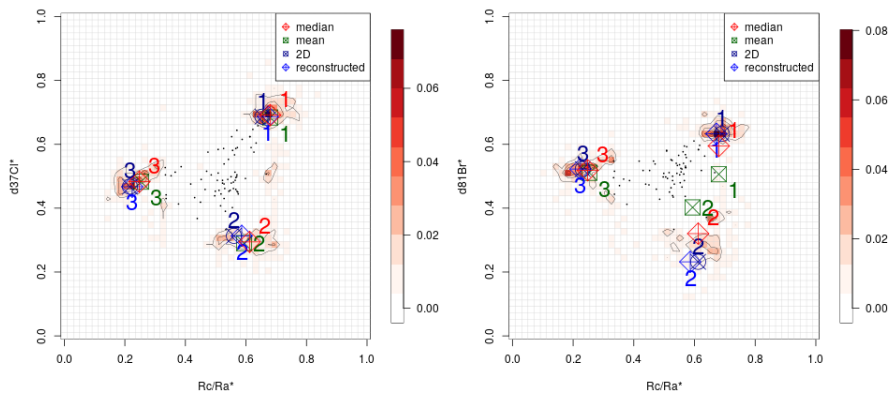


Figure 8: Level sets computed the first real dataset [Pinti et al., 2020]. The blue symbols represent the sources reconstructed by [Pinti et al., 2020], the green symbols are the centres of the clusters obtained by the k-means algorithm with three clusters on the pattern of sources, and the red symbols are the median points.

The proposed pattern of sources is not useful in this state because it can not be compared 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 15.

(a) first plane (b) second plane

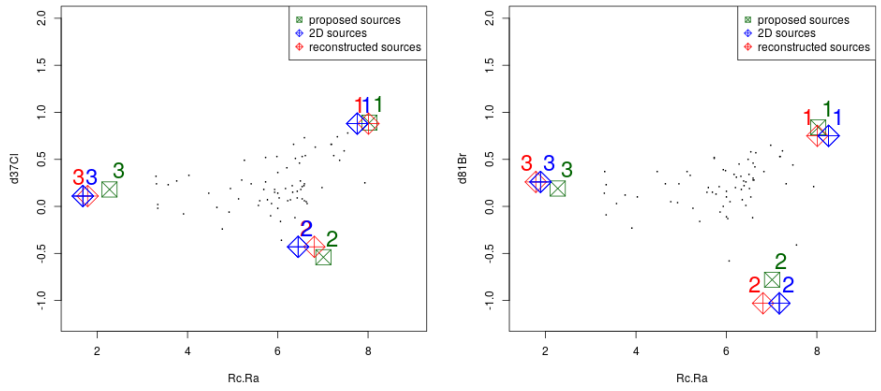


Figure 9: Composition of the sources proposed by the Hug model in the initial data space. The composition in δ^{37}Cl and δ^{81}Br are given in ‰.

Sources	Rc/Ra			δ^{37}Cl			δ^{81}Br		
	median	mean	sd	median	mean	sd	median	mean	sd
1 (mantle)	0.69	0.69	0.06	0.71	0.7	0.07	0.64	0.59	0.13
2 (subduction)	0.6	0.6	0.08	0.29	0.29	0.07	0.26	0.3	0.14
3 (crust)	0.23	0.24	0.07	0.48	0.47	0.06	0.52	0.52	0.06

Table 14: Proposed pattern for the first real dataset, made by clustering the sources in three clusters computed using a k-means algorithm, on the whole normalised data space.

sources	Rc/Ra	δ^{37}Cl in ‰	δ^{81}Br 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 15: Proposed pattern for the first real dataset made by the median points of clusters computed using a k-means with 3 classes on the initial data space (after reversing the normalisation).

The relative difference (in %) between the reconstructed sources and the proposed sources, after reversing the transformation, are given in Table 16.

Sources	Rc/Ra	δ^{37}Cl	δ^{81}Br	Mean Error Source
1	1.4	2.3	22.7	8.8
2	5.3	16.3	37.9	19.8
3	19.1	54.5	7.7	27.1
Mean Error Dimension	8.6	24.4	22.7	18.6

Table 16: Relative difference (in %) for the detected sources with respect to the sources estimated in [Pinti et al., 2020], and the mean error for each source and each dimension, for the first real dataset.

