A knowledge-driven way to interpret the isometric log-ratio transformation and mixture distributions of geochemical data

Xiangchong Liu, Wenlei Wang⁎, Yingru Pei, Pingping Yu

Institute of Geomechanics, Chinese Academy of Geological Sciences, Beijing, China
The Laboratory of Dynamic Diagenesis and Metallogenesis, Institute of Geomechanics, CAGS, China
Key Laboratory of Palaeomagnetism and Tectonic Reconstruction, Ministry of Natural Resources, Beijing, China

ARTICLE INFO

Keywords:
Isometric log-ratio transformation
Sequential binary partition
Compositional data
Mixture distributions
Duolong

ABSTRACT

When extracting geo-information from geochemical data, it is essential to consider the compositional nature and geological signatures of the data. Isometric log-ratio transformation (ILR) produces an orthonormal basis of geochemical data and accounts for the compositional nature of the data. However, it still often remains difficult to interpret ILR-transformed variables because of the lack of a geological basis for a data-driven approach; therefore, it is necessary to find some geological knowledge-based criteria to help enable a more understandable interpretation following ILR transformation. Characterized by certain elements and elemental ratios, the chronological order of geological processes can be used to construct interpretable ILR transformation. This concept was applied to extract geo-information from stream sediment geochemical data in the Duolong mineral district, Tibet, China. Furthermore, the expectation-maximization (EM) algorithm modified by a minimum message length criterion (MML) was employed to investigate mixture distributions of the geochemical data. In the study area, mafic rocks with high concentrations of Cr and Ni were emplaced earlier than the porphyry Cu–Au deposits which have stable Cu/Au ratios. Based on these criteria and hierarchical cluster analysis, sequential binary partition was constructed among the Cu, Au, Cr, and Ni concentrations. The ILR-transformed variables follow either a bi-normal or tri-normal distribution, the subpopulations of which were interpreted as fingerprints inherited from mafic magmatic processes, Cu–Au hydrothermal systems, and later geological processes, respectively. The high-average subpopulation of ILR-transformed Cu corresponds to anomalies associated with the porphyry and epithermal Cu–Au systems, and two areas are predicted to have high Cu–Au mineralization potential. This study demonstrates that geological knowledge-driven ILR transformation is a promising way to efficiently extract geo-information from geochemical data.

1. Introduction

The probability distribution of geochemical data is a classic and important topic in geochemistry because of its fundamental roles in tracing underlying geological processes, separating background and anomaly and performing further statistical analysis (e.g. Ahrens, 1954; Allegre and Lewin, 1995; Buccianti et al., 2018; Buccianti et al., 2006; Mateu-Figueras and Pawlowsky-Glahn, 2008; Reimann et al., 2005; Vistelius, 1960). One of the most broadly discussed distributions in exploration geochemistry is the normal distribution which has greatly supported geochemical data-based exploration in various fields (Albarede, 2003; Vistelius, 1960). However, for a long-term practice, it has been noted that frequency distributions of trace elements often fail to follow a normal distribution because of positive skewness, the presence of outliers, and mixtures of subpopulations (e.g. Reimann and Filzmoser, 2000; Zhang et al., 2005). Ahrens (1954) proposed that distributions of trace elements in crustal rocks follow a log-normal distribution, which is still introduced in many modern textbooks of geochemistry (e.g. Albarede, 2003; Parslow, 1984). Subsequently, many studies proposed and discussed that geochemical data follow a fractal distribution (e.g. Agterberg, 1995; Cheng et al., 1994; Mandelbrot, 1967; Turcotte, 1986; Zuo and Wang, 2016). Moreover, based on the concept that element concentrations of sampled rocks are often end products of multiple geological and geochemical processes, it was proposed that trace elements may follow a multi-model distribution or a multi-fractal distribution (e.g. Agterberg, 2007; Allegre and

⁎ Corresponding author at: Institute of Geomechanics, Chinese Academy of Geological Sciences, Beijing, China.
E-mail addresses: xcliu@cags.ac.cn (X. Liu), wenleiw@163.com (W. Wang).

https://doi.org/10.1016/j.gexplo.2019.106417
Received 8 January 2019; Received in revised form 18 October 2019; Accepted 11 November 2019
0375-6742/ © 2019 Elsevier B.V. All rights reserved.

