Earth Science Informatics (2022) 15:73–87
https://doi.org/10.1007/s12145-021-00694-3

RESEARCH ARTICLE

Characterization of hydrothermal alteration along geothermal wells using unsupervised machine-learning analysis of X-ray powder diffraction data

Kazuya Ishitsuka[1] · Hiroki Ojima[2] · Toru Mogi[3] · Tatsuya Kajiwara[4] · Takeshi Sugimoto[4] · Hiroshi Asanuma[5]

Received: 3 December 2020 / Accepted: 27 August 2021 / Published online: 1 October 2021

Abstract
Zonal distribution of hydrothermal alteration in and around geothermal fields is important for understanding the hydrothermal environment. In this study, we assessed the performance of three unsupervised classification algorithms—K-mean clustering, the Gaussian mixture model, and agglomerative clustering—in automated categorization of alteration minerals along wells. As quantitative data for classification, we focused on the quartz indices of alteration minerals obtained from rock cuttings, which were calculated from X-ray powder diffraction measurements. The classification algorithms were first examined by applying synthetic data and then applied to data on rock cuttings obtained from two wells in the Hachimantai geothermal field in Japan. Of the three algorithms, our results showed that the Gaussian mixture model provides classes that are reliable and relatively easy to interpret. Furthermore, an integrated interpretation of different classification results provided more detailed features buried within the quartz indices. Application to the Hachimantai geothermal field data showed that lithological boundaries underpin the data and revealed the lateral connection between wells. The method's performance is underscored by its ability to interpret multi-component data related to quartz indices.

Keywords Hydrothermal alteration · X-ray powder diffraction · Quartz index · Machine learning · Unsupervised classification

Introduction

Mapping the zonal distribution of hydrothermally altered minerals in and around a geothermal field is important for understanding the geothermal system and selecting a promising site. The distribution of alteration minerals can be linked to physical and chemical conditions, such as temperature and fluid acidity, which predominate in geothermal systems (e.g., Browne 1978; Reyes 1990; Yang et al. 2001; Lutz et al. 2011). The zonal distribution of alteration minerals can also be used to support interpretation of geophysical data, such as resistivity (e.g., Yoneda 2014), as petrophysical and mechanical properties are affected by secondary mineralization attributed to alteration and by host rock properties (e.g., Frolova et al. 2010, 2015; Wyering et al. 2014; Mielke et al. 2015; Delayre et al. 2020).

Machine learning is a suitable approach for automated zonation and characterization of multi-dimensional data, with several examples of its successful application to geophysical logs, such as seismic velocity, resistivity, gamma ray, and average neutron density porosity (e.g., Raeesi et al. 2012; Grana et al. 2017; Caté et al. 2017; He

Communicated by: H. Babaie

Kazuya Ishitsuka previous affiliation was Division of Sustainable Resources Engineering Hokkaido University Kita-ku, Sapporo, Japan

Kazuya Ishitsuka
ishitsuka.kazuya.4w@kyoto-u.ac.jp

1 Department of Urban Management, Kyoto University, Nishikyo-ku, Kyoto, Japan

2 Cooperative Program for Resources Engineering, Hokkaido University, Kita-ku, Sapporo, Japan

3 Volcanic Fluid Research Center, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan

4 Geothermal Engineering Co., Ltd., Takizawa, Iwate, Japan

5 Fukishima Renewable Energy Institute, National Institute of Advanced Industrial Science and Technology, Koriyama, Fukushima, Japan

et al. 2020; Feng 2020). Machine-learning classification of geophysical data can help with objectively classifying the subsurface's physical properties and has also been used to identify natural-resource reservoirs (e.g., oil/gas). However, the application of machine-learning classification to alteration minerals has been limited, with only a few studies that have focused on characterizing alteration minerals along wells via machine learning (e.g., Caté et al. 2018; Hood et al. 2018; Chen et al. 2018). Application of the machine-learning classification to characterize alteration minerals along wells would also be beneficial in terms of understanding the zonation of hydrothermal alteration in depth as well as comparing the classifications of geophysical logs.

X-ray powder diffraction (XRD) is a versatile technique based on the basal spacing of mineral crystals that facilitates the identification of mineral species. It has been used to document the species of alteration minerals in geothermal fields (e.g., Schiffman and Fridleifsson 1991; Fulignati et al. 1997; Inoue et al. 2004). One of the quantities obtained from XRD is the quartz index (QI), which is a measure of the ratio of the peak intensity of a mineral's XRD value to that of pure quartz minerals (Hayashi 1979; Takahashi et al. 2007). As XRD uses powder material, QI measurement can be performed based on rock cuttings. Cuttings are generally available along wells, and thus, the QI can be measured sequentially along any well. Additionally, compared to earlier studies on geochemical classification using machine learning with elemental composition data (e.g., Caté et al. 2018; Ueki et al. 2018), the QI indicates the mineral type, which may be useful in interpreting the environment in which the mineral was formed.

In this study, we used unsupervised classification methods to characterize hydrothermal alteration minerals based on the QI and temperature (Fig. 1). We used unsupervised classification as it does not require training data to define each class and therefore can be applied to fields for which the amount of training data is insufficient. These approaches facilitate automatic detection of a group of minerals that had experienced similar hydrothermal alteration processes, because each hydrothermal environment contains a conformable set of altered minerals. In this study, we tested three unsupervised classification algorithms: K-means clustering (MacQueen 1967), the Gaussian mixture model (GMM) (McLachlan and Peel 2000), and agglomerative clustering (AC) (Gower and Ross 1969; Lior and Maimon 2005). These unsupervised algorithms have previously demonstrated their effectiveness in classifying geophysical and geochemical data (e.g., Templ et al. 2008; Grana et al. 2017; Saporetti et al. 2018) but have not been evaluated for the classification of hydrothermal alterations along wells. We further proposed using a decision tree (Breiman 2001) to delineate each classified group's characteristics (Fig. 1). The application site was the Hachimantai geothermal field, located in Iwate Prefecture, northeast Japan (Fig. 2a) (NEDO 2007, 2008).

Local geology and data used in this study

Local geology

The geology of the Hachimantai geothermal field comprises a Quaternary formation that overlies a Tertiary formation (Kimbara 1985). The Tertiary formation consists of dacitic and andesitic tuff, and the Quaternary formation is approximately 450 m thick (Fig. 2b) (Kimbara 1985). The tonalite, which is encountered at depths of approximately 1500–1800 m, intrudes beneath the Tertiary formation (Fig. 2b) and is considered a heat source for the geothermal field. The metamorphism of the Tertiary wall rocks around the tonalite includes extensive development of biotite, cordierite, talc, magnetite, ilmenite, and other metamorphic minerals. The QI values in this study were obtained from cuttings along two neighboring wells called N19-HA-1 and N19-HA-2 that were drilled down to approximately 1750 and 1600 m, respectively. The QI values were reported in NEDO (2007, 2008). The ground facilities of the two wells are at almost the same position, but N19-HA-1 was drilled in an NW direction, whereas N19-HA-2 was drilled toward the NNW (NEDO 2007, 2008). Evidence of high-temperature fluid circulation can be found in the zonation of altered minerals around wells N19-HA-1 and