Eventually, applying the Hug model provides fairly consistent results with the geometrical approach proposed by [Pinti et al., 2020] on this rather simple dataset (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. As seen previously, working only on planes may induce bias. Thus, in [Pinti et al., 2020] at least 3 sources are detected (more sources may be needed) whereas the Hug model detect exactly 3 sources.

4.3.2 Second dataset :

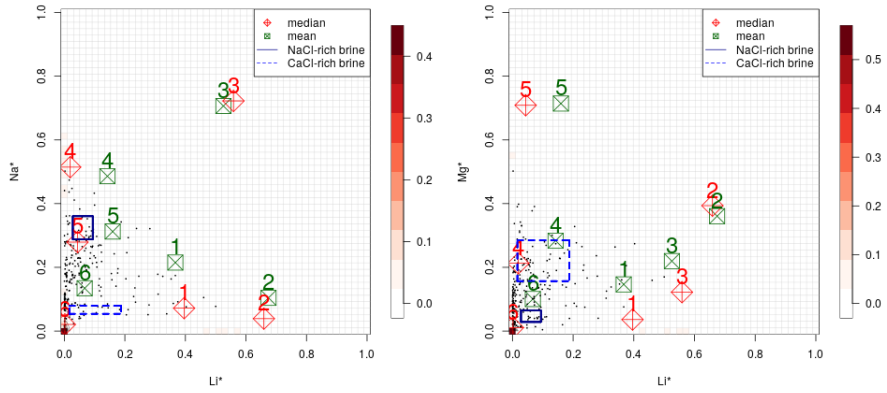
The second dataset considered is from [Richard et al., 2010], [Richard et al., 2016] and [Martz et al., 2019]. The dataset accounts for the composition of fluid inclusions from a series of hydrothermal uranium deposits of the Athabasca Basin (Canada). The concentrations of five chemical elements (lithium Li, sodium Na, magnesium Mg, potassium K and calcium Ca) are obtained by Laser Ablation-Inductively Coupled Plasma Mass Spectrometry (LA-ICPMS). Based on graphical interpretation of two-dimentional composition diagrams, previous authors have concluded that the data are spread between a "NaCl-rich-brine end-member" and a "CaCl_2-rich brine endmember" where NaCl-rich brines are definied by all fluid inclusions with [Na] $>$ 80000 ppm and CaCl_2-rich brines are defined as any fluid inclusions with [Na] $<$ 30000 ppm. The main conclusion was that the data result for a mixing of two sources. The full composition of the NaCl-rich and CaCl_2-rich brine end-members are given in Table 17. 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 Figures 10 and 11. On each plane, the model detects 3 areas with a rather high probability, indicating the presence of potential sources. By applying the sequential k-means with 3 clusters on each plane, 6 sources are detected.

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 17: 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).

(a) normalised plane 1

(b) normalised plane 2



(c) normalised plane 3

(d) normalised plane 4

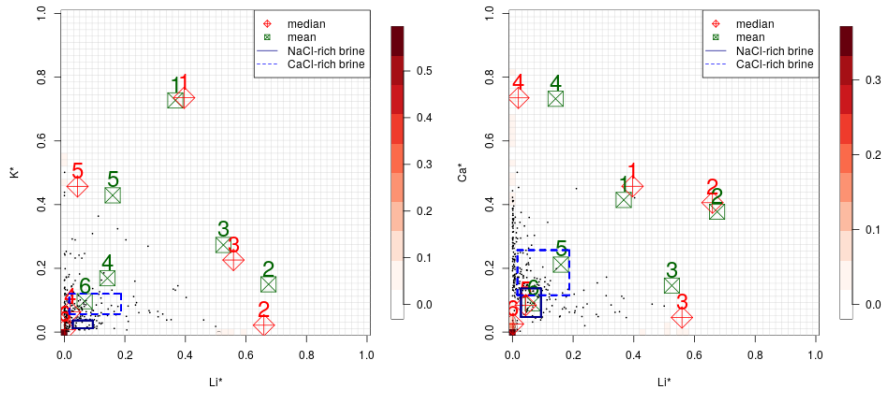
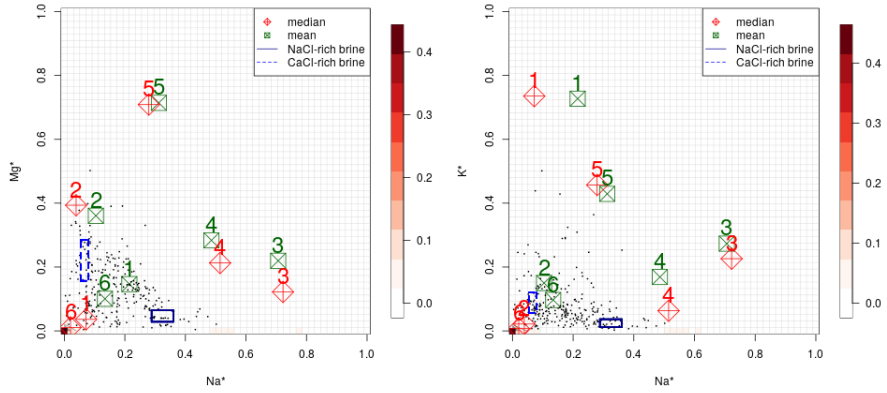