Please cite this article as: Xiangchong Liu, et al., Journal of Geochemical Exploration, https://doi.org/10.1016/j.gexplo.2019.106417
Lewin, 1995). Therefore, it can be concluded that finding a robust quantitative tool is a prerequisite for separating the mixture distributions of geochemical data and interpreting their geological signatures (e.g., Sinclair, 1974).

Another significant issue in geochemical exploration is the compositional nature of geochemical data (e.g., Bucciatti and Zuo, 2016). Geochemical data reported in the form of proportions (wt%, ppb, or ppm) are typical compositional data (Pawlowsky-Glahn and Bucciatti, 2011). Measured geochemical concentration values are relative rather than absolute and subject to a constant sum (e.g. 100% or 1) (Pawlowsky-Glahn and Egozcue, 2006). Consequently, analytical results based on standard statistical methods may obtain misleading conclusions, which has been referred as the closure effect (Pawlowsky-Glahn and Egozcue, 2006). From a mathematical perspective, the geochemical data discussed above are in the simplex and different from those in real Euclidean space (Pawlowsky-Glahn et al., 2015). That is: the balance is defined as: $b_i = \frac{\tau_i s_n \ln g(x_{i+}) - \ln g(x_{i-})}{\tau_i s_n + \tau_i s_n}$ (1)

where $\tau_i$ and $s_n$ are the numbers of variables termed as the numerator (i.e., $+1$) and the denominator ($-1$), respectively; $g(x_{i+})$ and $g(x_{i-})$ are geometric means of the numerator groups and denominator groups, respectively. In general, the balances are often constructed based on expert knowledge or data-driven methods. The data-driven methods introduced in previous studies include but not limited to variation matrix, hierarchical cluster analysis, and principle component analysis (e.g., Liu et al., 2018; Martin-Fernández et al., 2018; Thiombane et al., 2018). In this study, data-driven methods (e.g., log-ratio variation and hierarchical cluster analysis) and some geological criteria were jointly used to construct new interpretable variables using the SBP method.

2.2. Log-ratio variation and hierarchical cluster analysis

Let $x = [x_1, x_2, ... x_n]$ represent concentrations of $n$ elements from the geochemical data with $m$ samples. The concentration of the $i$th element can be expressed as $x_i = [x_{i1}, x_{i2}, ... x_{im}]$. The variation $t_{ij}$ measures the association of two elements (i.e. $\ln \frac{x_i}{x_j}$) (Aitchison, 1986):

\[ t_{ij} = \text{var}(\ln \frac{x_i}{x_j}) \] (2)

where $t_{ij} = 0$ only if $x_i$ and $x_j$ are linearly proportional.

Hierarchical cluster analysis based on the variation matrix of compositional data is used to cluster variables and aids in constructing ILR transformations (e.g. Liu et al., 2018). Using the open-source R packages 'compositions' and 'flashClust' (Boogaart and Tolosa-Delgado, 2013), hierarchical cluster analysis was applied to geochemical data from Duolong. The distance $d_{ij}$ of two groups was determined by the Ward2 method, which calculates the squared Euclidean distance of two groups (Murtagh and Legendre, 2014). The value $d_{ij} = 0$ means that the two groups have a perfect association. The two groups are of perfect independence as $d_{ij}$ moves towards $+\infty$. The strategy of hierarchical cluster analysis is that all elements are initially treated as isolated groups. Then, the two nearest groups are merged until all elements are pooled into one single group (Boogaart and Tolosa-Delgado, 2013).

2.3. An algorithm for fitting mixture distributions to geochemical data

The statistical distribution of geochemical data often contains at least two populations because of the influences of multiple geological and geochemical processes (e.g. Carranza, 2009; Grunsky and Smeek, 1999; Liu et al., 2011; Sinclair, 1991). Sinclair (1974) used probability plots to estimate the threshold between anomalous and background geochemical data. The expectation-maximization (EM) algorithm is a common method to fit finite mixture models to observed data (Hsu et al., 1986; McLachlan and Peel, 2004). However, this algorithm is sensitive to initial values of iterations because the likelihood function of a mixture model is not unimodal. For certain types of mixtures, the EM algorithm may converge to the boundary of the parameter space and further lead to meaningless estimation.