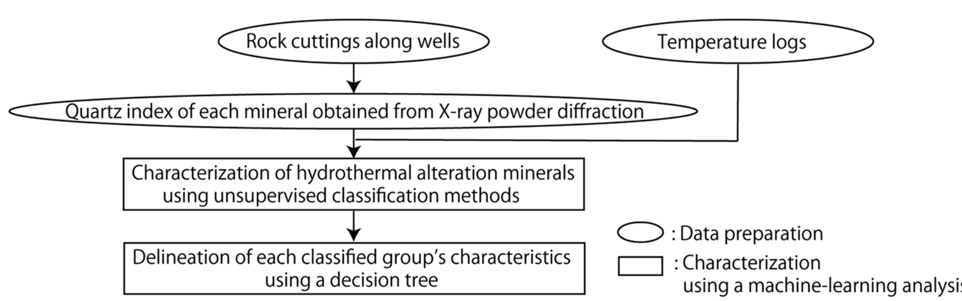


Fig. 1 The work flow used in this study

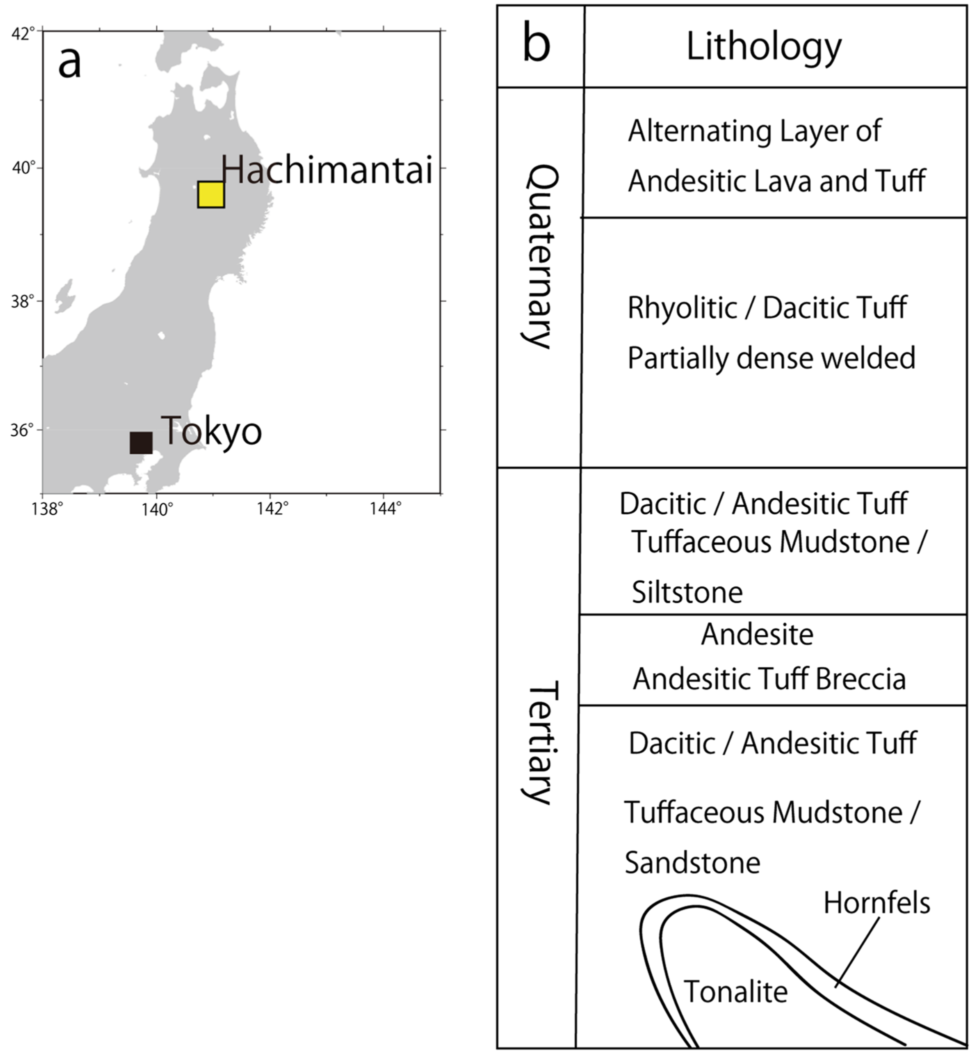


Fig. 2 **a** Location of the Hachimantai geothermal field. **b** Lithological characteristics of the Hachimantai geothermal field (based on NEDO 2008)

N19-HA-2. These can be sub-divided into the Silica-Rutile, Silica-Anatase, and Tridymite-Cristobalite zones (around well N19-HA-1). Temperature logs mainly show patterns of conduction, with maximum temperatures of approximately 300 °C at the bottoms of the wells (NEDO 2007, 2008).

X-ray powder diffraction data and the quartz index

The rock cuttings (e.g., Fig. 3) were first crushed manually using a pestle and then disaggregated using a planetary mill (NEDO 2007). XRD analyses were subsequently performed on basally and randomly oriented samples using a diffractometer with CuK α_1 radiation at 40 kV and 25 mA. The randomly oriented samples were scanned between 2–62° 2θ at a scan speed of 2° $2\theta/\mathrm{min}$. Minerals were identified based on the International Centre for Diffraction Data (ICDD) PDF-2 database.

The QI of a mineral is defined as follows (Hayashi 1979):

$$\mathrm{QI} = \frac{I_m}{I_q} \times 100 \tag{1}$$

where I_m is the peak intensity of the XRD value of a mineral and I_q is the peak intensity of the XRD value of pure silica. The peak intensity of XRD is influenced not only by the amount of minerals but also by the crystallinity and preferential growth of the crystal structure. We thus used the normalized values (see Sect. 3) for machine-learning classification and did not take the absolute QI value into account.

Table 1 presents the 24 mineral types for which QI values were calculated: four clay minerals (smectite, chlorite, sericite, kaolinite), two zeolite minerals (laumontite, wairakite), two silica minerals (tridymite, cristobalite), seven silicate minerals (clinopyroxene, epidote, prehnite, anthophyllite, biotite, cordierite, talc), five oxide minerals (magnetite, ilmenite, hematite, anatase, rutile), one sulfide mineral (marcasite), two sulfate minerals (anhydrite, alunite), and one carbonate mineral (calcite)

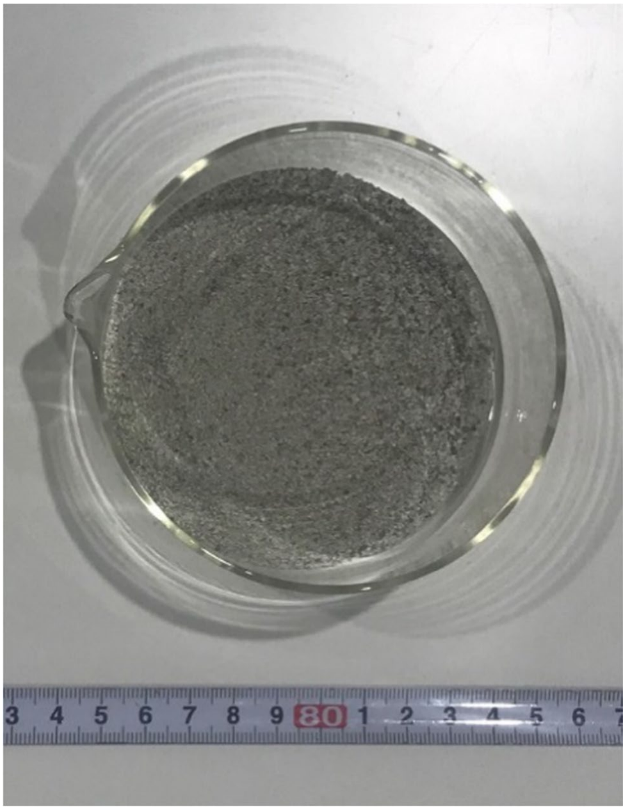


Fig. 3 Example of rock cuttings obtained from the N19-HA-1 well at a depth of 1165 m

(NEDO 2007, 2008). These minerals are not components of the host rock and are likely to have formed as a result of hydrothermal alteration. A total of 43 QI values were obtained based on XRD of cuttings from N19-HA-1, and a total of 45 QI values were obtained from N19-HA-2 (NEDO 2008). The average depth spacing of the values was approximately 38 m. Depth profiles based on the QI showed that each mineral had its own characteristic depth range (Fig. 4). We confirmed that the standard deviations of QI values were in the range of 0.01–0.17, as calculated from three measurements using different samples at the same depth. Such low values indicate the validity of the QI values.

Synthetic data for the evaluation of classification methodologies

We compared and evaluated three classification methods—K-means clustering, the GMM, and AC—using synthetic QI data. In this synthetic dataset, we assumed four different mineral distributions along a well down to 1000 m with a depth spacing of 10 m. Mineral distributions exhibited a Gaussian shape across the depth range, with mean μ_i $(i = 1, \cdots 4)$, standard deviation σ_i $(i = 1, \cdots 4)$, and maximum amplitude Amp_i $(i = 1, \cdots 4)$, as shown in Table 2. Gaussian noise with a mean of 0 and a standard deviation of 0.25 was added to each data

Table 1 List of 24 minerals identified from X-ray diffraction

Clay minerals	Smectite	Chlorite	Sericite	Kaolinite			
Zeolite minerals	Laumontite	Wairakite					
Silica minerals	Tridymite	Cristobalite					
Silicate minerals	Clinopyroxene	Epidote	Prehnite	Anthophyllite	Biotite	Cordierite	Talc
Oxide minerals	Magnetite	Ilmenite	Hematite	Anatase	Rutile		
Sulfide minerals	Marcasite						
Sulfate minerals	Anhydrite	Alunite					
Carbonate minerals	Calcite						

