TECHNICAL NOTE

GEOFCM: a new method for statistical classification of geochemical data using spatial contextual information

Kenta YOSHIDA^{*}, Tatsu KUWATANI^{*,**}, Atsushi YASUMOTO^{***},
Satoru HARAGUCHI^{*}, Kenta UEKI^{*} and Hikaru IWAMORI^{*,†,‡}

^{*}Japan Agency for Marine–Earth Science and Technology (JAMSTEC), Yokosuka 237–0061, Japan

^{**}PRESTO, Japan Science and Technology Agency (JST), Kawaguchi 332–0012, Japan

^{***}Department of Geology and Mineralogy, Kyoto University, Kyoto 606–8502, Japan

[†]Earthquake Research Institute, The University of Tokyo, Tokyo 113–0032, Japan

[‡]Department of Earth and Planetary Sciences, Tokyo Institute of Technology, Tokyo 152–8550, Japan

Conventional clustering algorithms such as *k*-means and fuzzy c-means (FCM) cluster analysis do not fully utilize the spatial distribution information of geologic samples. In this paper, we propose GEOFCM, a new clustering method for geochemical datasets with location coordinates. A spatial FCM algorithm originally constructed for image segmentation was modified for application to a sparse and unequally-spaced dataset. The proposed algorithm evaluates the membership function of each sample using neighboring samples as a weighting function. To test the proposed algorithm, a synthetic dataset was analyzed by several hyper-parameter settings. Applying this algorithm to a geochemical dataset of granitoids in the Ina–Mikawa district of the Ryoke belt shows that samples collected from the same geological unit are likely to be classified as the same cluster. Moreover, overlapping geochemical trends are classified consistently with spatial distribution, and the result is more robust against noise addition than standard FCM analysis. The proposed method is a powerful tool to use with geological datasets with location coordinates, which are becoming increasingly available, and can help find overviews of complicated multidimensional data structure.

Keywords: Multivariate statistical analysis, Cluster analysis, Geocoding, Ryoke granitoid

INTRODUCTION

Geochemical data from geological samples show compositional trends that strongly reflect the material differentiation and assimilation that take place during geological processes. As a result, these trends often comprise groups in a multidimensional compositional space and are distributed in real space as geological units ranging from millimeters to kilometers in scale. Spatial contextual information and chemical affinities could provide fundamental information about the sources and generation processes associated with the samples.

Multivariate statistical analysis can be a powerful tool for detecting the characteristic features of multidimensional geochemical datasets, and it is employed in areas of study including environmental sciences, petro-

gy, and resource exploration (e.g., Templ et al., 2008; Kuwatani et al., 2014; Ofner et al., 2015; Yasukawa et al., 2016; Ueki and Iwamori, 2017; Kuwatani et al., 2018; Ueki et al., 2018; Yoshida et al., 2018). Geochemical data including chemical composition, petrographic description, and location coordinates are increasingly provided in the literature, making them available for analysis approaches that deal with large-scale, multidimensional data (Igarashi et al., 2016). Some recent studies shed light on the geocoding, i.e., numerical conversion of the location coordinates of geologic samples that are often provided as sampling maps. Most geologic samples have their sample location as potential information that is generally provided as two-dimensional longitude/latitude-coordinates. Yoshida et al. (2018) performed geocoding of the sampling map from the literature and statistical analyses on the corresponding geochemical data, comparing the clustering results from the geochemical data and its areal distributions. Recently, Haraguchi et al. (2017) performed

geocoding of published geochemical data since the 1980s and constructed a large location-based geochemical database called 'DODAI'. Two international projects have tried to integrate published geochemical data with location coordinates: GEOROC maintained by the Max Planck Institute (<http://georoc.mpch-mainz.gwdg.de/georoc/>) and PetDB maintained by EarthChem, which is part of the Interdisciplinary Earth Data Alliance (IEDA) (<http://www.earthchem.org/petdb>). The increasing availability of location-based geochemical databases may thus offer a chance to fully utilize location coordinates, in addition to the well-established geochemical/petrological works, for data analyses; however, appropriate multivariate statistical analyses should be required to correctly identify the structure of the data from these large sets.

Cluster analysis, one of the fundamental techniques of unsupervised machine learning, classifies a dataset into a certain number of groups based on its characteristics (e.g., Hanesch et al., 2001; Templ et al., 2008). The clusters extracted are commonly either discrete groups or sections of continuous data trends. This simple analytical technique was employed in microtextural analysis combined with electron probe microanalyser (Lanari et al., 2014; Yasumoto et al., in press), characterization of mantle-derived materials (Iwamori et al., 2017), environmental study (Hanesch et al., 2001), and identifying geochemical evolution during progressive metamorphism (Yoshida et al., 2018).