Figueiredo and Jain (2002) modified the EM algorithm using a minimum message length criterion (MML), with their algorithm (MML-
EM) avoiding the drawbacks of the EM algorithm. The $i$th element $x_i$ follows a $k$-component finite mixture distribution if its probability density function $p(\theta)$ can be written as (Figueiredo and Jain, 2002):

$$p(\theta) = \sum_{j=1}^{k} \lambda_j f(\mu_j, \sigma_j)$$

where $\theta = (k, \lambda_j, \mu_j, \sigma_j)$ is the parameter set; $k$ stands for the number of the components; $f(\mu_j, \sigma_j)$ stands for the probability density of the $j$th component; $\mu_j$ is the mean; $\sigma_j$ is the standard variation; $\lambda_j$ is the weight. The weight $\lambda_j$ must satisfy

$$\sum_{j=1}^{k} \lambda_j = 1, \lambda_j \geq 0$$

Similar to the EM algorithm, the first step of the MML-EM algorithm is to estimate the conditional expectation of the log-likelihood and further to update the parameter set $\theta$ by the maximum likelihood criterion. Second, the parameter set's message length is calculated (for its definition see Eq. (8) in Figueiredo and Jain, 2002). Third, iteration is applied to update $\theta^{t+1}$ from the previous estimation $\theta^t$ until the message length is minimized. The optimal parameter set $\theta$ will be the one with the minimum message length. The MML-EM algorithm is insensitive to initial values of the parameter set $\theta$ (Figueiredo and Jain, 2002). Liu et al. (2011) compared the MML-EM algorithm with the minimum message length. The MML-EM algorithm is in-

3. Case study

3.1. Study area and geological knowledge for modeling

The Duolong mineral district is located on the south rim of the southern Qiangtang Terrane and the north part of Bangongco–Nujiang suture zone (Fig. 1). The exposed magmatic rocks are granite porphyry, granodiorite porphyry, diorite, diorite porphyrite, basalt, and gabbro. The strata in this district consist of the Upper Triassic Riganpeicuo Formation (T3r), the Lower Jurassic Quse Formation (J1q), the Middle Jurassic Sewa Formation (J2s), the Lower Cretaceous Meirigecuo Formation (K1m), the Upper Cretaceous Abushan Formation (K1a), and the Upper Oligocene Kangtuo Formation (E1k). The Riganpeicuo Formation (T3r) is mainly composed of limestone in the northeast of Duolong. The Quse Formation (J1q) is composed of dark gray mudstone and feldspar–quartz siltstone. The Meirigecuo Formation (K1m) is mainly composed of continental-face intermediate–basic volcanic rocks containing andesite, andesitic volcanic breccia, rhyolitic tuff, and tholeiite (Li et al., 2012).

Ten porphyry and epithermal Cu–Au deposits discovered in the last decade are the Dibao, Nadun, Bolong, Duobuza, Rongna, Naruo, Saijiao, Sena, Tiegelong, and Gaerqin deposits (Li et al., 2012). Formation of these deposits is genetically related to porphyritic granitoids emplaced below these deposits (Li et al., 2016a, 2016b; Zhu et al., 2015). The hydrothermal alteration is divided into potassic, intermediate argillic and propylitic alteration zones from outward and upward of the ore-bearing porphyry (Li et al., 2016b). This is consistent with the typical alteration model of porphyry Cu–Au deposits (Sillitoe, 2000). Surficial argillic and propylitic alteration zones are direct indicators of ancient magmatic–hydrothermal systems at depth (Wang et al., 2015). In addition, a few placer gold deposits were also discovered far from the ten porphyry and epithermal Cu–Au deposits (see Fig. 1 in Li et al., 2008).