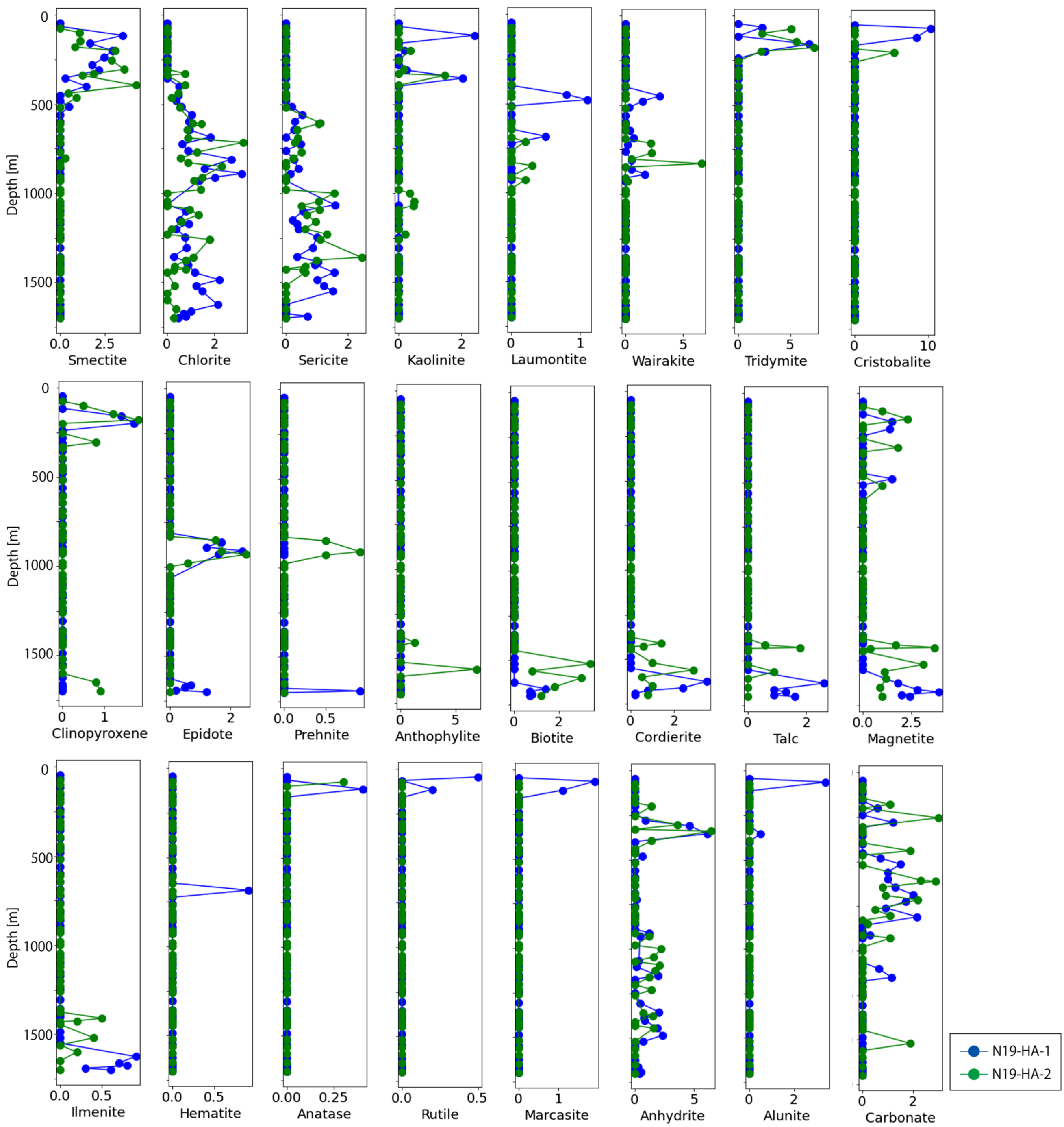


Fig. 4 Depth profiles of the quartz index (QI) of all minerals evaluated in this study. The horizontal axis indicates the QI, and the vertical axis indicates depth in the N19-HA-1 (blue dots) and N19-HA-2 (green dots) wells

Table 2 List of parameters used to create synthetic data

	Mineral 1	Mineral 2	Mineral 3	Mineral 4
μ_i(m)	200	450	700	780
σ_i(m)	80	80	30	80
Amp_i	1.5	1.0	0.7	0.5

point, to simulate the fluctuation in QI value. The depth profile of the synthetic data indicates that each mineral is distributed within a characteristic depth interval, and some depth intervals contain multiple minerals (Fig. 5a). We assumed temperature increases with depth at a constant rate of 0.2 degrees C/m.

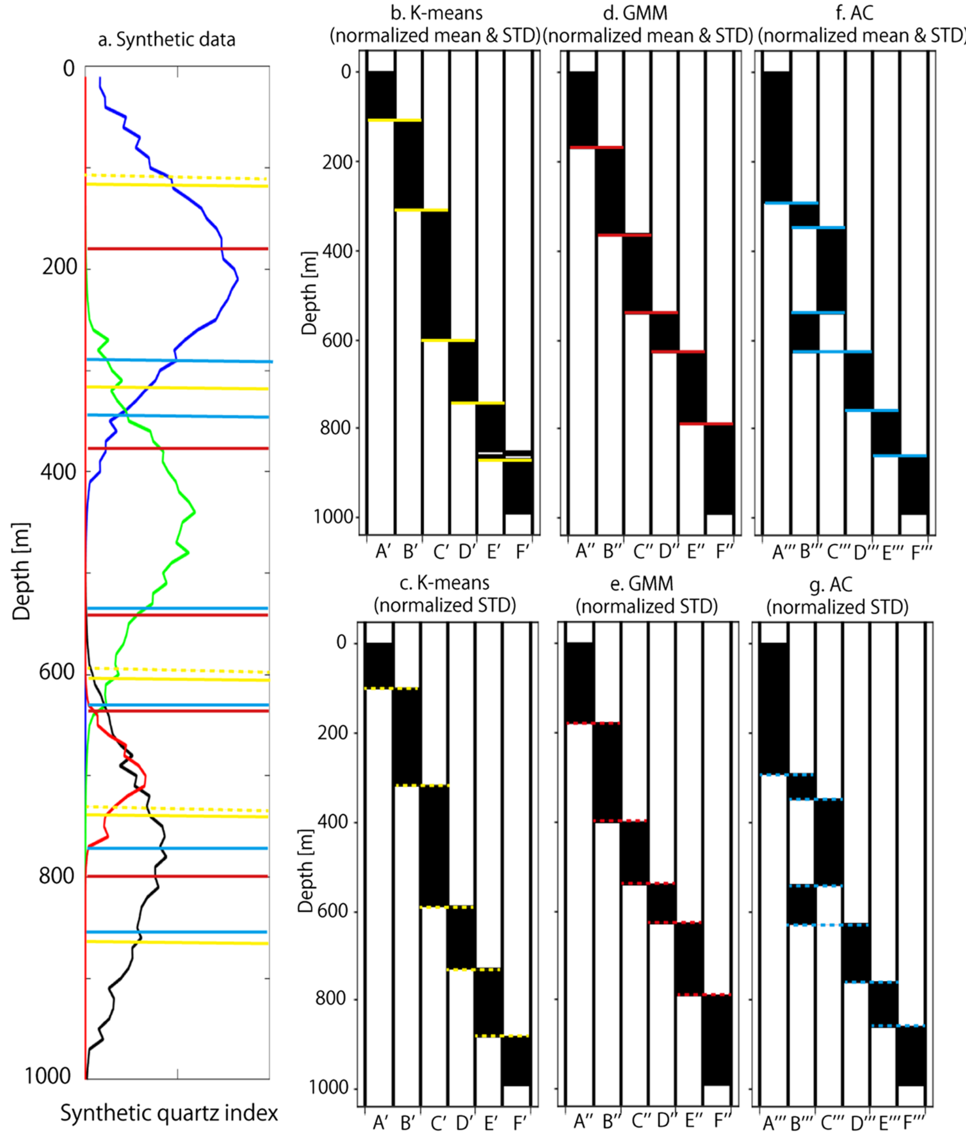


Fig. 5 Unsupervised classification based on **a** synthetic QI data using **b**, **c** K-means clustering, **d**, **e** the Gaussian mixture model (GMM), and **f**, **g** agglomerative clustering (AC). Synthetic data with different noise amounts were used, and classification was performed with **b**, **d**, **f** normalization of the mean and standard deviation (STD) or **c**, **e**, **g** normalization of the STD only. Blue, green, red and black curves in **a** indicate the synthetic QI values of different mineral types

Methods for classification and interpretation

As described above, K-means clustering, the GMM, and AC were applied to categorize the characteristics of hydrothermal alteration along wells. The data used for classification were the QI, temperature, and depth. Decision tree classification was then applied to the QI datasets to delineate each category's quantitative characteristics. This process was implemented using Python 3.7.3 and scikit-learn 0.19.2, a machine-learning library for Python (Pedregosa et al. 2011).

Preprocessing

Data classification typically entails preprocessing of observed data. A common preprocessing method is to normalize both the mean and standard deviation of the data. The effectiveness of this method is well-known, but normalization of the mean value produces negative QI values that are not found in the original data. To assess the impact of this preprocessing strategy, we applied additional preprocessing to normalize only the standard deviation of the observed data and compared the classification results of these two preprocessing methods.

