Earth, Planets and Space

FULL PAPER Open Access

Quantitative logging data clustering with hidden Markov model to assist log unit classification

Suguru Yabe[1*], Yohei Hamada[2], Rina Fukuchi[3], Shunichi Nomura[4], Norio Shigematsu[1], Tsutomu Kiguchi[1] and Kenta Ueki[5]

Abstract

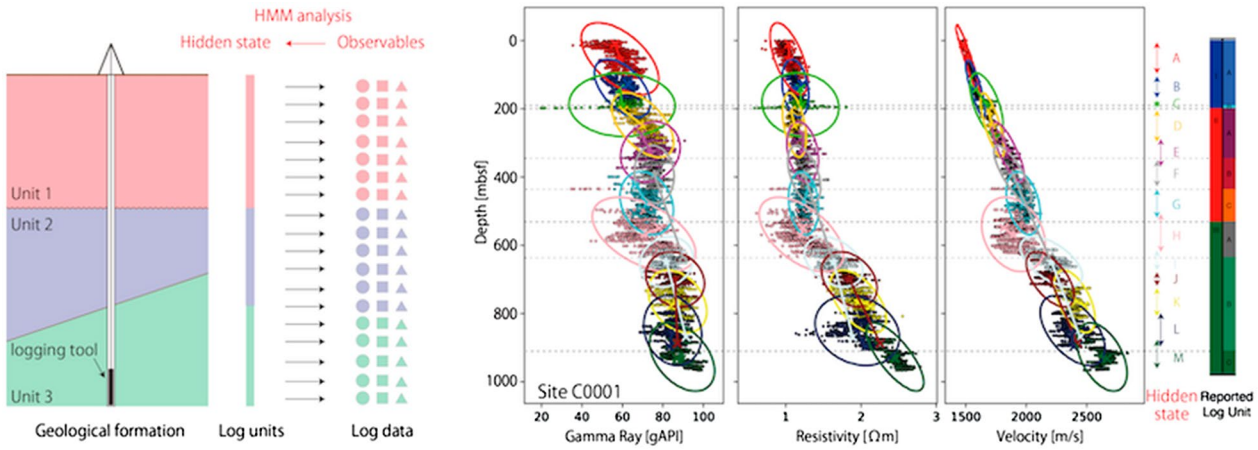
Revealing subsurface structures is a fundamental task in geophysical and geological studies. Logging data are usually acquired through drilling projects, which constrain the subsurface structure, and together with the description of drill core samples, are used to distinguish geological units. Clustering is useful for interpreting logging data and making log unit classification and is usually performed by manual inspection of the data. However, the validity of clustering results with such subjective criteria may be questionable. This study proposed the application of a statistical clustering method, the hidden Markov model, to conduct unsupervised clustering of logging data. As logging data are aligned along the drilled hole, they and the geological structure hidden behind such sequential datasets can be regarded as observables and hidden states in the hidden Markov model. When log unit classification is manually conducted, depth dependency of logging data is usually focused. Therefore, we included depth information as observables to explicitly represent depth dependency of logging data. The model was applied to the following geological settings: the accretionary prism at the Nankai Trough, the onshore fault zone at the Kii Peninsula (southwest Japan), and the forearc basin at the Japan Trench. The optimum number of clusters were searched using a quantitative index. The clustering results using the hidden Markov model were consistent with previously reported classifications or lithological descriptions; however, our method allowed a more detailed division of logging data, which is useful to interpret geological structures, such as a fault or a fault zone. Therefore, the use of the hidden Markov model enabled us to clarify assumptions quantitatively and conduct clustering consistently for the entire depth range, even for different geological sites. The proposed method is expected to have wider applicability and extensibility for other types of data, including geochemical and structural geological data.