Two pieces of geological knowledge about the Duolong mineral district have been employed as a guide to construct interpretable ILR transformations. The first one is the chronological order of geological processes in Duolong. The Cu–Au deposits in this area are dated at 118–115 Ma (see references in Xu et al., 2017), later than the inter-

mediate and felsic intrusions (116–128 Ma) (Chen et al., 2013; Li et al., 2013; Li et al., 2014; Li et al., 2015; She et al., 2009; Wei et al., 2016). Two pieces of evidence suggest that these mafic rocks were also emplaced earlier than the Cu–Au ore-forming processes in Duolong. The first is that the mafic rocks in Duolong were interpreted to be fragments of older oceanic crust (Duan et al., 2013). The second is that the zircon U–Pb dating ages (126–127 Ma) of those mafic rocks are older than the molybdenite Re-Os ages (121.2 ± 1.2 Ma) and the sericite 40Ar–39Ar ages (115.2 ± 1.2 Ma) of the Cu–Au mineralization (Li et al., 2011; Lin et al., 2017; Song et al., 2018; Xu et al., 2017). Therefore, the chronological order of those geological processes was as follows: the mafic magmatic processes, the intermediate and felsic magmatic processes, and the Cu–Au mineralization processes.

The second piece of geological information used for the ILR transformation is that porphyry Cu–Au deposits have stable Cu/Au ratios at different scales. Microanalysis of trace elements of fluid inclusions in two of the largest porphyry Cu–Au deposits in the world suggests that
the Cu/Au ratios (~10^4) of mineralizing fluids are identical to the bulk Cu/Au ratios of ore bodies (Ulrich et al., 1999). That similarity is because these two elements are transported and precipitated together during the ore-forming processes (Ulrich et al., 1999). The Cu/Au ratios of the porphyry Cu–Au deposits in Duolong within a range of 2 × 10^3 – 3.9 × 10^4 (Song et al., 2018) have the same order of magnitude as the data reported by Ulrich et al. (1999). Thus, the Cu/Au ratios of the stream sediment in Duolong can be used to extract geo-information associated with the ore-forming processes and later geological processes.

Exploratory datasets used in this study include geological data and stream sediment geochemical data. The geological data include the locations of ten porphyry and epithermal Cu–Au deposits, fault traces, lithological units, and outcrops of mafic rocks. The geochemical data at 1:50,000 scale with 0.5 km spatial resolution are composed of 3217 samples, and concentration values of 15 trace elements (Cu, Au, Pb, Zn, Cr, Ni, Mn, Ag, Sn, W, Mo, As, Sb, Bi, and Hg) are recorded for each sample. More detailed descriptions of these datasets can be found in Wang et al. (2017).

3.2. The SBP-based modeling process

From the variation matrix (Table 1), the log-ratios of Cr over Ni have a variation of 0.05 indicating that their ratios vary little compared with the mean. In addition, Cr and Ni have relatively large variations with the other thirteen elements, especially Au, Cu, Bi, Mo, As, and Sb (Table 1). Consequently, Cr and Ni are separated first in the dendrogram (Fig. 4). The ore element Cu against Ag and Mo has low variations of 0.14 and 0.19, respectively. The variation of Cu/Au of 0.5 is therefore, the dendrogram calculated from the variation matrix reflects that Au is removed from the remaining thirteen elements, and Hg is separated in the next order. The remaining multi-element associations are Sn–W–Zn–Mn, Mo–Cu–Bi, and Pb–Ag–As–Sb.

The spatial distribution of Cr is very similar to that of Ni (Fig. 2), but significantly different from those of Cu and Au (Fig. 3). Cr and Ni concentrations are very low in the most parts of Duolong and strong anomalies are present mainly in the northeast corner where limestone and serpentinitized olivinite are found in outcrops (Figs. 1 and 2). Given that mafic and ultramafic rocks are often characterized by high concentrations of Cr and Ni (Oze et al., 2004), high concentrations of Cr and Ni in the northeast corner were interpreted to be related to the serpentinitized olivinite. The small-scale scattered anomalies of Cr and Ni in the other areas may be associated with small-scale basalt and gabbro exposed in Duolong (Fig. 1). Thus, the mafic magmatic processes that occurred before the Cu–Au mineralization in Duolong are characterized by high concentrations of Cr and Ni.