Figure 10: Level sets computed the second real dataset. The blue rectangles made by the continuous line and the dotted line represent respectively the NaCl-rich brine and the CaCl_2-rich brine presented in [Richard et al., 2016].

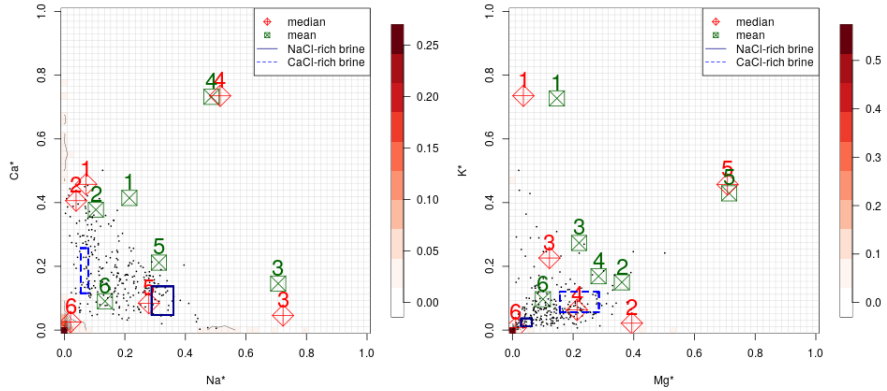
(a) normalised plane 5

(b) normalised plane 6



(c) normalised plane 7

(d) normalised plane 8



(e) normalised plane 9

(f) normalised plane 10

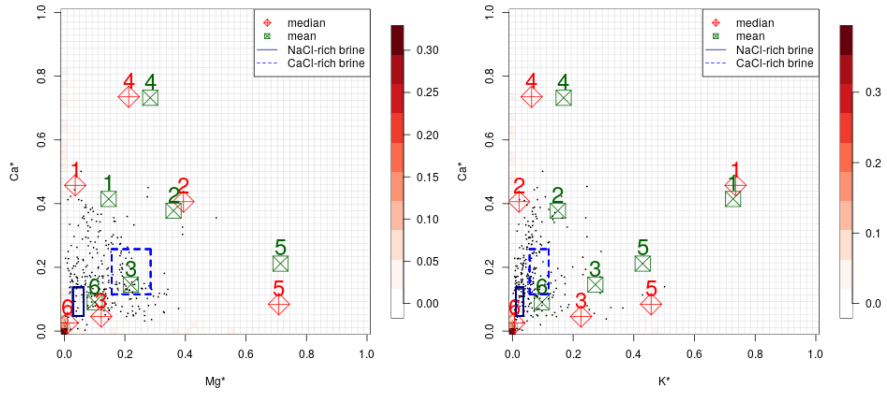


Figure 11: Level sets computed the second real dataset. The blue rectangles made by the continuous line and the dotted line represent respectively the NaCl-rich brine and the CaCl_2-rich brine presented in [Richard et al., 2016].

The hierarchical cluster algorithm is also applied to confirm the number of sources. As seen in Figure 12, 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.