K-means clustering

Suppose that the QI values are measured for N samples along a depth profile, and the data are represented as the matrix $\vec{x} = \left(\vec{x_1}, \cdots, \vec{x_n}, \cdots, \vec{x_N}\right)$ with the dimension of $N \times D$, where $\vec{x_n}$ is a D-dimensional vector $\vec{x_n} = \left(x_{n,1}, \cdots, x_{n,d}, \cdots, x_{n,D}\right)^T$ that contains the QI values of D different minerals. We now consider the classification of QI values from different samples into K classes. We assign a sample to the k^{th} class when the sample's QI is at the minimum distance from the mean of the data in the k^{th} class μ_k. This classification is attained by minimizing the following objective function J:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \|\vec{x_n} - \vec{\mu_k}\|^2 \tag{2}$$

where $r_{n,k} \in \{0,1\}$ is the coefficient that becomes 1 if x_n is classified into the k^{th} cluster and otherwise becomes 0. $r_{n,k}$ and μ_k are generally unknown, and the classification problem can be defined as the estimation of optimal $r_{n,k}$ and μ_k that minimize the objective function. This procedure is called "K-means clustering" (MacQueen 1967).

Gaussian mixture model

The GMM is a type of unsupervised learning in Bayesian modeling that assumes Gaussian distribution of classes (McLachlan and Peel 2000). The Gaussian mixture is defined as follows:

$$p(\vec{x_n}) = \sum_{k=1}^{K} \pi_k N\left(\vec{x_n} | \mu_k, \Sigma_k\right) \tag{3}$$

where $N\left(\vec{x} | \mu_k, \Sigma_k\right)$ is the Gaussian distribution of the k^{th} mixture component with a mean of μ_k and a mean and covariance matrix of Σ_k. π_k is the mixing coefficient that sums to 1 ($\sum_{k=1}^{K} \pi_k = 1$) such that the integral of the Gaussian mixture $p(\vec{x_n})$ is 1. In terms of classifying data $\vec{x}$, the GMM estimates the optimal μ_k, Σ_k, and π_k by fitting the Gaussian mixture and categorizes $\vec{x}$ into K clusters. Therefore, the objective function to be minimized is as follows:

$$\widehat{\mu_k}, \widehat{\Sigma_k}, \widehat{\pi_k} = \operatorname{argmax} \left\{ \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k N\left(x_n | \mu_k, \Sigma_k\right) \right) \right\} \tag{4}$$

where $\widehat{\mu_k}, \widehat{\Sigma_k}$, and $\widehat{\pi_k}$ indicate the optimal values of the mean, covariance matrix, and mixing coefficient, respectively. Because Eq. (4) cannot be solved analytically, numerical optimization is generally employed. In this study, the expectation–maximization algorithm was used for numerical optimization (Dempster et al. 1977; McLachlan and Krishnan 1997).

The K-means classification method is theoretically interpreted as cases wherein the covariance of the Gaussian mixture becomes zero and the classification depends on the distance from the mean (Bishop, 2006). This theoretical implication can be understood based on the distinction whereby the GMM method estimates both the mean and covariance matrix of the k^{th} class, whereas the K-means algorithm estimates the mean only.

Furthermore, the GMM calculates the probability of the classification from the mean and covariance matrix, and the probability can be used to obtain the optimal number of classes based on an information criterion, such as the Akaike information criterion (AIC) (Akaike 1974) adopted here.

$$\mathrm{AIC}\left(\vec{m}\right) = -2\ln L\left(\vec{x} | \vec{m}\right) + 2M \tag{5}$$

where $\vec{m}$ is the parameters to be estimated, and M is the number of parameters. AIC enables the comparison of models that include different numbers of classes, and the optimal number of classes has a minimum AIC value. The AIC values were obtained in cases where the number of classes ranged from 1 to 20.

Agglomerative clustering

AC classifies data by measuring the distances between data points (Lior and Maimon 2005). At the beginning of this algorithm, each data point forms its own class, leading to n single-object classes. Based on the measured distances, two classes with the shortest distance merge. The distances are measured again with the newly formed class, and a new class is created with the data points or clusters with the shortest distances. This procedure is repeated until the number of classes is identical to the a priori determined value or all distances are over a certain threshold. In this algorithm, the classification result is influenced by the choice of distance measure. In this study, Ward's method was applied (Ward 1963), which defines the distance of two clusters d_{z_i,z_j} as follows:

$$d_{z_i,z_j} = \sum \left(x_{z_i \cup z_j} - \mu_{z_i \cup z_j}\right)^2 - \sum \left(x_{z_i} - \mu_{z_i}\right)^2 - \sum \left(x_{z_j} - \mu_{z_j}\right)^2 \tag{6}$$

where z_i indicates the i^{th} cluster and $z_i \cup z_j$ indicates a new cluster that is obtained after merging z_i and z_j. x_{z_i} and μ_{z_i} are the data points in cluster z_i and the mean of the data in cluster z_i, respectively. Ward's method minimizes the variance of data in a new cluster with respect to the variance of data in two existing clusters.

Decision tree

A decision tree is a type of unsupervised classification that uses a flowchart-like graph to represent each test of attributes to explain known classes (Breiman 2001). More specifically, heuristic criteria are applied to attributes as a plausible boundary of different classes, and the tests are applied continuously until classification is complete or the number of tests reaches a certain threshold. Heuristic criteria were selected based on the Gini coefficient G in this study, as follows:

$$G = 1 - \sum_{k}^{K} p(k)^2 \tag{7}$$

where K is the total number of classes and $p(k)$ is the ratio of data that are categorized into a class. One of the major advantages of the decision tree is that the classification flow is explainable because the optimal heuristic criteria can be visualized via a flowchart. For the optimization of the decision tree, K-fold cross-validation was applied, which is a versatile and simple evaluation technique in machine learning. In K-fold cross-validation, the available dataset is divided into K subsets, with one subset used for validation and the other subsets used for training for supervised classification. Data in this study were divided into 10 non-overlapping subsets, resulting in tenfold cross-validation (James et al. 2014).

Results

Classification results using synthetic data

Figure 5b–g shows the classification results of three algorithms (i.e., K-means, the GMM, and AC), with each algorithm producing two sets of results representing the different preprocessing methods (i.e., normalization of the mean and standard deviation or normalization of the standard deviation). The boundaries of each class are depicted as horizontal lines in Fig. 5a. Based on the AIC values obtained from the GMM, six classes were determined to be optimal. Therefore, classification with six classes was performed for all three algorithms. Figure 5b–g shows that the classification results differed depending on the algorithm used, whereas the preprocessing strategies did not significantly influence the classification results. The reason for the variance in classification results among algorithms might have been each algorithm's different mathematical background. For example, Euclidean distance was used in the K-means algorithm, whereas Eq. (6) was used as a distance measure in the AC algorithm. Because no universal answer to unsupervised classification exists, different mathematical backgrounds can yield classification results from different viewpoints.

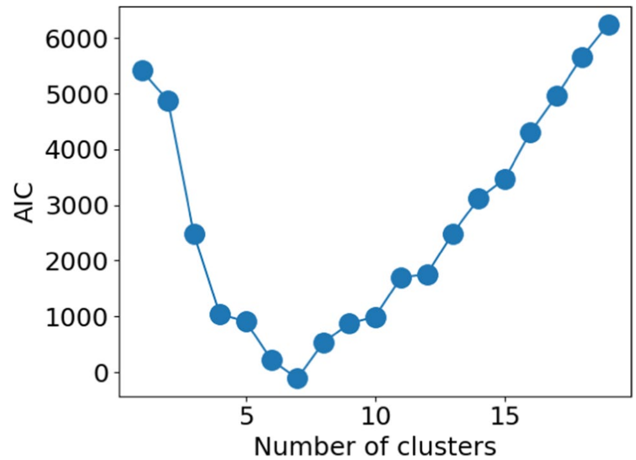


Fig. 6 Akaike information criterion (AIC) values as a function of the number of classes obtained from the GMM

Thus, an integrated interpretation of the classification from the different unsupervised algorithms may facilitate a more detailed extraction of the features inherent in the data.

Of the three classification results, the result of GMM classification is relatively easy to interpret. The first class (A′ and A″) corresponds to the shallowest synthetic QI (a blue line in Fig. 5a), and the second class (B′ and B″) roughly corresponds to the depth overlapping the shallowest and next shallowest synthetic QI values (blue and green lines in Fig. 5a). The following classes produced by the GMM also correspond to the QIs of various mineral types. Notably, this GMM classification result was similar to that of K-means clustering, possibly because the two algorithms share a similar mathematical background. On the other hand, the AC classification result was unique. The first class (A‴) covered a broader depth range compared with those of the K-means (A′) and GMM (A″) classifications, which corresponds to the shallowest synthetic QI (a blue line in Fig. 5a), and the classes at subsequent depths were similar to those from either the K-means or GMM classification. For example, the depth ranges for the classes C‴ and C″ were approximately equal, and the depth ranges for the classes D‴, E‴, and F‴ were similar to those for D′, E′, and F′. Overall, we showed that the classification results of the three algorithms were consistent with the distribution of the synthetic QI values. Thus, in the following application to real data, we used all three algorithms and compared their results to obtain the features buried within the data.