The data analysis above and the two pieces of geological knowledge noted above (see Section 3.1) were used to construct the sequential binary partitions shown in Table 2. The ILR-transformed variables are as follows:

\[ b_1 = \ln \frac{\sqrt{\text{CuCr}}}{\sqrt{\text{AuNi}}} - \ln \frac{\text{CuCr}}{\text{AuNi}} = \frac{1}{2} \ln \frac{\text{Cu}}{\text{Au}} + \frac{1}{2} \ln \frac{\text{Cr}}{\text{Ni}} \]  

\[ b_2 = \frac{1}{\sqrt{2}} \ln \frac{\text{Cu}}{\text{Cr}} \]  

\[ b_3 = \frac{1}{\sqrt{2}} \ln \frac{\text{Au}}{\text{Ni}} \]

Because Cr is highly linearly proportional to Ni (see Table 1), \( b_1 \) will be positively correlated with the log-ratio of Cu/Au, which describes geochemical signatures associated with hydrothermal Cu–Au systems and later geological processes in Duolong. In addition, \( b_2 \) can be interpreted as geochemical signatures caused by the mixture of mafic magmatic processes (Cr concentrations), hydrothermal Cu–Au systems (Cu concentrations), and later geological processes. Similarly, \( b_3 \) represents the mixture of mafic magmatic processes (Ni concentrations), hydrothermal Cu–Au systems (Au concentrations), and later geological processes. As described in the next section, the mixture distributions of these three variables were fitted using the MML-EM algorithm to extract the geo-information associated with geological processes before and after the Cu–Au mineralization in Duolong.

3.3. Mixture distributions of balances

The MML-EM algorithm was employed to fit the probability density functions of the geochemical data from the Duolong mineral district. The crossover points of the probability density functions of mixture distributions were further used to constrain the range of each subpopulation under the concept of probability.

The probability density curves (Fig. 5) show that \( b_1 \) follows a bi-normal distribution. The two components have weights of 0.63 and 0.37, respectively (Table 3). The highly weighted component has a higher average than the component with a lower weight. The probability density curves of the two components intersect at 1.35. The data points lower than 1.35 were grouped to the low-average subpopulation and the others were grouped to the high-average one. From the spatial distribution of these two subpopulations, the high-average one is scattered in the most parts of the Duolong area, whereas the low-average one is distributed around a few Cu–Au deposits (Fig. 6).