Keywords: Hidden Markov model, Logging data, Clustering, Unit classification

*Correspondence: s.yabe@aist.go.jp

[1] Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology, Tsukuba Central 7, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8567, Japan
Full list of author information is available at the end of the article

Graphical Abstract



Main Text

Introduction

Revealing subsurface structures is a fundamental task in various geophysical and geological studies. For example, examining stratigraphy and fault distributions helps geologists interpret the temporal evolution of geological formations (Tanaka et al. 2001; Strasser et al. 2009). Exploring the spatial variation in physical properties along major faults can lead to an improved understanding of the seismic behavior of faults (Moore and Saffer 2001; Saffer 2007; Kimura et al. 2007; Lockner et al. 2011; Sutherland et al. 2017). In the petroleum industry, identification of antiform structures is crucial to detect possible oil reservoirs (Harding 1974). Geological surveys on the surface, geophysical exploration, and drilling for the recovery of drill core samples and logging data are the main strategies used to estimate subsurface structures. Each method has different types of data and resolutions, which are complementary to each other. For efficient interpretations, combination of various kinds of data from different strategies, known as "core-log-seismic integration, has been suggested (Cerchiari et al. 2018). An appropriate statistical framework is required to combine various datasets with different units and qualities, although such an approach has not been established.

During scientific drilling, such as the International Ocean Discovery Program (IODP), acquired logging data are often used to classify logs into log units as basic information for interpreting the geological structure at the site. In the case of IODP, the standard criteria for log unit classification are based on qualitative and quantitative analyses (Expedition 314 Scientists 2009a). Qualitative analyses include "identification of the boundaries separating sections of different log responses and concomitant rock properties" (Expedition 314 Scientists 2009a), which have been conducted subjectively based on discussions among onboard scientists. Although quantitative analyses have been performed to validate the qualitative analyses, they only include "investigating the percentile ranges and distribution of absolute values within the visually defined logging units" (Expedition 314 Scientists 2009a), which does not rely on any statistical models. Although log unit classifications are usually reasonable, their quantitative validity should be justified by some statistical methods. Based on a statistical method, automated clustering provides a team of scientists with quantitative basis to discuss geological interpretations of the obtained data.

Townend et al. (2013) conducted a principal component analysis for wire-line logging data to characterize the detailed structure of the hanging wall of the Alpine Fault, New Zealand. They used seven types of logging data [natural gamma ray (NGR), borehole diameter, neutron porosity, compensated density, *P*-wave velocity, electrical resistivity, and spontaneous potential] as inputs. Their results showed that seven-dimensional logging data are well represented by the first, second, and third principal components, which represent the electric, acoustic, and NGR characteristics, respectively. However, their unit classification depended on the lithological units from the core descriptions. Logging data for each lithological unit overlapped in the principal component domain, resulting in imperfect reconstruction of the lithology of the cored

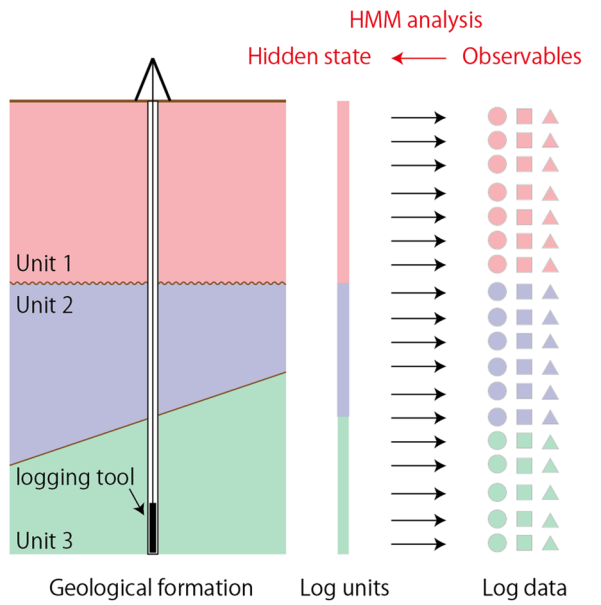


Fig. 1 Schematic concept of Hidden Markov model for logging data clustering

section. A similar approach was performed by Townend et al. (2017).

Another statistical clustering method with the Markov chain assumption has been applied to lithology classification using geophysical exploration data (e.g., Eidsvik et al. 2004; Larsen et al. 2006; Hammer et al. 2012; Lindberg and Omre 2014; Feng et al. 2018). A hidden Markov model (HMM) is a relatively simple and useful approach among this line of studies (Schumann 2002; Jeong et al. 2014; Tian et al. 2021). HMM is often used as a supervised machine-learning technique to predict unknown lithology after training the model with datasets of known lithology. However, logging data clustering for log unit classification, which is the main focus of this study, cannot be accomplished with such supervised HMM.