K-means cluster analysis (KCM), one of the most classical and well-known cluster analyses, divides a dataset based simply on the distance between each data point in a multidimensional space (MacQueen, 1967). Independent component analysis (ICA), in contrast, describes data using a set of base vectors that maximize non-Gaussianity and detects independent data trends (Comon, 1994). Recently, Iwamori et al. (2017) proposed a new analytical scheme using KCM, ICA, and principal component analysis, suggesting that KCM combined with a decorrelation technique, known as data whitening, would provide a classification criterion comparable to ICA. They showed that KCM with whitening is an easy and convenient way to perform data mining in multivariate geochemical datasets, applying it to mantle derived materials such as mid-ocean ridge basalt + arc basalt, ocean island basalt + continental basalt, and enriched basalts.

KCM is generally appropriate for the isotropic and distinct clusters (Nagy, 1968). However, geologic samples often show continuous data distributions derived from geologic process such as crystal fractionations (Fig. 1). This may result in variable clusters, particularly in terms of points located between two or more clusters, which would be derived from data uncertainties.

Figure 1. Schematic of the spatially separated, but geochemically overlapped data shown in chemical data space. Filled and open circles indicate data from different geological units. If the geochemical trend show overlaps, it is difficult to classify the data into two geologic units based only on geochemistry.

One possible solution to this problem is a spatial clustering method that coordinates data points laterally as the weighting function of the cluster assignment. This method was developed in medical imaging studies that segmented considerably noisy magnetic resonance images (MRI) (e.g., Ahmed et al., 2002; Chuang et al., 2006). Fuzzy c-means (FCM) clustering is a soft clustering technique which allows each data point to have membership values to the related clusters (Bezdek, 1974; 1981), i.e., FCM evaluates the membership value that shows the probability in which a data point will be attributed to a specific cluster. The clustering criteria for FCM is comparable to those for KCM, and both analyses show the similar results when FCM assigns each data point to the cluster with the maximum membership (e.g., Ghosh and Dubey, 2013). The proposed spatial clustering method, named as spatial FCM (sFCM), treated MRI data in which neighboring pixels were highly correlated and had a high probability of being part of the same cluster. Geologic samples with sample localities listed as location coordinates have a similar characteristic: samples in spatially close associations are expected to be attributed to the same cluster.

The previously proposed sFCM (Chuang et al., 2006) is designed to deal with image data composed of equally spaced pixels. However, the distributions of geologic samples are unequally spaced, due mainly to the presence of outcrops. In the present paper, we propose a modified objective function for sFCM that can be applied to unequally spaced data. The modified method is applicable for any geochemical dataset that contains location coordinates. The proposed algorithm was tested its performance by using a synthetic dataset composed of both

two-dimensional variables and location coordinates. We also apply the method to a geochemical dataset of natural samples constructed by Haraguchi et al. (2017), of granitic rocks distributed in the Ina-Mikawa area, central Japan. The proposed method can be calculated in the ‘GEOFCM’ Excel® workbook provided as a supplementary material online (<https://doi.org/10.2465/jmps.171127>) and our website (<http://dsap.jamstec.go.jp>).

CALCULATION ALGORITHM

Review of Fuzzy c-means (FCM) and spatial FCM clustering

FCM clustering uses a partitional optimization algorithm that optimizes the classification of the dataset by assigning data points to each cluster using fuzzy memberships. Assume that a multidimensional dataset $X = (x_1, x_2, \dots, x_N)$ is assigned to a total of c clusters. Each point x_i represents multidimensional data (vectors) such as the chemical compositions of rocks. The FCM algorithm tries to minimize the cost function using iterative optimization calculations, as follows:

$$J = \sum_{k=1}^c \sum_{i=1}^N (u_{ki})^m \|x_i - \mu_k\|^2 \quad (1),$$

where u_{ki} refers the membership of data point x_i in the k th cluster, and μ_k is the geometric center (centroid) of the k th cluster. $\|\cdot\|$ represents a norm metric and $m (>1)$, called a fuzzifier, which controls the fuzziness of the partitioning calculation. As m increases, the differences between the respective clusters become blurred. Conversely, FCM will reduce to KCM when $m = 1$. In this study, $m = 1.5$ is used in the all calculations.

The cost function approaches its minimum when data points close to the centroids are allotted high membership values. During the optimization calculation, the centroids and cluster membership functions are updated using the following expressions, beginning with an initial guess:

$$\mu_k = \frac{\sum_i^N (u_{ki})^m x_i}{\sum_i^N (u_{ki})^m} \quad (2)$$

and

$$u_{ki} = \left[\sum_j^c \left(\frac{\|x_i - \mu_k\|}{\|x_i - \mu_j\|} \right)^{\frac{2}{m-1}} \right]^{-1} \quad (3).$$

Iterative optimization convergence can be detected by comparing the change in the centroids at two successive iterations (Fig. 2a). After the convergence, each data point is assigned to the cluster with the maximum membership.