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Cu</th>
<th>Pb</th>
<th>Zn</th>
<th>Cr</th>
<th>Ni</th>
<th>Mn</th>
<th>Ag</th>
<th>Sn</th>
<th>W</th>
<th>Mo</th>
<th>As</th>
<th>Sb</th>
<th>Bi</th>
<th>Hg</th>
<th>Au</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cu</td>
<td>0.26</td>
<td>0.20</td>
<td>0.96</td>
<td>0.84</td>
<td>0.28</td>
<td>0.14</td>
<td>0.23</td>
<td>0.25</td>
<td>0.19</td>
<td>0.27</td>
<td>0.33</td>
<td>0.23</td>
<td>0.48</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>Pb</td>
<td>0.10</td>
<td>0.10</td>
<td>0.89</td>
<td>0.79</td>
<td>0.19</td>
<td>0.06</td>
<td>0.18</td>
<td>0.21</td>
<td>0.22</td>
<td>0.15</td>
<td>0.21</td>
<td>0.19</td>
<td>0.40</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Zn</td>
<td>0.09</td>
<td>1.00</td>
<td>0.74</td>
<td>0.61</td>
<td>0.10</td>
<td>0.10</td>
<td>0.12</td>
<td>0.19</td>
<td>0.26</td>
<td>0.16</td>
<td>0.22</td>
<td>0.20</td>
<td>0.36</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>Cr</td>
<td>0.86</td>
<td>0.96</td>
<td>0.04</td>
<td>0.05</td>
<td>0.67</td>
<td>0.87</td>
<td>0.65</td>
<td>0.69</td>
<td>0.82</td>
<td>0.81</td>
<td>0.80</td>
<td>0.90</td>
<td>0.59</td>
<td>1.34</td>
<td></td>
</tr>
<tr>
<td>Ni</td>
<td>0.92</td>
<td>0.98</td>
<td>0.41</td>
<td>0.42</td>
<td>0.75</td>
<td>0.75</td>
<td>0.55</td>
<td>0.60</td>
<td>0.74</td>
<td>0.70</td>
<td>0.71</td>
<td>0.81</td>
<td>0.52</td>
<td>1.23</td>
<td></td>
</tr>
<tr>
<td>Mn</td>
<td>3.04</td>
<td>3.15</td>
<td>2.14</td>
<td>2.18</td>
<td>2.95</td>
<td>0.18</td>
<td>0.15</td>
<td>0.20</td>
<td>0.30</td>
<td>0.21</td>
<td>0.25</td>
<td>0.31</td>
<td>0.32</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Ag</td>
<td>5.85</td>
<td>5.75</td>
<td>6.57</td>
<td>6.71</td>
<td>5.95</td>
<td>8.90</td>
<td>0.16</td>
<td>0.18</td>
<td>0.19</td>
<td>0.16</td>
<td>0.21</td>
<td>0.16</td>
<td>0.39</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>Sn</td>
<td>2.62</td>
<td>2.52</td>
<td>3.52</td>
<td>3.48</td>
<td>2.71</td>
<td>5.66</td>
<td>3.24</td>
<td>0.11</td>
<td>0.19</td>
<td>0.19</td>
<td>0.22</td>
<td>0.20</td>
<td>0.26</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>2.77</td>
<td>2.67</td>
<td>3.67</td>
<td>3.63</td>
<td>2.87</td>
<td>5.82</td>
<td>3.08</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.24</td>
<td>0.25</td>
<td>0.21</td>
<td>0.29</td>
<td>0.68</td>
</tr>
<tr>
<td>Mo</td>
<td>3.55</td>
<td>3.45</td>
<td>4.45</td>
<td>4.41</td>
<td>3.64</td>
<td>6.59</td>
<td>2.31</td>
<td>0.16</td>
<td>0.17</td>
<td>0.17</td>
<td>0.27</td>
<td>0.28</td>
<td>0.23</td>
<td>0.39</td>
<td>0.63</td>
</tr>
<tr>
<td>As</td>
<td>0.18</td>
<td>0.08</td>
<td>1.08</td>
<td>1.04</td>
<td>0.27</td>
<td>3.22</td>
<td>5.67</td>
<td>2.44</td>
<td>2.59</td>
<td>3.37</td>
<td>0.14</td>
<td>0.19</td>
<td>0.37</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>Sb</td>
<td>3.22</td>
<td>3.12</td>
<td>4.12</td>
<td>4.08</td>
<td>3.32</td>
<td>6.27</td>
<td>2.63</td>
<td>0.61</td>
<td>0.45</td>
<td>0.32</td>
<td>0.35</td>
<td>0.25</td>
<td>0.36</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>Bi</td>
<td>4.43</td>
<td>4.33</td>
<td>5.33</td>
<td>5.29</td>
<td>4.52</td>
<td>7.47</td>
<td>1.42</td>
<td>1.81</td>
<td>1.66</td>
<td>0.88</td>
<td>4.25</td>
<td>1.21</td>
<td>0.45</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>Hg</td>
<td>7.14</td>
<td>7.03</td>
<td>8.03</td>
<td>7.99</td>
<td>7.23</td>
<td>10.18</td>
<td>1.28</td>
<td>4.52</td>
<td>4.36</td>
<td>3.59</td>
<td>6.96</td>
<td>3.91</td>
<td>2.71</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Au</td>
<td>2.51</td>
<td>2.41</td>
<td>3.41</td>
<td>3.37</td>
<td>2.61</td>
<td>5.56</td>
<td>3.34</td>
<td>0.11</td>
<td>0.26</td>
<td>1.04</td>
<td>2.33</td>
<td>0.71</td>
<td>1.52</td>
<td>4.62</td>
<td></td>
</tr>
</tbody>
</table>

Note that the variation of ln \( \frac{X}{Y} \) equals that of ln \( \frac{Y}{X} \).
In contrast, $b_2$ follows a tri-normal distribution (Fig. 7), and the probability density functions of the three subpopulations intersect at $-0.25$ and $-1.27$. The transformed data higher than $-0.25$, between $-0.25$ and $-1.27$, and lower than $-1.27$ were grouped to the high-average, medium-average, and low-average subpopulations, respectively. From the spatial distribution of the three subpopulations (Fig. 8), the low-average subpopulation lies mainly in the northeastern corner of Duolong. The medium-average subpopulation occupies most parts of the area (Fig. 8). The high-average subpopulation is close to the Cu–Au deposits (Fig. 8).