This study proposed the application of a statistical clustering method with an unsupervised HMM for logging data clustering. Unsupervised clustering methods usually require analysts to subjectively determine various criteria, such as the number of clusters to be divided. A statistical approach enables us to represent the assumed clustering criteria quantitatively and provides a quantitative basis and validity for log unit clustering, assisting in making geological interpretations and determining log unit classification.

In the next section, we provide a formulation of the clustering method using HMM. Thereafter, the results for the three applications at different geological settings are presented. Finally, we discuss the stability of the clustering results and the applicability and extensibility of the proposed method. The proposed data-driven method could be used to determine the appropriate number of clusters. The obtained clusters were not only consistent with previously reported log units, but also provided interpretations for finer structures.

Hidden Markov model

HMM is a statistical approach that estimates a sequence of unobservable (hidden) states. It is widely used for analyzing sequential data and has applications in voice recognition (Jelinek 1997) and bioinformatics (Baldi et al. 2001). At every sequential step, the hidden state may remain in the same state or change to other states according to transition probabilities, which depend on the current state (i.e., Markov process). Although we cannot observe hidden states directly, they are estimated from observable quantities recorded at every sequential step. These observables are generated according to a generation probability distribution, which depends on the hidden state. The HMM is suitable for estimating geological structures (e.g., Schumann, 2002; Eidsvik et al. 2004). The geological structures to be estimated cannot be directly measured (i.e., hidden states); instead, they must be interpreted based on various measurements of rocks (i.e., observables). In our problem setting of logging data, the hidden states and observables correspond to log units and logging data, respectively (Fig. 1). This study proposes the application of HMM to logging data and inverting a sequence of log units.

For simplicity, this study considered logging data with a constant sampling interval: all types of logging data used in the analysis were acquired at the same depth. Additionally, we did not consider missing data. Although the methodology can be extended to such incomplete data sets, in this study, if the original datasets were incomplete, we created complete data sets by interpolation. The types of logging data are expected to be quantities that represent the physical properties of geological formation, such as electrical resistivity, natural gamma ray (NGR), velocity, and porosity. In addition, we used depth information as observables in the HMM. As was observed in the applications of the model, logging data sometimes showed significant depth dependency. In such cases, the values of the logging data had different meanings according to their depth. Therefore, in this study, we added depth data to observables in the HMM-based clustering to represent depth dependency.

The HMM is formulated as follows (e.g., Bishop 2006). We used D-dimensional vectors of logging data x_i for $i = 1, \ldots, N$, where N is the total number of sequential data. The number of hidden states (log units) is assumed to be K. K-dimensional vectors $z_i \text{for} i = 1, \ldots, N$ represent the hidden state at the i-th sequential step such that

they have 1 for the k-th component ($z_{i,k}$) and 0 for others when the i-th sequential step is at the k-th hidden state ($k = 1, \ldots, K$). The probability of choosing the hidden state in the first sequential step can be written with a discrete distribution as follows:

$$P(z_1|\pi) = \prod_{k=1}^{K} \pi_k^{z_{1,k}}, \tag{1}$$

where π is a K-dimensional vector that represents the probability of the k-th hidden state to be the hidden state in the first sequential step. The prior distribution for Eq. (1) can be written with a Dirichlet distribution as follows:

$$P(\pi) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_{0,k}-1}, \tag{2}$$

where $C(\alpha_0)$ represents normalizing coefficients and can be rewritten as

$$C(\alpha_0) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_{0,k})}{\prod_{k=1}^{K} \Gamma(\alpha_{0,k})}. \tag{3}$$

Similarly, the transition probability of the hidden state at each sequential step and its prior distribution are written with discrete and Dirichlet distributions as follows:

$$P(z_n|z_{n-1}, A) = \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{z_{n-1,j} z_{n,k}}, \tag{4}$$

$$P(A) = \prod_{j=1}^{K} C(\alpha_j) \prod_{k=1}^{K} A_{jk}^{\alpha_{jk}-1}, \tag{5}$$

where A is a $K \times K$ matrix, whose jk component represents the transition probability from the j-th hidden state to the k-th hidden state.