In the standard FCM algorithm, the probability (membership function) is determined solely on the distances between data points in multidimensional space. To apply this to MRI image segmentation, Chuang et al. (2006) proposed sFCM method to exploit the spatial relationship among the data points. Their algorithm added a calculation of spatial function after the u_{ki} value was determined by Equation (3) (Fig. 2a). The newly introduced spatial function, in which the membership values of neighboring data points are utilized to weight their probability, is formulated as follows:

$$h_{ki} = \sum_{j \in NB(x_i)} u_{kj} \quad (4),$$

where $NB(x_i)$ represents the set of data points neighboring x_i . As Chuang et al. (2006) applied this algorithm to MRI data composed of equally spaced pixels, they formulated $NB(x_i)$ as a 5×5 square window centered on pixel x_i (Fig. 2b). The spatial function of a data point for a cluster grew when the majority of its neighboring points belonged to the same clusters. The membership function is modified using the spatial function as follows:

$$u'_{ki} = \frac{(u_{ki})(h_{ki})^q}{\sum_{l=1}^c (u_{li})(h_{li})^q} \quad (5),$$

where q is introduced to control the relative importance of the membership and spatial functions. For the image analyses, this formula reduces the weighting of a noisy cluster based on the labels of neighboring data points.

The clustering calculation is performed in two steps (Fig. 2a). First, the membership functions u_{ki} are calculated in the same way as in the conventional FCM using multidimensional data. Second, u_{ki} are mapped in the real space and the spatial function is calculated. The next iteration will proceed with the new, modified, membership with the spatial function (u'_{ki}). Note that when $q = 0$, the calculation is identical to the conventional FCM without any location information.

Extension to the unequally spaced dataset: GEOFCM

In general, the distributions of geologic samples are unequally spaced, due mainly to the presence of outcrops. To treat unequally spaced dataset, such as rock samples, we introduced a modified spatial function as follows:

Figure 2. (a) Procedure to calculation of conventional and spatial FCM. C06 refers Chuang et al. (2006). (b) and (c) Schematic of the neighboring window of (b) previously established method applied for equally spaced data such as image pixels (Chuang et al., 2006). In this case, a total of 5×5 pixels surrounding to the target pixel x_i are considered for the weighting function of the cluster membership. (c) For unequally spaced data, such as geologic samples, points within a certain radius r around x_i are considered for the weighting function of the cluster membership.

$$h'_{ki} = \frac{1}{n} \sum_{j \in NB_r(x_i)} u_{kj} \quad (6),$$

where $NB_r(x_i)$ is the set of neighboring data points x_j located within the distance r in the spatial domain and n is the number of data points in $NB_r(x_i)$, inclusive of x_i itself (Fig. 2c). One difference between this and the spatial windows used in the Chuang et al. (2006) formula is that the number of data points neighboring x_i is not always the same. Therefore, the spatial function proposed in this study (Eq. 6) averages the neighboring membership functions (u_{ki}). The amount of neighboring data considered for the spatial function will increase as r increases. Setting a larger r value will approach a result that all

samples would prefer to belong to the cluster with the highest population; therefore, r should be limited to define “neighbors” depending on the spatial density of the samples. Using this formula, the membership function u' is calculated as in Chuang et al. (2006); thus, centroid μ is updated with the calculated u' (Fig. 2a).

Validity of the results obtained was evaluated using the fuzzy partitioning coefficient (Bezdek, 1974), which is one of several validity indices for fuzzy clustering and is calculated with the following expression:

$$V_{pc} = \frac{\sum_k^c \sum_i^N (u_{ki})^2}{N} \quad (7).$$

The partitioning coefficient provides a quantitative meas-

Figure 3. (a) The synthetic data used for the test of GEOFCM algorithm, showing two clusters spatially distinct (left column), but show crossing trends in A-B space (center column). IC1 and IC2 in the center column indicate independent component (IC) calculated by ICA. Projection on the IC space is also shown in the right column. ICA was performed with FastICA (Hyvärinen, 1999). (b)–(f) Analyses of the synthetic data under different analytical conditions: (b) $q = 0$, which is equivalent to the standard FCM; (c) and (d) $q = 1.5$ and 5 at constant $r (= 30)$ showing the effect of changing q . (e) and (f) $r = 15$ and 600 at constant $q (= 1.5)$ showing the effect of changing r . For detail, please see the text.

ure of how fuzzy the calculated memberships are; in other words, the best clustering is obtained when V_{pc} is maximized such that samples are attributed to respective single clusters.

RESULTS

To testify the proposed algorithm, GEOFCM, we applied it to a synthetic dataset containing 400 data points with location coordinates. The synthetic dataset used here is shown in Figure 3a, with location coordinates ranging ~ 132 – 140 for X and ~ 27 – 39 for Y , modeling longitude and latitude, respectively. Two distinct clusters are prepared, both of which contain 200 data points. Model geochemical data are composed of two dimensions, A and B. The data of the cluster 1 show spreading values of A with almost constant B, whereas those of the cluster 2 show negative linear correlation between A and B. Two trends overlap at around $A = 43$ and $B = 50$.