Finally, $b_3$ contains three normal subpopulations, whose probability density functions meet at $-1.55$ and $-2.80$ (Fig. 9). These two crossover points divide the data into three subpopulations. Their spatial distribution shows that the low-average subpopulation lies mainly in the northeast of Duolong. The high-average subpopulation is widely distributed around the ten Cu–Au deposits (Fig. 10).

Fig. 2. The spatial distribution of the concentrations of Cu and Au in the Duolong. The concentrations of Cu and Au have a unit of parts-per-million and part-per-billion, respectively.

Fig. 3. The spatial distribution of the concentrations of Cr and Ni in the Duolong. The concentrations of Cu and Au have a unit of parts-per-million.
3.4. Discussion

Geochemical elements are often analyzed to characterize the geochemical signatures of certain geological processes (White, 2013). Consequently, the chronological order of geological processes constrained by radiometric dating and other dating methods (Reiners et al., 2017; Schmitz and Kuiper, 2013) can be a knowledge-based criterion for constructing SBP-based ILR-transformed balances (i.e., variables). In the Duolong mineral district, mafic rocks with high concentrations of Cr and Ni were emplaced earlier than the porphyry and epithermal Cu–Au deposits characterized by stable Cu/Au ratios. This case study shows that elemental ratios indicative of geological processes also have great potential to construct interpretable ILR-transformation and extract geo-information.

Commonly used geological criteria often group elements with similar geochemical properties when employing the SBP method. For example, Cu and Au are put into one group, whereas Cr and Ni are grouped as another group. The three new variables are

\[
v_1 = \ln \frac{\text{Cu}}{\text{Cr}} - \ln \frac{\text{Au}}{\text{Ni}}
\]

\[
v_2 = \frac{1}{2} \ln \frac{\text{Cu}}{\text{Cr}} + \frac{1}{2} \ln \frac{\text{Au}}{\text{Ni}}
\]

\[
v_3 = \frac{1}{2} \ln \frac{\text{Cu}}{\text{Cr}} - \frac{1}{2} \ln \frac{\text{Au}}{\text{Ni}}
\]

In \(v_1\), \(\frac{1}{2} \ln \frac{\text{Cu}}{\text{Cr}}\) and \(\frac{1}{2} \ln \frac{\text{Au}}{\text{Ni}}\) are linearly proportional to \(b_2\) (Eq. (4)) and \(b_3\) (Eq. (5)), respectively. Thus, \(v_1\) mixes the geo-information carried by \(b_2\) and \(b_3\); \(v_2\) traces the compositional changes of mafic rocks with little valuable information for the Cu–Au mineralization in Duolong; and \(v_3\) is...
Therefore, less valuable geo-information for prospecting Cu–Au mineralization would be extracted from these variables than the ones used in this study.

Ten porphyry and epithermal Cu–Au deposits have been discovered in the Duolong mineral district, and it is still commonly believed that this area has great potential for Cu–Au resources. From current results, the Cu–Au deposits in Duolong are located within areas of the high-average subpopulation of $b_3$ and the low-average subpopulation of $b_1$ (Figs. 6, 8, and 10). Among these three subpopulations, the high-average subpopulation of $b_3$ (Fig. 10) is more widely distributed than the high-average subpopulation of $b_2$ (Fig. 8) and the low-average subpopulation of $b_1$ (Fig. 6). One of the reasons for this issue is that Au-bearing species can be mobilized and dispersed far from the porphyry and epithermal Cu–Au deposits because of their high mobility within a wide range of temperature and redox conditions (Trigub et al., 2017; Williams-Jones et al., 2009; Zhu et al., 2011). The low-average subpopulation of $b_1$ may also correspond to the secondary enrichment of Au after the porphyry and epithermal Cu–Au mineralization. It can be