Classification results based on the quartz indices of Hachimantai geothermal field samples

For preprocessing, the standard deviation of the data was set to a value of 1 to normalize the data variation. The AIC values obtained from the GMM were larger for low numbers

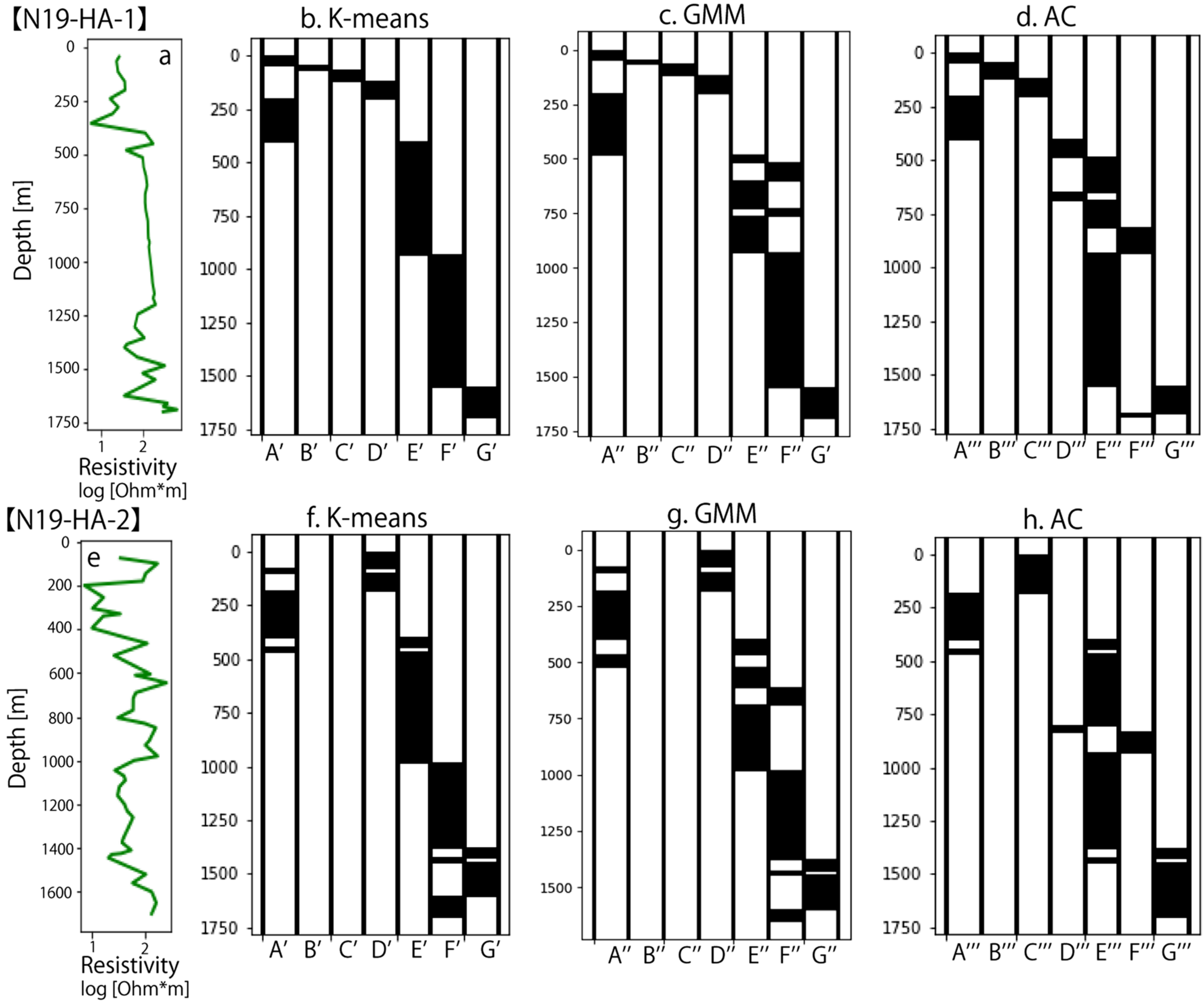
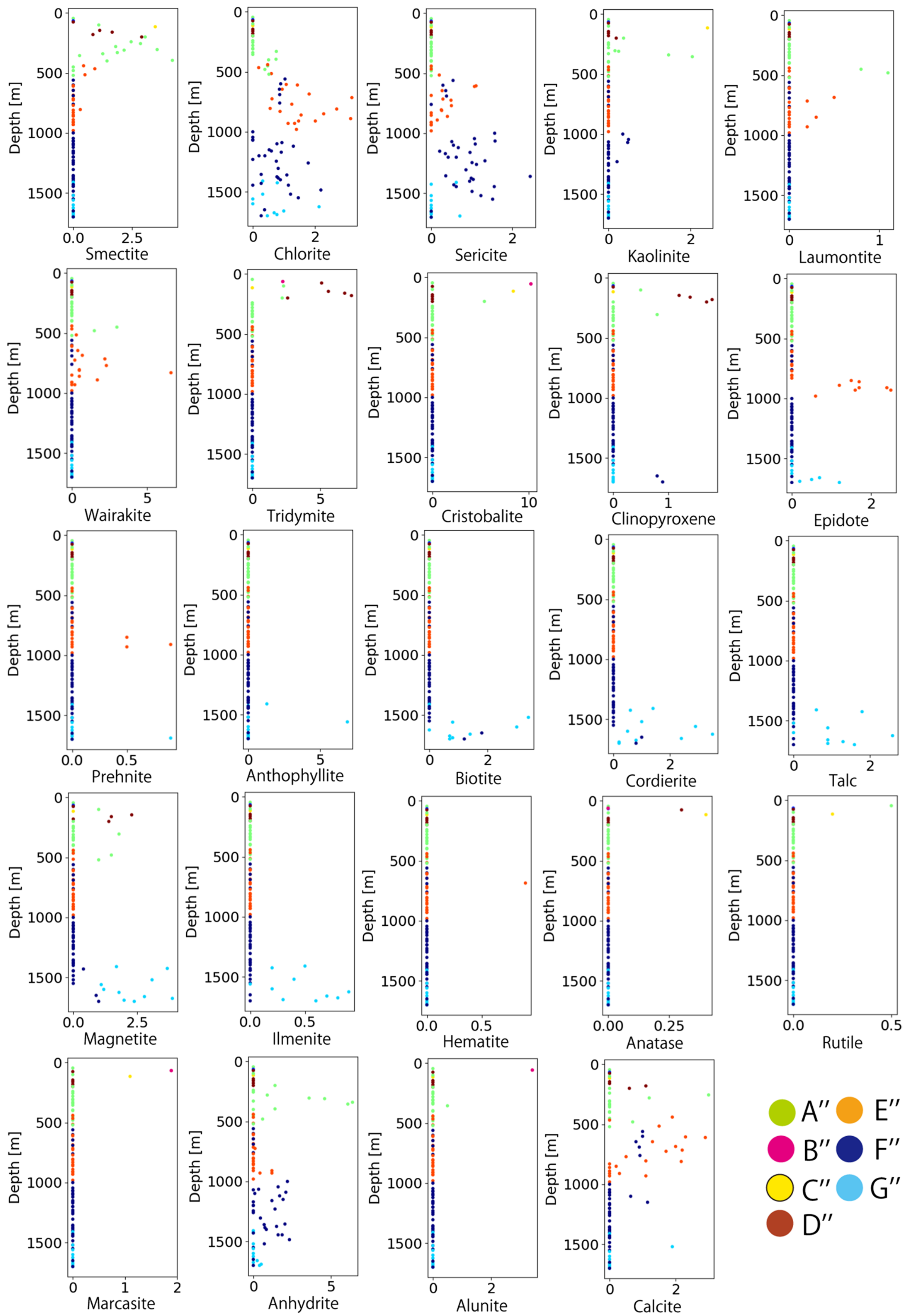


Fig. 7 Classification results of well logs (N19-HA-1 and N19-HA-2) using **b**, **f** K-means clustering, **c**, **g** the GMM, and **d**, **h** AC. The letters A–G denote the name of each cluster. Single, double, and triple quotations indicate that the classes were obtained from the K-means, GMM, and AC algorithms, respectively. Resistivity logs **a**, **e** are also attached to the classification results for comparison

of classes, and the value decreased with an increase in the number of classes (Fig. 6). However, if the number of classes exceeds 7, the AIC values increase again (Fig. 6). Because low AIC values relate to a statistically optimal model, the class number with the minimum AIC value is optimal, which in this study corresponds to seven classes (Fig. 6). Because the optimal distribution of classes approximately followed depth, the classes were labeled A to G (shallow to deep) (Fig. 7). Single, double, and triple quotation marks attached to the class names in Fig. 7 indicate that the classes were obtained using the K-means, GMM, and AC algorithms, respectively.