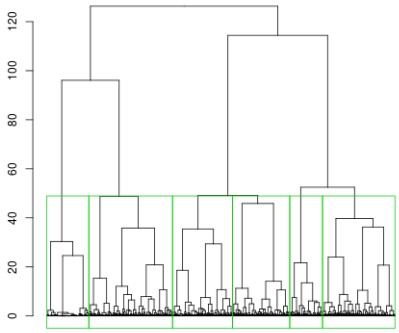


Figure 12: Dendrogram obtain from a hierarchical clustering that minimised the within-cluster variance. Each green rectangle contains the simulated sources that belong to one of the six clusters.

The proportion of sources in the 6 biggest clusters obtained by a k-means with 7, 8 and 9 clusters, is given in Table 18. The proportion is greater than 70%, hence the hypothesis of 6 clusters can be reasonably confirmed.

	7	8	9
proportion	0.90	0.79	0.72

Table 18: Proportion of sources in the 6 clusters containing the most simulated sources when the k-means is applied with 7, 8 and 9 clusters.

The proposed pattern of sources is given in Table 19.

	Li*			Na*			Mg*			K*			Ca*		
	med	mean	sd	med	mean	sd	med	mean	sd	med	mean	sd	med	mean	sd
1	0	0.15	0.21	0.03	0.15	0.2	0.01	0.08	0.14	0.55	0.55	0.14	0.03	0.12	0.18
2	0	0.05	0.12	0.02	0.1	0.17	0.59	0.59	0.14	0.02	0.18	0.23	0.04	0.15	0.2
3	0	0.09	0.18	0.55	0.55	0.19	0.01	0.09	0.16	0	0.08	0.15	0.01	0.04	0.06
4	0	0.15	0.22	0.01	0.04	0.07	0.01	0.13	0.19	0.01	0.11	0.18	0.63	0.64	0.15
5	0.57	0.61	0.15	0.05	0.15	0.2	0.07	0.22	0.24	0.01	0.09	0.16	0.02	0.1	0.16
6	0	0.1	0.2	0.56	0.58	0.15	0.02	0.17	0.23	0.02	0.16	0.24	0.61	0.63	0.16

Table 19: Proposed pattern for the second real dataset, made by clustering the sources in six clusters computed using a k-means algorithm, on the whole normalised data space.

The proposed pattern of sources is presented in the original data space in the Table 20 and the Figures 13 and 14.