Figure 3b shows the clustering result obtained from standard FCM ($q = 0$), in other words, clustering result only based on A and B. Since model geochemical data of A and B overlap at considerable degree, many points are misidentified.

Figures 3c and 3d show results from GEOFCM using different q values. In the present cases, the radius of neighboring function is set to 30 km since the average distance between nearest points is approximately 15. By introducing spatial contextual information, clustering results are drastically improved. Figure 4 shows number of misidentification and V_{pc} for different q and r . As shown in Figure 4a, the number of misidentification drastically decreases with increasing q . With $q = 1.5$, only ten points of cluster 1 are misidentified as cluster 2.

On the other hand, Figures 3e and 3f show the results using different neighboring radius at constant q ($q = 1.5$). By using $r = 15$ km, that is similar value to the average distance between nearest points, the result is very similar to that without location coordinates. With increase of r , the number of misidentification decrease. With the value larger than 30, the number of misidentification becomes fairly constant. However, at larger r (>400), the number of misidentification increases and V_{pc} decreases again.

DISCUSSION

Affinity of the method

As shown in Figure 3, the inappropriate hyper-parameters significantly lead serious misidentifications of the clusters. Iwamori et al. (2017) indicated that KCM with whitening of the dataset can provide the clustering results

identical to the ICA. According to Iwamori et al. (2017), two crossing trends in a multivariate space can be distinguished by both ICA and KCM, although points around the intersection of the trends should be treated as a ‘mixing cluster.’ In the case of present example, intersection of two trends cannot be distinguished correctly (Fig. 3b). As we set only two clusters for the calculation, Figure 3b shows smaller cluster 1 and larger cluster 2. This classification can be well recognized in the IC space that is shown in the right column of Figure 3. Two clusters obtained from standard FCM ($q = 0$) is determined based on IC1 shown in the right column of Figure 3b. Cluster 1 is recognized as the points with $IC1 < -10.8$ and for cluster 2, vice versa. However, given two trends substantially overlap (Fig. 3a) and therefore, it is difficult to distinguish them in the IC space. By introducing spatial contextual information, this problem is reasonably solved, keeping its consistency to the ICA result. At the intersection of two trends, the classification criteria are modified by the spatial information and the results gain consistency with the spatial distribution of the samples (Figs. 3c and 3d). Figure 4a indicates that increase of q , i.e., the increase of weighting of the spatial information, will lead the clustering result with higher precision and less fuzzy partitioning. However, with larger q values, clustering results become strongly controlled by the spatial distribution of the data point, and geochemical characteristics will be blinded. Therefore, q should be minimal that can modify the classification appropriately. For a q value larger than 1.5, the number of misidentification is less than ten (out of total 400 points) and V_{pc} value larger than 0.95; this indicates that when $q = 1.5$, the result may represent the true characteristic with $>99\%$ of precision.

Comparing Figures 3c and 3e, the radius for the neighboring function (Eq. 6) should be large enough compared to the average distance between nearest points. When r is similar to the average distance between nearest points, the number of points considered for neighboring function may be so small that the modification is too weak. The r value slightly larger than the average distance between nearest points (20 in Fig. 4b) still produces relatively large number of misidentification. Figure 4b indicates that r requires twice or larger value of the average distance between nearest points. Conversely, very large neighboring function also brings worse results (Fig. 3f). Studied synthetic data are distributed over ~ 4 and 6° along vertical (latitude) and horizontal (longitude) directions, respectively. This means each cluster spreads over $\sim 450 \times 570$ km as the actual distance. If the neighboring function exceeds minor axes of the clusters, considerable number of the data points belonging to different clusters may inevitably incorporate to the modification of mem-

Figure 4. Number of misidentification (left axis) and V_{pc} (right axis) for different hyper-parameters. Small number of misidentification indicates precise clustering and large V_{pc} exhibits less fuzzy partitioning. (a) With increasing q and constant r ($= 30$), indicating $q \geq 1.5$ may provide good results. (b) With increasing r and constant q ($= 1.5$), indicating $r \geq 30$ may provide good results.

bership values, inducing incorrect classifications. Therefore, the radius of the neighboring function should be small enough compared to the actual breadth of the individual clusters

Consequently, the present result on the synthetic data set suggests that $q \geq 1.5$ and $r \geq 2 \times$ (average distance between nearest points) will provide fairly good results. However, we need careful evaluation on the calculated results changing with hyper-parameters.