We assumed normal distribution for the generation probability of observables as follows:

$$P(x_n|z_n, \mu, \Lambda) = \prod_{k=1}^{K} G\left(x_n|\mu_k, {\Lambda_k}^{-1}\right)^{z_{n,k}}, \tag{6}$$

where G represents the normal distribution, and μ_k and Λ_k are the mean vector and precision matrix of the generation probability for the k-th hidden state, respectively. Its prior probability was assumed to be a Gaussian–Wishart distribution, which is a conjugate prior distribution for the normal distribution as follows:

$$P(\mu, \Lambda) = \prod_{k=1}^{K} G\left(\mu_k|m_0, (\beta_0 \Lambda_k)^{-1}\right) W(\Lambda_k|w_0, \upsilon_0), \tag{7}$$

where β_0 and υ_0 are scalar hyperparameters, and m_0 and w_0 are the vector and matrix hyperparameters, respectively. The function W represents Wishart distributions that can be represented as below:

$$W(\Lambda_k|w_0, \upsilon_0) = B|\Lambda_k|^{(\upsilon_0 - D - 1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(w_0^{-1}\Lambda_k\right)\right), \tag{8}$$

where B is a normalizing factor, and $\mathrm{Tr}(\bullet)$ is trace of a matrix. Although we assumed the normal distribution as the generation probability of observables and the Gauss–Wishart distribution as its prior distribution in this study, other probability distributions can be used according to the characteristics of the logging data used.

The joint probability can be written using Eqs. (1), (2), and (4), (5), (6), (7) as follows:

$$P(x, z, \pi, A, \mu, \Lambda) = P(z_1|\pi)P(\pi)\left\{\prod_{n=2}^{N} P(z_n|z_{n-1}, A)\right\}P(A)\left\{\prod_{n=1}^{N} P(x_n|z_n, \mu, \Lambda)\right\}P(\mu, \Lambda). \tag{9}$$

The model was optimized by monitoring the evidence value $P(x)$, which is the marginal probability with respect to the model parameters $Z = (z, \pi, A, \mu, \Lambda)$. As the evidence is difficult to maximize directly, it can be rewritten using variational approximation as follows (Jordan et al. 1998; Jaakkola 2001):

$$\mathrm{Ln}P(x) = L(q) + KL(q\,\|P), \tag{10}$$

where KL is the Kullback–Leibler divergence, which measures a distance from a probability distribution P to another one q.

$$KL(q\,\|P) = -\int q(Z) ln\left(\frac{P(Z|x)}{q(Z)}\right) \mathrm{d}Z. \tag{11}$$

Here, q is the posterior distribution of Z given x.

$$L(q) = \int q(Z) \ln \frac{P(x,Z)}{q(Z)} \mathrm{d}Z. \tag{12}$$

As $KL(q\|P)$ is non-negative, $L(q)$ represents the infimum of $\ln P(x)$. Assuming,

$$q(z,\pi,A,\mu,\Lambda) = q(z)q(\pi,A,\mu,\Lambda) \tag{13}$$

$L(q)$ can be maximized using the iterative E–M algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997). We used the forward–backward algorithm (Baum 1972; Rabiner 1989) in the E step of the E–M algorithm to calculate the expectations of the model parameters. Optimization is regarded to converge when the increase in $L(q)$ is less than 10^{-4}. As the solutions obtained through the E–M algorithm depend on the initial assumptions on μ_k and Λ_k, we prepared 100 sets of initial average vectors μ_k stochastically using the K-means++method (Arthur and Vassilvitskii 2007). The initial Λ_k was set to w_0. We adopted the best model among them based on the approximated evidence values $L(q)$. After the optimization, the most possible sequence of hidden states was estimated using the Viterbi algorithm (Viterbi 1967). The depth ranges of the clusters were characterized by 5th and 95th percentiles.

The HMM described above contains six hyperparameters ($K, \alpha, \beta_0, \upsilon_0, m_0$ and w_0). The hyperparameter α in Eqs. (2) and (5) controls how often hidden states (in other words, log unit) change to another state at each sequential step. In this study, all components of the vector $\alpha_j (j = 0, \ldots, K)$ were set to one, which