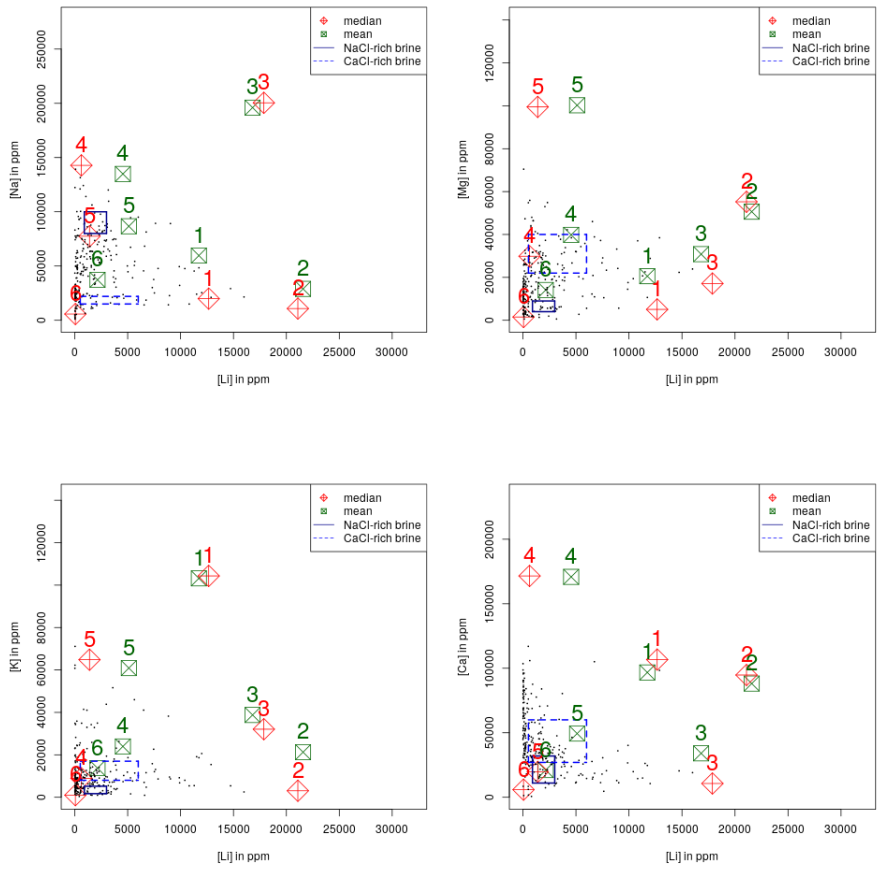


Figure 13: Composition of the proposed sources for the second real dataset, made by clustering the sources in six clusters computed using a k-means algorithm, on the first four planes (after reversing the normalisation). The blue rectangle and the dotted rectangle represent respectively the NaCl-rich brine and the CaCl_2-rich brine presented in [Richard et al., 2016].