Application to the natural dataset of granitoids

Here an example of GEOFCM is shown, using a geochemical dataset of basement rocks, compiled by Haraguchi et al. (2017). The study area is about 80×80 km and

Figure 5. Distribution of granitic rocks in the area of interest (modified from the Seamless Geologic map, Geological Survey of Japan, 2015). Abbreviations for the granitoid bodies are as follows: T, Tenryukyo granite; Tt, Takato granite; Oc, Ochiai granite; Ks, Kisokoma granodiorite; Ic, Ichida and Seinanji granite; In, Inagawa and Ikuta granite; S, Shinshiro tonalite; Km, Kamihara-Hiji tonalite; Bs, Busetsu and Kadoshima granite; Ky, Kiyosaki granodiorite; and N, Naegi-Agematsu granite. The position of the median tectonic line is also shown.

contains the Ryoke granitoids occurring with the Ryoke metamorphic rocks (Fig. 5). The Ryoke granitoids are generally classified into the Older and Younger groups: the former consists of syn- to late-tectonic plutons occurring in high-grade metamorphic zones while the latter consists of post-tectonic plutons that cross-cut the structures of associated metamorphic rocks and Older Ryoke granitoids (e.g., Ryoke Research Group, 1972). Th-U-total Pb isochron dating of monazite shows that the Older and Younger Ryoke granitoids were emplaced during 100–85 and 85–75 Ma, respectively (Suzuki et al., 1994; Kutsukake, 2002). Recently, Takatsuka et al. (2018) measured the U-Pb zircon ages of granitoids collected from the Kamihara tonalite body and Tenryukyo granite in the southern part of the studied area and showed that the southwestern part of the Tenryukyo granite is younger than others (~ 81 –75 Ma) and has a different origin, suggesting that age data is a powerful tool to classify the plutons in this area, although the dataset used here does not contain the ‘younger part’ of the Tenryukyo body. There are a total of 11 recognized granitoid bodies in the study area (Fig. 5): the Older Ryoke granitoid bodies (Kami-

Figure 6. The results of GEOFCM as applied to a geochemical dataset of the studied area, with the calculation settings (a) $q = 0$, equivalent to the standard FCM, and (b) $q = 1.5$. Clustering results also applied for the noise-added data with calculation setting of (c), when $q = 0$, equivalent to the standard FCM, and (d) $q = 2$.

hara, Tenryukyo, and Kiyosaki) and the Younger Ryoke granitoid bodies (Inagawa, Shinshiro, Busetsu, Naegi-Agematsu, Ichida, Kisokoma, Takato, and Ochiai) which include some small bodies not shown in Figure 5. However, samples collected from the Kiyosaki granitoid body are missing from the dataset.

Among the dataset compiled by Haraguchi et al. (2017), about 14 articles and 226 sample datasets are related to the area of interest (Yuhara and Kagami, 1995; Kutsukake, 1997; Yuhara and Kagami, 1999; Ishihara and Wu, 2001; Yuhara et al., 2004a, 2004b; Ishihara, 2005; Yuhara and Kagami, 2006; Ishihara and Murakami, 2006; Yuhara and Kagami, 2007; Ishihara and Chappell, 2007; Yuhara and Kagami 2008; Yuhara, 2011; Yuhara and Kagami, 2012). Treatment of compositional data generally faces the constant-sum constraint, although this can be avoided using logratio transformation (e.g., Aitchison, 1986). However, this study focused on trace elements that are relatively free from the constant-sum constraint and chose Rb, Sr, Y, Zr, Nb, and Ba to test the applicability of the proposed method.

Before performing the GEOFCM calculation, we performed decorrelation of the dataset to avoid the scaling problem of the dataset, i.e., standardization and whitening of the dataset following the procedure in Iwamori et al.

(2017).

Figure 6a shows the clustering result obtained using a standard FCM ($q = 0$), which is equivalent to the result of KCM with whitening. Figure 6b shows the result using $q = 1.5$. The average distance between the nearest neighbors was approximately 1.1 km, and the $NBr(x_i)$ was calculated using $r = 5$ km for all calculations so that several samples were included in the neighboring radius of every other sample. The number of clusters was set to 10 for all calculations as previous studies reported about 10 different granitoid bodies in the studied dataset, although some granitoid bodies are missing from the analyzed data set.

Figure 6a shows the result of standard FCM which is roughly consistent with the previously reported classification, although there are elaborate distributions of some clusters. Cluster 10 is present in both the Kamihara body (in the southwestern part of the study area, Older Ryoke granitoids) and the Ichida body (in the northeastern part of the study area, Younger Ryoke granitoids). The Kamihara body is subdivided into four clusters (cluster 2, cluster 3, cluster 8, and cluster 10) within ~ 20 km. The Inagawa body shows the complex intricacies of cluster 2, cluster 3, cluster 4, and cluster 5, while at the eastern part, cluster 8 also occurs within the Inagawa body. The Kisokoma body is mostly occupied by cluster 9. The classification

Figure 7. Clustering results of Sr plotted against Rb. The calculation settings are (a) $q = 0$, which is equivalent to the standard FCM, and (b) $q = 1.5$.

of Takato body is very complex, including clusters 1, 3, 6, and 7. Overall, many points are distributed irregularly compared to their neighbors.