Classification using K-means clustering and the GMM was similar. The shallow layer of N19-HA-1, down to approximately 400–500 m (in the Quaternary layer), was divided into the classes A′, B′, C′, and D′ (GMM: A″, B″, C″, and D″) (Fig. 7b and c). However, the same depth range in N19-HA-2 was categorized into the classes A′ and D′ (A″ and D″) (Fig. 7f and g). Depths from 450 m to 1300–1500 m were categorized as classes E′ and F′ (E″ and F″) (Fig. 7b, c, f and g). Most of the depth distribution for classes E′ and E″ was shallower than that for classes F′ and F″, and was found in the first, second, and third Tertiary layers (Fig. 1b). Most of the minerals in classes F′ and F″ are found in the fourth Tertiary layer. Classes G′ and G″ included the deepest minerals, with depths of > 1500 m in N19-HA-1 and 1400 m in N19-HA-2 (Fig. 7b, c, f and g). Notably, the classification result is consistent with the resistivity logs (Fig. 7a and e). For example, classes A′, B′, C′, and D′ (A″, B″, C″, and D″) correspond to a low resistivity zone ($< 10^{1.5}$ Ω•m) in both


Depth [m]
Smectite
Chlorite
Sericite
Kaolinite
Laumontite
Wairakite
Tridymite
Cristobalite
Clinopyroxene
Epidote
Prehnite
Anthophyllite
Biotite
Cordierite
Talc
Magnetite
Ilmenite
Hematite
Anatase
Rutile
Marcasite
Anhydrite
Alunite
Calcite
A″ E″
B″ F″
C″ G″
D″

◂**Fig. 8** Cross-plots of depth and QI values of minerals in N19-HA-1 and N19-HA-2 well samples based on GMM classification. Note that the different colors represent different classes

N19-HA-1 and N19-HA-2, and the resistivities of classes E′ and E″ were approximately 10^2 Ω•m (Fig. 7).

Based on classification using the AC algorithm, the shallow part of N19-HA-1, down to approximately 400–500 m, was grouped into three classes (A‴, B‴ and C‴ in Fig. 7d). Comparison of the classification results across AC, K-means clustering, and the GMM showed that class B‴ included classes B′, B″, C′, and C″ from the K-means and GMM classifications (Fig. 7b, c and d). On the other hand, the class D‴ from AC classification did not correspond to any of the classes of the K-means and GMM classifications (Fig. 7). Class E‴ of AC roughly corresponded to classes F′ (F″) and E′ (E″) of the K-means (GMM) classification, whereas class F‴ did not correspond to any classes of the K-mean and GMM classifications (Fig. 7). Class G‴ of AC was comparable with class G′ (G″) of the K-means (GMM) classification (Fig. 7). Therefore, compared with the K-means and GMM classification results, unique information obtained from the AC algorithm were found in classes D‴ and F‴.

Based on the analysis of the synthetic data in Sect. 3.1 and the agreement with K-means clustering, we relied on GMM classification and illustrated cross-plots of the cluster classification using the GMM (Fig. 8). Figure 8 shows that most samples in class A″ were associated with large QI values corresponding to smectite and a certain amount of anhydrite. The low resistivity of class A″ minerals may be attributable to smectite. Class B″ contained relatively high concentrations of cristobalite, marcasite, and alunite compared to class C″, which contained kaolinite in addition to smectite (Fig. 8). Class D″ contained smectite, but to a lesser extent than in class A″, and characteristically contained tridymite, clinopyroxene, and magnetite (Fig. 8). The clay minerals in class E″ were largely chlorite and small amounts of sericite (Fig. 8). In addition, the QI values of wairakite and carbonate in class E″ were larger than those in the other clusters. Epidote was only observed at depths below that of class E″ (Fig. 8). Class F″ was characterized by larger amounts of sericite than those found in the other clusters, although smaller amounts of chlorite and anhydrite were also present in class F″ (Fig. 8). Class G″ also contained a certain amount of chlorite and was characterized by the presence of epidote, prehnite, biotite, cordierite, talc, magnetite, and ilmenite (Fig. 8).

Quantitative evaluation using a decision tree

A decision tree was further used to clarify the characteristics of the GMM classes (Fig. 9). In such a flowchart, information about the criteria used for classification is shown, such as the number of data points that satisfied each criterion and the number of data points within and outside each class. For example, in the class A″ decision tree, the first box shows that the total number of data points is 88, of which 16 were categorized into class A″, whereas 72 were not (Fig. 9). When a QI criterion for smectite of 1.005 was applied, 15 data points exceeded the criterion threshold, whereas the other 73 did not (Fig. 9). Of the 15 data points above the criterion threshold, 11 were in class A″, whereas four were not categorized into class A″. Subsequently, when a tridymite value of 2.45 was applied, 12 data points (including the 11 in class A″) were below the threshold (Fig. 9). Therefore, most of the data in class A have QI characteristics of smectite (>1.005) and tridymite (≤ 2.45) (Table 3). The decision trees support our qualitative interpretation described above, with the addition of quantitative QI information.

Other decision trees further delineated the class characteristics. For example, the class B″ decision tree showed that one dataset classified into the category had a cristobalite QI exceeding 9.35, and the class C″ decision tree showed that the category was associated with a kaolinite QI exceeding 5.978 (Fig. 9) (Table 3). In addition, the class D″ decision tree showed that the category was associated with a tridymite QI of more than 2.45 (Fig. 9) (Table 3), whereas the class E″ decision tree revealed more complex characteristics in the category. When the QI values of wairakite were >0.1 and those of laumontite were <0.65, 12 data points of 21 in class E″ were extracted (Fig. 9). However, when the wairakite QI values were <0.1, nine data points of 21 were still considered to be in class E″ (Fig. 9). The complexity of the category agrees with the qualitative observations of the cross-plots in Fig. 8. Moreover, the class F″ decision tree showed that sericite was a major mineral that characterized the class. When the QI of sericite was >0.28, most data points (30 of 34) in class F″ were extracted (Fig. 9) (Table 3). Interestingly, these 30 data satisfy the conditions of a wairakite QI value ≤ 0.1, cordierite QI value ≤ 0.1, and carbonate QI value ≤ 1.65 (Fig. 9) (Table 3). The class G″ decision tree indicated that nine out of 10 data points in class G had an ilmenite QI exceeding 0.1 (Fig. 9) (Table 3).

As Fig. 6c and g illustrate, classes B″ and C″ appear only at shallow depths in well N19-HA-1, but not in N19-HA-2. Because classes B″ and C″ are characterized by larger amounts of cristobalite and kaolinite, respectively, the shallower part of N19-HA-1 had experienced relatively higher temperatures and more acidic conditions than did N19-HA-2. Some classes successfully captured the characteristics of clay minerals, such as the characteristically large amounts of smectite in class A″ and the larger kaolinite, epidote, and sericite amounts characteristic of classes C″, E″, and F″. Clay species are generally controlled by the temperature and acidity of the hydrothermal environment. Therefore, the results presented