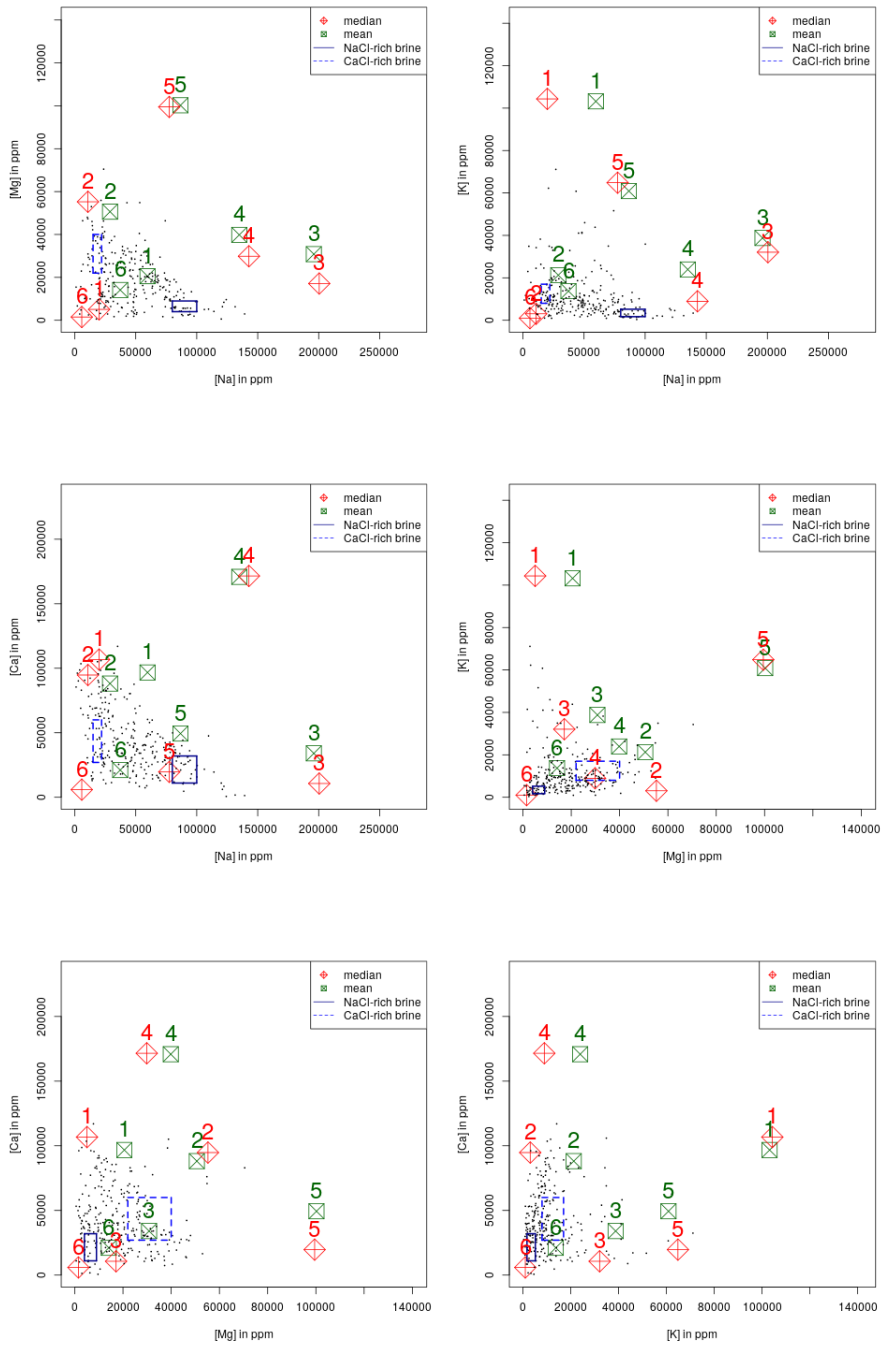


Figure 14: Composition of the proposed sources for the second real dataset, made by clustering the sources in six clusters computed using a k-means algorithm, on the last six planes (after reversing the normalisation). The blue rectangle and the dotted rectangle represent respectively the NaCl-rich brine and the $CaCl_2$-rich brine presented in [Richard et al., 2016].

On each plane the areas are: a first area (low abscissa values, low ordinate values), a second area (low abscissa values, high ordinate values) and a last area (high abscissa values, low ordinate values).

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 20: Composition of the proposed sources, made by clustering the sources in six clusters computed using a k-means algorithm, on the whole data space (after reversing the normalisation).

In [Richard et al., 2010], [Richard et al., 2016] and [Martz et al., 2019], the data are supposed the result of a mixing between two brine sources. The Hug model propose six sources. None of the six sources proposed by the HUG model matches consistently the composition of the NaCl-rich or the CaCl_2-rich brine end-members. The first explanation is that the HUG model detects sources that are close but outside the convex hull of the data, while the NaCl-rich and CaCl_2-rich brine represent composition end-members. The second explanation is that the HUG model uses consistent determination of sources across all considered elemental compositions (five dimensions and ten planes) while previous graphical detection used only four planes independently (Li versus Na; Ca versus Na; K versus Na and Mg versus Na).

As the HUG model detects sources without projections biases, one should reconsider the previous interpretations of the composition of fluid inclusions in the Athabasca basin. While the continuum of data is still compatible with mixing of different sources: the nature of the sources detected with HUG can be questioned: some of the detected sources may actually be linked to compositional data that suffered from perfectible analytical quality and therefore may be only considered as simple artefact. The HUG model can therefore be also used for detection of exotic data. In any case, the nomenclature and composition of the NaCl and CaCl_2-rich brine end-members must be reconsidered. In order to make the difference between the detected source that are real sources and the detected sources that result from exotic data, one should reconsider raw analytical data, which is beyond the scope of this manuscript.

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 multidimensional nature 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 estimated 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 multidimensionality 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 datasets. Numerical experiences done using known real datasets 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, the sensitivity of the model to the prior and the quality of the data is studied. New procedures for inferring the model parameters are also 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 an expert.

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. Currently, a thorough study concerning data uncertainties using Bayesian inference is under development.

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.

[Blum, 2010] Blum, M. G. (2010). Approximate bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187.

[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.

A Appendix: proof

Proof. Let a configuration $\mathbf{s} \in \Omega$. By hypothesis there is $M_1 \in \mathbb{R}$ such as

$$M_1 = \theta_1 \min_{t \in [0, \nu(W)]} \left\{ \left| \frac{t}{g^{(l)}(\mathbf{d})} - 1 \right| \right\}. \tag{19}$$

Hence

$$U^{(l)}(\mathbf{s}|\theta, \mathbf{d}) \geq M_1 + \theta_3 n(\mathbf{s}) + \theta_4 n_r^{(l)}(\mathbf{s}) \geq \theta_3 n(\mathbf{s}) + \theta_4 n_r^{(l)}(\mathbf{s}). \tag{20}$$

The energy function of the Hug model with $K = 2$ dominates the energy function of a Strauss point process. A Strauss point process is integrable if $\theta_4 > 0$ [Geyer, 1999, van Lieshout, 2000, Møller and Waagepetersen, 2003]. For $\mathbf{s} \in \Omega$ and $\xi \in W$ such as $s_i \neq \xi, \forall s_i \in \mathbf{s}$, its Papangelou conditional intensity is dominated by

$$\lambda^*(\mathbf{s}, \xi) = \exp[-\theta_3] \exp[-\theta_4]^{n_r(\mathbf{s} \cup \xi) - n_r(\mathbf{s})} \leq \exp[-\theta_3]. \tag{21}$$

The Hug model with $K = 2$ is integrable if $\theta_4 > 0$. □