The effect of the spatial function is clearly seen in Figure 6b, as the membership function is modified according to the memberships of neighboring points. In the Inagawa body, the complex intricacies decreased as the spatial weighting (the q value) increased. The GEOFCM algorithm, with its higher q value, had a stronger smoothing effect. This can ignore specific characteristics distributed within a small distance; therefore, evaluation with different spatial weightings is required to precisely determine the geochemical structure of the area of interest.

Figure 7 shows the plot of Sr versus Rb with the results calculated from GEOFCM. The relationship between the two elements in the study area is expressed mainly in two major trends; trend 1, which showed a steep decrease in Sr as Rb increased (approximately Rb <150 ppm) and trend 2, which showed a moderate decrease in Sr as Rb increased (approximately Rb >150 ppm). The clusters corresponding to trend 2 is observed mainly in the Naegi body and the northwestern part of the Inagawa body (cluster 4 and cluster 5 in Fig. 6), which was classified as cluster 5 in the GEOFCM results but partially assigned to another cluster (cluster 4) by the standard FCM ($q = 0$). The lowest Rb content of cluster 5 in Figure 7b is approximately Rb = 150 ppm, whereas cluster 4 in

Figure 7a lay on the range of Rb = 100–250 ppm. This could indicate that trend 1 and trend 2 overlap in a certain part of the Inagawa body and the standard FCM, that is calculated only based on geochemical data, cannot distinguish the boundary of trend 1 and trend 2. In the case of KCM, Iwamori et al. (2017) also shows a similar example that the intersection of two (or more) different trends cannot be distinguished and is classified as a mixture cluster. However, the GEOFCM algorithm revealed the boundary between the two overlapping trends based on the combination of geochemical characteristics and spatial information. Kutsukake (2002) reported a trend similar to the trend 1 in samples from the Tenryukyo body and indicated that it could be explained well by the fractionation of plagioclase, as Sr incorporates Ca-site of plagioclase. Ishihara and Wu (2001) reported that the Naegi body contained very high quantities of Rb and indicated that it was more fractionated than other granitoid bodies in the south. These two different trends are successfully distinguished by using spatial contextual information.

CONCLUSION

Many clustering algorithms have been proposed to utilize spatial data. The present results propose the applicability of a simple clustering algorithm, which combines an FCM algorithm with a spatial weighting function, to geo-

chemical datasets with location coordinates. This method effectively modified the clustering results obtained from geochemical data and helped evaluate the geochemical structure related to spatial information. The increasing availability of geochemical datasets with location coordinates, from sources including compilation of rock composition databases, scientific ocean floor drilling, and resource mining, provides a good opportunity to study the spatially distributed affinities and generation history of geologic samples. Since the results of the GEOFCM algorithm are weighted by their spatial distribution, several types of noise found in geochemical datasets, such as the inhomogeneity of hand specimens, analytical uncertainty, employed decimal points, improvement of instruments over ages, and so on, can be suppressed. Although hyper-parameter tuning, i.e., choosing the appropriate m , r , and q in Equations (1), (5), and (6) is required to obtain feasible classification results, the results of the calculations may indicate new knowledge having consistency with geological maps constructed in previous studies.

As suggested by Iwamori et al. (2017), combining cluster analysis with either standardization and/or whitening of the dataset may provide an overview of the complicated structure of multidimensional datasets. To deal with overlapping trends, using spatial contextual information can be a key to classify geochemical data correctly. The method presented in this paper may help geologists use statistical analysis to mine geological datasets for hidden data structures, based on current understanding of geology.

ACKNOWLEDGMENTS

The authors thank Shun-suke Horiuchi and Mika Mohamed Abdelbaky Seif El-Nasr at JAMSTEC for their help and discussion on treating the data. This work was partly supported by JSPS KAKENHI Grants Nos. JP16K17835 for KY, No. 4503 (No. JP25120005 for TK) and JP15K20864 for TK, and JP17H02063 for KU. Financial support was also provided by JST PRESTO (Grant No. JPMJPR1676) for TK. The authors also thank members of the Joint Usage/ Research Center Program Nos. 2015-B-04 and 2018-B-01 from the Earthquake Research Institute, The University of Tokyo for fruitful discussions. This paper also benefited from constructive reviews by T. Ikeda and an anonymous reviewer, and editorial handling by T. Kato. We would like to thank Editage (www.editage.jp) for English language editing.

SUPPLEMENTARY MATERIAL

The proposed method can be calculated in the Excel® workbook ‘GEOFCM’ that is available online from

<https://doi.org/10.2465/jmps.171127>. Note that the calculation will work on Visual Basic for Application macro and the users are required to enable macros. If bug-fixes and/or further improvements will be added in the future, the latest version of the calculation sheet will be provided at our website (<http://dsap.jamstec.go.jp>).