<table>
<thead>
<tr>
<th>Variables</th>
<th>Subpopulations</th>
<th>Weight</th>
<th>Average</th>
<th>Standard variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln \frac{Cu}{Cr}$</td>
<td>1</td>
<td>0.63</td>
<td>1.72</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.37</td>
<td>1.49</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.05</td>
<td>1.62</td>
<td>0.58</td>
</tr>
<tr>
<td>$\frac{1}{2} \ln \frac{Cr}{Au}$</td>
<td>1</td>
<td>0.25</td>
<td>−0.09</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.70</td>
<td>−0.62</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.05</td>
<td>−3.00</td>
<td>0.58</td>
</tr>
<tr>
<td>$\frac{1}{2} \ln \frac{Au}{Ni}$</td>
<td>1</td>
<td>0.32</td>
<td>−1.43</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.52</td>
<td>−2.01</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.16</td>
<td>−2.13</td>
<td>2.49</td>
</tr>
</tbody>
</table>
explained by findings of a few placer gold deposits far away from the ten porphyry and epithermal Cu-Au deposits. Therefore, the high-average subpopulation of $b_2$ and the low-average subpopulation of $b_1$ can be pathfinders for the Cu-Au mineralization in Duolong.

The high-average subpopulation of $b_2$ concentrated around the discovered Cu-Au deposits is indicative of potential Cu-Au mineralization in Duolong, and two target areas were identified with high prospecting potential (Fig. 8). The first is located in southeastern Duolong, and the second is in the northwestern corner. The first prospecting zone where an outcrop of diorite was recently identified was also predicted to have a high potential for porphyry Cu-Au deposits by applying fractal analysis (Wang et al., 2017). Given the close relationship between the porphyry deposits and intermediate-felsic intrusions (cf. Sillitoe, 2010), this prospecting zone was demonstrated to be a priority for further studies to evaluate its mineralization potential. The second prospecting zone is covered by Quaternary sediment and shows a sparser pattern of the high-average subpopulation of $b_2$ than the first one. Whether this pattern is related to buried Cu-Au deposits still needs more geological and geophysical constraints to determine.

4. Conclusions

The SBP-based ILR transformation and an algorithm for fitting mixture models were used to extract geo-information from stream sediment data in the Duolong mineral district, northern Tibet, China. ILR-transformed variables were constructed from Cu, Au, Cr, and Ni concentrations based on hierarchical cluster analysis and two geological criteria. The first criterion was that magmatic activities with high Cr and Ni concentrations occurred earlier than porphyry and epithermal Cu-Au deposits with low Cr and Ni concentrations. The second criterion is that porphyry Cu-Au deposits were characterized by stable Cu/Au ratios. The ILR-transformed variables follow a bi-normal or tri-normal distribution that records geo-information associated with magmatic activities, porphyry and epithermal Cu-Au deposits, and later geological processes in Duolong. Two target areas were consequently identified and predicted to have Cu and Au potentials. This SBP-based ILR transformation driven by the above geological criteria can extract more detailed geo-information from geochemical data, which is valuable for understanding the geological processes and prospecting for mineral resources. Therefore, this study demonstrates a modeling process that uses geological criteria to obtain interpretable results following log-ratio transformation and presents an example for further work using compositional data analysis in mineral exploration.

Declaration of competing interest

I, Wenlei Wang, declared that I do not have any financial and personal relationships with other people or organizations that could inappropriately influence (bias) their work or state.

Acknowledgements

This work was financially supported by grants from the National Natural Science Foundation of China (grant numbers 41872206, 41772353, 41602088), a China Geological Survey project (grant number DD20179142), and a Basic Research Fund for Central Research Institutes (grant number JYYWF20180602). The authors thank Mario A. T. Figueiredo and Anil K. Jain, for sharing the code of the MML-EM algorithm (http://www.hxx.it.pl/~mtf/). The clustering analysis was run using the R packages ‘composition’ and ‘flashClust’. Two anonymous reviewers and Renguang Zuo are appreciated for helpful comments that improved the original manuscript significantly.