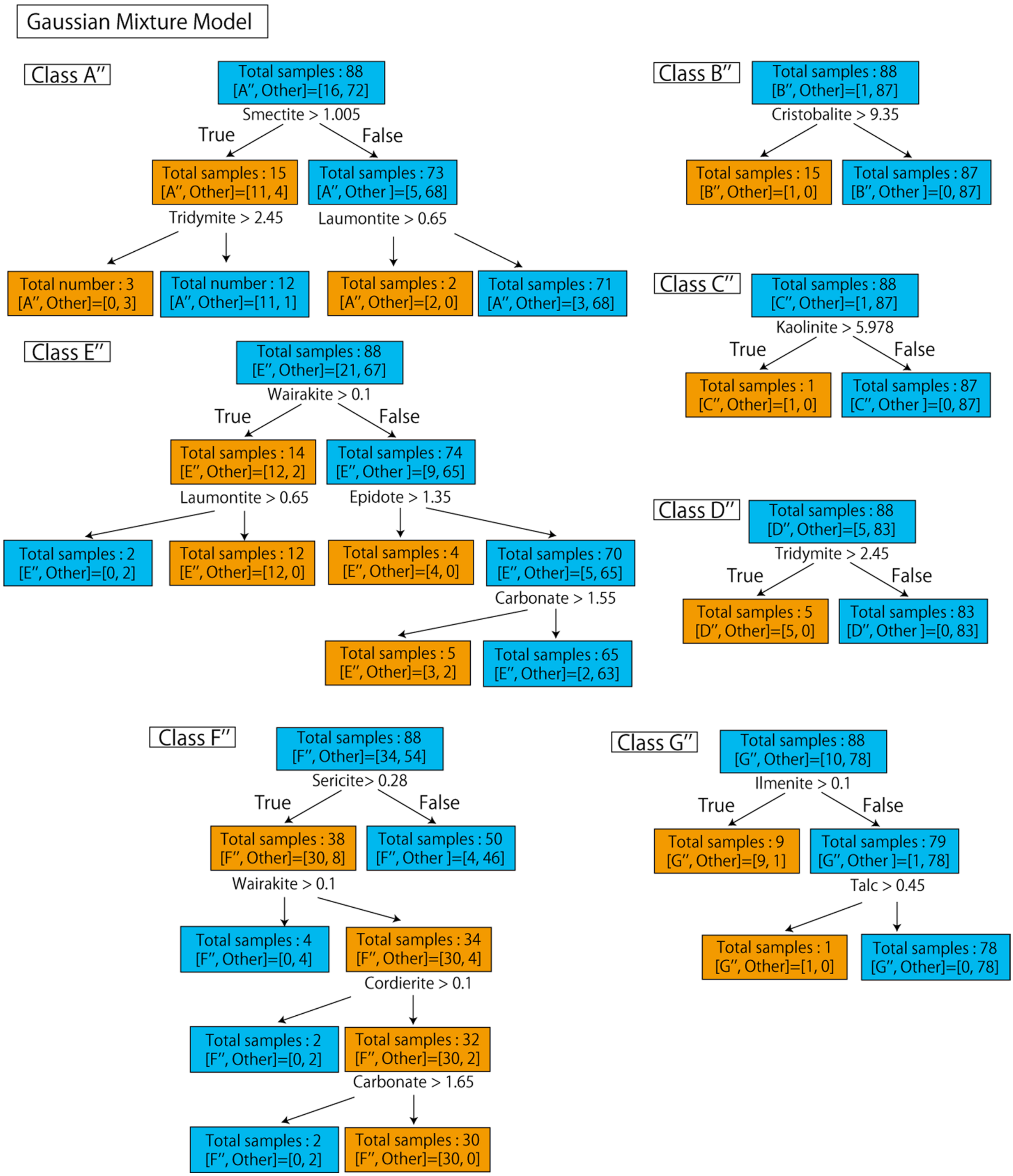


Fig. 9 Flowcharts representing decision trees of the classification results from the GMM, where classes A″–G″ correspond to the names of the classes shown in Fig. 7

demonstrate that the analytical method is capable of capturing variations in temperature and acidity in the hydrothermal system. Comparison of depth within a given group illustrates the features of alteration. The deepest class (class G″) most likely corresponds to hornfels, which are formed by contact metamorphism of mafic igneous rock.

Table 3 Summary of the minerals and their QIs that characterize each class

Class	K-means	Class	GMM	Class	AC
A′	Smectite > 0.88 Tridymite ≤ 2.45	A″	Smectite > 1.005 Tridymite ≤ 2.45	A‴	Smectite > 1.195 Tridymite ≤ 2.4
B′	Cristobalite > 9.35	B″	Cristobalite > 9.35	B‴	Cristobalite > 6.9
C′	Kaolinite > 2.235	C″	Kaolinite > 5.978	C‴	Tridymite > 2.25
D′	Tridymite > 2.45	D″	Tridymite > 2.45	D‴	Laumontite > 0.4 or Laumontite ≤ 0.4 Wairakite > 4.45
E′	Wairakite > 0.1 or Wairakite ≤ 0.1 Carbonate > 0.1 Chlorite > 0.82	E″	Wairakite > 0.1 Laumontite ≤ 0.65 or Wairakite ≤ 0.1 Epidote > 1.35	E‴	Sericite > 0.175 Talc ≤ 0.3
F′	Sericite > 0.505 Cordierite ≤ 0.1 Carbonate ≤ 0.82	F″	Sericite > 0.28 Wairakite ≤ 0.1 Cordierite ≤ 0.1 Carbonate ≤ 1.65	F‴	Epidote > 0.95 Biotite ≤ 0.35
G′	Ilmenite > 0.1	G″	Ilmenite > 0.1	G‴	Cordierite > 0.1

In well N19-HA-2, class G″ appeared shallower than it did in N19-HA-1, indicating that intrusive rock, which is considered a heat source, is more shallowly distributed around well N19-HA-2 (Fig. 7). However, the depths of the boundaries between smectite, chlorite, and sericite are almost identical.

As described in Sect. 4.2, classes D‴ and F‴ from AC exhibited unique characteristics compared with the GMM classification. The class D‴ decision tree showed

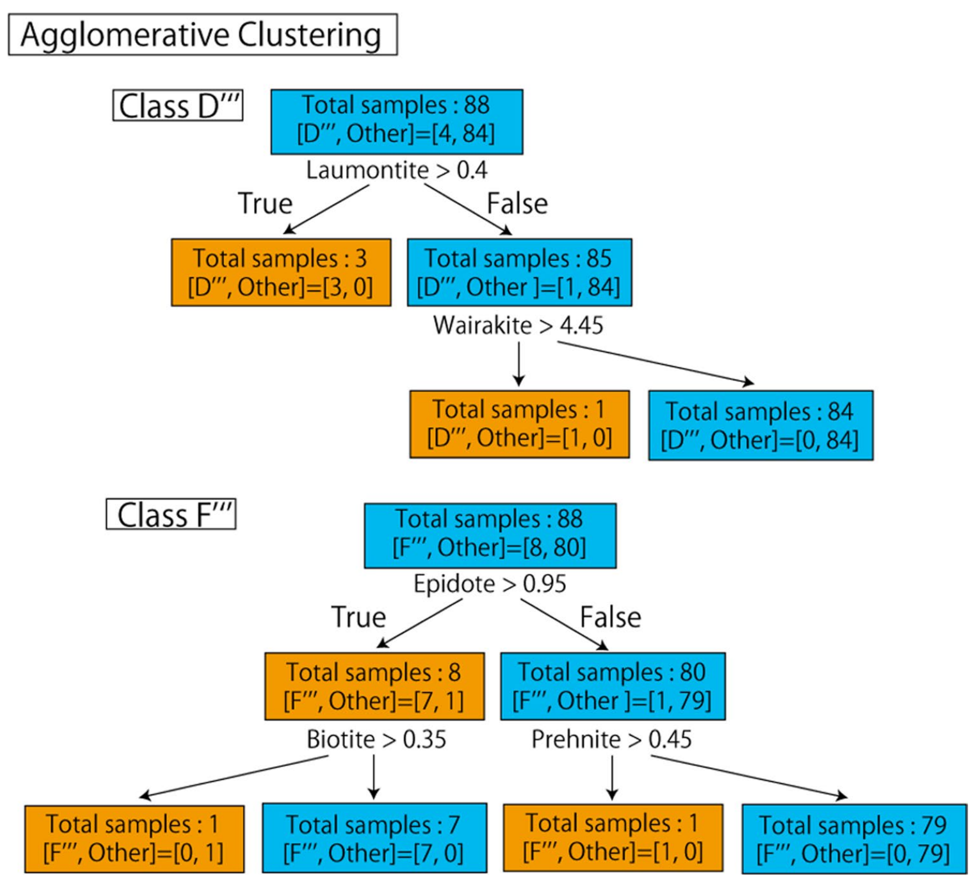


Fig. 10 Flowcharts representing decision trees of the classification results from AC. Classes D‴ and F‴ correspond to the names of the classes shown in Fig. 7

that laumontite and wairakite were important minerals for characterizing this class (Fig. 10). When the QI of laumontite was > 0.40, most data points (3 of 4) in class D‴ were extracted (Fig. 10) (Table 3). When the QI of laumontite was ≤ 0.40 but the QI of wairakite was > 4.45, the data were classified into class D‴ (Fig. 10) (Table 3). The class F‴ decision tree showed that most data were classified as this class when the QI of epidote was > 0.95 and the QI of biotite was ≤ 0.35 (Fig. 10) (Table 3). These features of classes D‴ and F‴ were not clearly identified by the GMM. Thus, our classification results demonstrated that AC is useful for extracting information that complements the GMM and K-means clustering. As the summary of the results by a decision tree, the minerals and their QI thresholds that characterize each class derived from the three classification algorithms are presented in Table 3.

Conclusions

In this study, we examined the performance of three unsupervised classification algorithms—K-means clustering, the GMM, and AC—in automatically classifying the QI values and temperature logs of geothermal wells. These methods enable categorization of zones with similar mineral characteristics. In particular, K-means clustering and the GMM provided similar classification results and were used to verify the classification results. On the other hand, AC provided unique classification outcomes that were not apparent in the results of the above two algorithms. Furthermore, the characteristics of each class could be delineated using a decision tree, which has the advantage of generating a comprehensive flowchart. The classification of QI values at the Hachimantai geothermal field revealed the connectivity of the two geothermal wells as they share similar characteristics. Moreover, the classification analysis detected higher acidity in N19-HA-1 compared with N19-HA-2. As the QI can be measured from rock cuttings and the algorithms do not require training data, the proposed approach is applicable to other boreholes in geothermal fields as well as boreholes in any other Earth science and engineering projects.