REFERENCES

- Ahmed, M.N., Yamany, S.M., Mohamed, N., Farag, A.A. and Moriarty, T. (2002) A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 21, 193–199.
- Aitchison, J. (1986) *The statistical analysis of compositional data*. pp. 416, Chapman and Hall, London.
- Bezdek, J.C. (1974) Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, 7, 57–71.
- Bezdek, J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.
- Chuang, K.-S., Tzeng, H.-L., Chen, S., Wu, J. and Chen, T.-J. (2006) Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*, 30, 9–15.
- Comon, P. (1994) Independent component analysis, A new concept? *Signal Processing*, 36, 287–314.
- Geological Survey of Japan, AIST (2015) Seamless digital geological map of Japan 1: 200,000. May 29, 2015 version. Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology.
- Ghosh, S. and Dubey, S.K. (2013) Comparative analysis of K-Means and Fuzzy C-Means algorithms. *International Journal of Advanced Computer Science and Applications*, 4, 35–39.
- Hanesch, M., Scholger, R. and Dekkers, M.J. (2001) The application of fuzzy c-means cluster analysis and non-linear mapping to a soil data set for the detection of polluted sites. *Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy*, 26, 885–891.
- Haraguchi, S., Ueki, K., Yoshida, K., Kuwatani, T. and Iwamori, H. (2017) Geochemical database of Japanese islands for data-driven science: problem and solution of published domestic data. Abstract of JpGU-AGU 2017, MG129-P01.
- Hyvärinen, A. (1999) Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Transactions on Neural Networks*, 10, 626–634.
- Igarashi, Y., Nagata, K., Kuwatani, T., Omori, T., Nakanishi-Ohno, Y. and Okada, M. (2016) Three levels of data-driven science. *Journal of Physics: Conference Series*, 699, 012001.
- Ishihara, S. (2005) Geology of the Hiraiwa fluorite deposits and their relevance to the fluorine metallogenic province in Japan. *Bulletin of the Geological Survey of Japan*, 56, 167–176.
- Ishihara, S. and Wu, C. (2001) Genesis of Late Cretaceous-Paleogene Granitoids with Contrasting Chemical Trends in the Chubu District, Central Japan. *Bulletin of the Geological Survey of Japan*, 52, 471–491.
- Ishihara, S. and Murakami, H. (2006) Characteristics of REE distribution in granitoids of SW Japan: Miocene plutonic rocks at Ashizuri-misaki and late Cretaceous granitoids of the Sanyo Belt of SW Japan. *Bulletin of the Geological Survey of Japan*, 57, 89–103.
- Ishihara, S. and Chappell, B.W. (2007) Chemical compositions of the late Cretaceous Ryoke granitoids of the Chubu District,

- central Japan-Revisited. *Bulletin of the Geological Survey of Japan*, 58(9/10), 323–350.
- Iwamori, H., Yoshida, K., Nakamura, H., Kuwatani, T., Hamada, M., Haraguchi, S. and Ueki, K. (2017) Classification of geochemical data based on multivariate statistical analyses: Complementary roles of cluster, principal component, and independent component analyses. *Geochemistry, Geophysics, Geosystems*, doi: 10.1002/2016GC006663.
- Kutsukake, T. (1997) Petrology and geochemistry of a calcic and ferrous granitoid pluton: the Mitsuhashi Granite in the Ryoke Belt, southwest Japan. *Journal of Mineralogy, Petrology and Economic Geology*, 92, 231–244.
- Kutsukake, T. (2002) Geochemical characteristics and variations of the Ryoke Granitoids, Southwest Japan: Petrogenetic implications for the plutonic rocks of a magmatic arc. *Gondwana Research*, 5, 355–372.
- Kuwatani, T., Nagata, K., Okada, M., Watanabe, T., Ogawa, Y., Komai, T. and Tsuchiya, N. (2014) Machine-learning techniques for geochemical discrimination of 2011 Tohoku tsunami deposits. *Scientific Report*, 4, 7077, 1–6.
- Kuwatani, T., Nagata, K., Yoshida, K., Okada, M. and Toriumi, M. (2018) Bayesian probabilistic reconstruction of metamorphic *P-T* paths using inclusion geothermobarometry. *Journal of Mineralogical and Petrological Sciences*, 113, 82–95.
- Lanari, P., Vidal, O., De Andrade, V., Dubacq, B., Lewin, E., Grosch, E.G. and Schwartz, S. (2014) XMapTools: A MATLAB®-based program for electron microprobe X-ray image processing and geothermobarometry. *Computers & Geosciences*, 62, 227–240.
- MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.
- Nagy, G. (1968) State of the art in pattern recognition. *Proceedings of the IEEE*, 56, 836–863.
- Ofner, J., Kamilli, K.A., Eitenberger, E., Friedbacher, G., Lendl, B., Held, A. and Lohninger, H. (2015) Chemometric analysis of multisensory hyperspectral images of precipitated atmospheric particulate matter. *Analytical Chemistry*, 87, 9413–9420.
- Ryoke Research Group (1972) The mutual relations of the granitic rocks of the Ryoke metamorphic belt in central Japan. *Earth Science (Chikyu Kagaku)*, 26, 205–216.
- Suzuki, K., Morishita, T., Kajizuka, I., Nakai, Y., Adachi, M. and Shibata, K. (1994) CHIME ages of monazites from the Ryoke metamorphic rocks and some granitoids in the Mikawa-Tono area, Central Japan. *Bulletin of Nagoya University Furukawa Museum*, 10, 17–38 (in Japanese with English abstract).
- Takatsuka, K., Kawakami, T., Skrzypek, E., Sakata, S., Obayashi, H. and Hirata, T. (2018) Age gap between the intrusion of gneissose granitoids and regional high-temperature metamorphism in the Ryoke belt (Mikawa area, central Japan). *Island Arc*, doi: 10.1111/iar.12224.
- Templ, M., Filzmoser, P. and Reimann, C. (2008) Cluster analysis applied to regional geochemical data: Problems and possibilities. *Applied Geochemistry*, 28, 2198–2213.
- Ueki, K. and Iwamori, H. (2017) Geochemical differentiation processes for arc magma of the Sengan volcanic cluster, Northeastern Japan, constrained from principal component analysis. *Lithos*, 290–291, 60–75.
- Ueki, K., Hino, H. and Kuwatani, T. (2018) Geochemical discrimination and characteristics of magmatic tectonic settings; A machine-learning-based approach. *Geochemistry, Geophysics, Geosystems*, 19, <https://doi.org/10.1029/2017GC007401>
- Yasukawa, K., Nakamura, K., Fujinaga, K., Iwamori, H. and Kato, Y. (2016) Tracking the spatiotemporal variations of statistically independent components involving enrichment of rare-earth elements in deep-sea sediments. *Scientific Reports*, 6:29603.
- Yasumoto, A., Yoshida, K., Kuwatani, T., Nakamura, D., Svojtka, M. and Hirajima, T. (2018) A rapid and precise quantitative electron probe chemical mapping technique and its application to ultrahigh-pressure eclogite from the Moldanubian Zone of the Bohemian Massif (Nové Dvory, Czech Republic). *American Mineralogist*, in press.
- Yoshida, K., Kuwatani, T., Hirajima, T., Iwamori, H. and Akaho, S. (2018) Progressive evolution of whole-rock composition during metamorphism revealed by multivariate statistical analyses. *Journal of Metamorphic Geology*, 36, 41–54.
- Yuhara, M. (2011) Trace and Rare Earth Elements Compositions of Metamorphic Rocks from the Ina District of the Ryoke Metamorphic Belt, Southwest Japan Arc. *Fukuoka University Science Reports*, 41, 39–50.
- Yuhara, M. and Kagami, H. (1995) Cooling history of the Katsuma quartz diorite in the Ina district of the Ryoke belt, Southwest Japan Arc. *The Journal of the Geological Society of Japan*, 101, 434–442.
- Yuhara, M. and Kagami, H. (1999) Chronological and isotope geological study of the Takato granite in the Ina district of the Ryoke belt, Southwest Japan Arc. *The Journal of the Geological Society of Japan*, 105, 181–192.
- Yuhara, M., Miyazaki, T. and Kagami, H. (2004a) Geochronological and isotope geological study of the Kise Granodiorite in the Ina district of the Ryoke Belt, Southwest Japan Arc. *Fukuoka University Science Reports*, 34, 51–65.
- Yuhara, M., Miyazaki, T., Takahashi, T. and Kagami, H. (2004b) The change of chemical and Sr, Nd isotopic compositions of Cretaceous granitic rocks during weathering process. *Japanese Magazine of Mineralogical and Petrological Sciences*, 33, 185–196.
- Yuhara, M. and Kagami, H. (2006) Geochronological and Isotope Geological Study of the Kisokoma Granodiorite and Ichida Granite in the Ina District of the Ryoke Metamorphic Belt, Southwest Japan Arc. *Fukuoka University Science Reports*, 36, 37–61.
- Yuhara, M. and Kagami, H. (2007) Geochronological and Isotope Geological Study of Mafic Igneous Rocks in the Ina District of the Ryoke Metamorphic Belt, Southwest Japan Arc. *Fukuoka University Science Reports*, 37, 57–78.
- Yuhara, M. and Kagami, H. (2008) Geochronological and Isotope Geological Study of the Suisyozan Mafic Mass in the Miho Area of the Ryoke Metamorphic Belt, Southwest Japan Arc. *Fukuoka University Science Reports*, 38, 75–88.
- Yuhara, M. and Kagami, H. (2012) Geochronological and Rb-Sr and Sm-Nd isotopic study of mafic rocks in the Ryoke Metamorphic Belt of the Mikawa District, Southwest Japan Arc. *Fukuoka University Science Reports*, 42, 37–55.

Manuscript received November 27, 2017

Manuscript accepted April 6, 2018

Published online June 7, 2018

Manuscript handled by Takenori Kato