Abbreviations AC: Agglomerative clustering; AIC: Akaike information criterion; GMM: Gaussian mixture model; ICDD: International centre for diffraction data; QI: Quartz index; XRD: X-ray powder diffraction

Funding Part of this study was conducted within a program of the National Institute of Advanced Industrial Science and Technology (Fukushima Renewable Energy Institute) to support industry in the disaster area of the 2011 Great East Japan Earthquake. K.I. was also funded partly by the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant no. 20K15219).

Data availability The datasets analyzed during the current study are available in http://geothermal.jogmec.go.jp/gathering/nedo.html

Declarations

Competing interests The authors declare no competing interests.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Browne PRL (1978) Hydrothermal alteration in active geothermal fields. Ann Rev Earth Planets Sci 6:229–250

Caté A, Perozzi L, Gloaguen E, Blouin M (2017) Machine learning as a tool for geologists. Lead Edge 36(3):215–219. https://doi.org/10.1190/tle36030215.1

Caté A, Schetselaar E, Mercier-Langevin P, Ross RS (2018) Classification of lithostratigraphic and alteration units from drillhole lithogeochemical data using machine learning: a case study from the Lalor volcanogenic massive sulphide deposit, Snow Lake, Manitoba, Canada. J Geochem Explor 188:216–228. https://doi.org/10.1016/j.gexplo.2018.01.019

Chen S, Hattori K, Grunsky EC (2018) Identification of sandstones above blind uranium deposits using multivariate statistical assessment of compositional data, Athabasca Basin, Canada. J Geochem Explor 188:229–239. https://doi.org/10.1016/j.gexplo.2018.01.026

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39(1):1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Delayre C, Mas PP, Sardini P, Cosenza P, Thomas A (2020) Quantitative evolution of the petrophysical properties of andesites affected by argillic alteration in the hydrothermal system of Petite Anse-Diamant, Martinique. J Volnanol Geotherm Res 401:106927. https://doi.org/10.1016/j.volgeores.2020.106927

Feng R (2020) Lithofacies classification based on a hybrid system of artificial neural networks and hidden Markov models. Geophys J Int 221:1484–1498. https://doi.org/10.1093/gji/ggaa083

Frolova JV, Ladygin VM, Rychagov SN (2010) Petrophysical alteration of volcanic rocks in hydrothermal systems of the Kuril-Kamchatka Island Arc, Proceedings of World Geothermal Congress 2010, Bali, Indonesia.

Frolova, JV, Gvozdeva IP, Kuznetsov NP (2015) Effects of hydrothermal alterations on physical and mechanical properties of rocks in the Geysers Valley (Kamchatka Peninsula) in connection with landslide development, Proceedings of World Geothermal Congress 2015, Melbourne, Australia.

Fulignati P, Malfitano G, Sbrana A (1997) The Pantelleria caldera geothermal system: Data from the hydrothermal minerals. J Volnanol Geotherm Res 75:251–270. https://doi.org/10.1016/S0377-0273(96)00066-2

Grana D, Fjeldstad T, More H (2017) Bayesian Gaussian mixture linear inversion for geophysical inverse problems. Math Geosci 49:493–515. https://doi.org/10.1007/s11004-016-9671-9

Gower JC, Ross GJS (1969) Minimum spanning trees and single linkage cluster analysis. J R Stat Soc Ser C Appl Stat 18(1):54–64. https://doi.org/10.2307/2346439

Hayashi M (1979) Quantitative descriptions of cores and cuttings from geothermal wells. J Geotherm Res Soc Jpn 1(2):103–116. https://doi.org/10.11367/grsj1979.1.103 (**(in Japanese with English abstract)**)

He M, Gu H, Wan H (2020) Log interpretation for lithology and fluid identification using deep neural network combined with MAHAKIL in a tight sandstone reservoir. J Petrol Sci Eng 194:107498. https://doi.org/10.1016/j.petrol.2020.107498

Hood SB, Cracknell MJ, Gazley MF (2018) Linking protoloth rocks altered equivalents by combining unsupervised and supervised machine learning. J Geochem Explor 186:270–280. https://doi.org/10.1016/j.gexplo.2018.01.002

Inoue A, Meunier A, Beaufort D (2004) Illite-smectite mixed-layer minerals in felsic volcaniclastic rocks from drill cores, Kakkonda, Japan. Clay Clay Miner 52(1):66–84. https://doi.org/10.1346/CCMN.2004.0520108

James G, Witten D, Hastie T, Tibshirani R (2014) An Introduction to statistical learning: With applications in R. Springer, New York

Kimbara K (1985) An overview of the geothermal system in the Sengan geothermal area, northern Japan. J Geotherm Res So Jpn 7(3):189–200. https://doi.org/10.11367/grsj1979.7.189 (**(in Japanese with English abstract)**)

Lior R, Maimon O (2005) Clustering methods, Data mining and knowledge discovery handbook. Springer US, 321–352.

Lutz SJ, Zutshi A, Robertson-Tait A, Drakos P, Zemach E (2011) Lithologies, hydrothermal alteration, and rock mechanical properties in wells 15–12 and BCH-3, Bradys hot springs geothermal field, Nevada. GRC Trans 35(1):469–476

MacQueen JB (1967) Some methods for classification and analysis of multivariate observations, Proc. Fifth Berkeley Symp. on Math Statist. and Prob., 281–297.

McLachlan GJ, Krishnan T (1997) The EM algorithm and its extensions. Wiley

McLachlan G, Peel D (2000) Finite mixture models. Wiley

Mielke P, Nehler M, Bignall G, Sass I (2015) Thermo-physical rock properties and the impact of advancing hydrothermal alteration – a case study from the Tauhara geothermal field, New Zealand. J Volnanol Geotherm Res 301:14–28. https://doi.org/10.1016/j.jvolgeores.2015.04.007

New Energy and Industrial Technology Development Organization (NEDO) (2007) Mid-term Report on geothermal promotion survey at the Hachimantai area (2nd), 387 pp. (in Japanese)

New Energy and Industrial Technology Development Organization (NEDO) (2008) Report on geothermal promotion survey at the Hachimantai area (3rd), 562 pp. (in Japanese)

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Raeesi M, Moradzadeh A, Ardejani FD, Rahimi M (2012) Classification and identification of hydrocarbon reservoir lithofacies and their heterogeneity using seismic attributes, logs data and artificial neural networks. J Pet Sci Eng 82–83:151–165. https://doi.org/10.1016/j-petrol.2012.01.012

Reyes AG (1990) Petrology of Philippine geothermal systems and the application of alteration mineralogy to their assessment. J Volnanol Geotherm Res 43:279–309. https://doi.org/10.1016/0377-0273(90)90057-M

Saporetti CM, da Fonseca LG, Pereira E, de Olieira LC (2018) Machine learning approaches for petrographic classification of carbonate-siliciclastic rocks using well logs and textual information. J Appl Geophys 155:217–225. https://doi.org/10.1016/j.jappgeo.2018.10.013

Schiffman P, Fridleifsson GO (1991) The smectite-chlorite transition in drillhole NJ-15, Nesjavellir geothermal field, Iceland: XRD, BSE and electron microprobe investigations. J Metamorphic Geol 9:679–696. https://doi.org/10.1111/j.1525-1314.1911.tb00558.x

Takahashi R, Matsuda H, Okrugin M, Ono S (2007) Epithermal gold-silver mineralization of the Asachinskoe deposit in South Kamchatka, Russia. Resour Geol 57(4):354–373. https://doi.org/10.1111/j.1751-3928.2007.00034.x

Templ M, Filzmoser P, Reimann C (2008) Cluster analysis applied to regional geochemical data: problems and possibilities. Appl Geochem 23:2198–2213. https://doi.org/10.1016/j.apgeochem.2008.03.004

Ueki K, Hino H, Kuwatani T (2018) Geochemical discrimination and characteristics of magmatic tectonic settings: a machine-learning-based approach. Geochem Geophys Geosys 19(4):1327–1347. https://doi.org/10.1029/2017GC007401

Ward JH (1963) Hierarchical grouping to optimize and objective function. J Am Stat Assoc 58:236–244

Wyering LD, Villeneuve MC, Wallis IC, Siratovich PA, Kennedy BM, Gravley DM, Cant JL (2014) Mechanical and physical properties of hydrothermally altered rocks, Taupo Volcanic Zone, New Zealand. J Volnanol Geotherm Res 288:76–93. https://doi.org/10.1016/j.jvolgeores.2014.10.008

Yang K, Browne PRL, Huntington JF, Walshe JL (2001) Characterising the hydrothermal alteration of the Broadlands-Ohaaki geothermal system, New Zealand, using short-wave infrared spectroscopy. J Volnanol Geotherm Res 106:53–65. https://doi.org/10.1016/S0377-0273(00)00264-X

Yoneda T (2014) Mineralogical properties of clay minerals and its usefulness as an index in exploration of natural resources. J Min Mater ProcessInst Jpn 130:1–8 (**(in Japanese with English abstract)**)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.