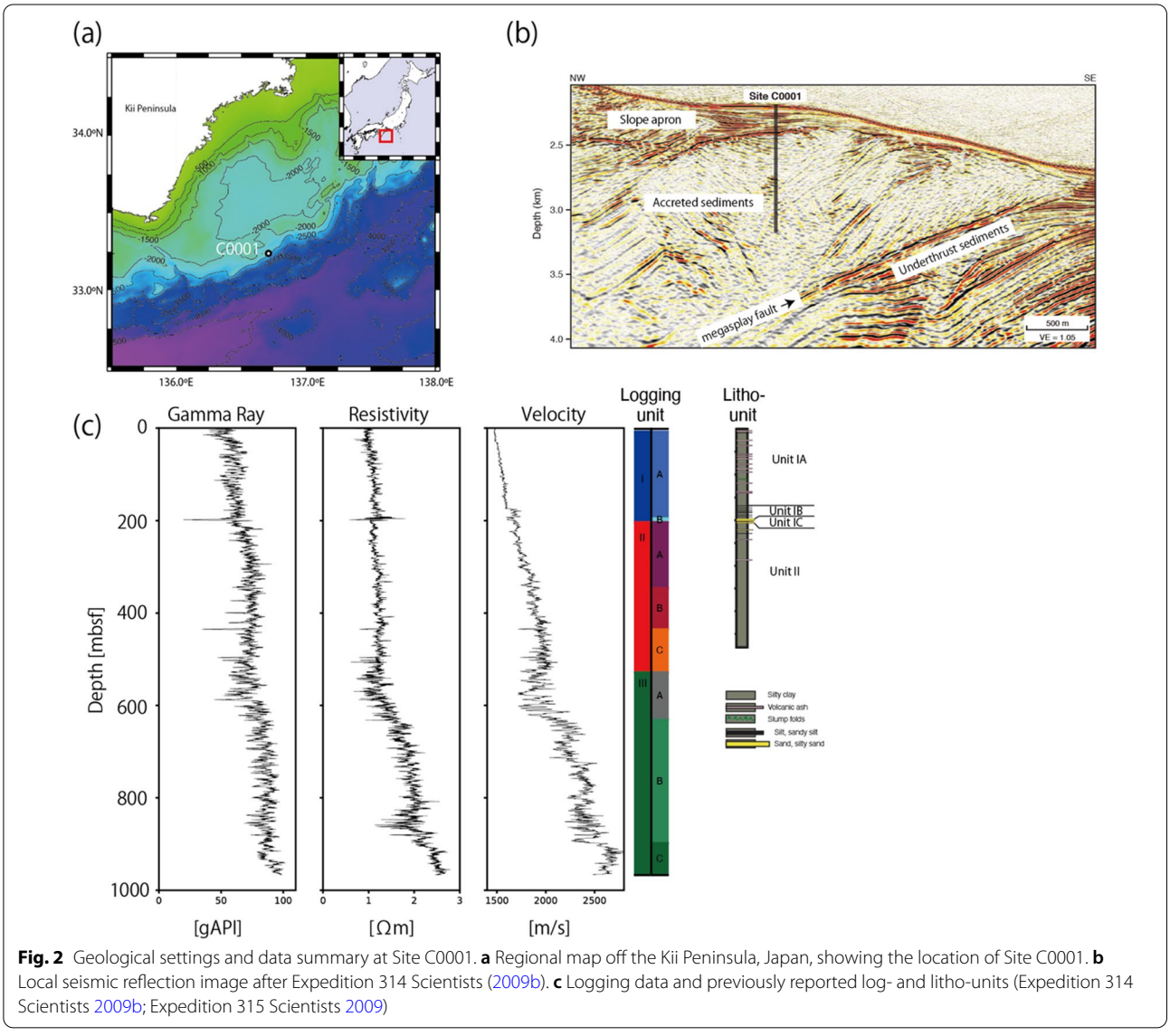


Fig. 2 Geological settings and data summary at Site C0001. **a** Regional map off the Kii Peninsula, Japan, showing the location of Site C0001. **b** Local seismic reflection image after Expedition 314 Scientists (2009b). **c** Logging data and previously reported log- and litho-units (Expedition 314 Scientists 2009b; Expedition 315 Scientists 2009)

represented a uniform distribution of the Dirichlet distribution, i.e., prior does not have any information on how to change hidden states. The hyperparameters β_0, υ_0, m_0 and w_0 in Eq. (7) control generation probability of observables (i.e., logging data and depth). In particular, the hyperparameters m_0 and w_0 are related to expectations of mean value and precision matrix of the prior distribution. We assumed those values based on input data such that m_0 was set as a median vector of the input data and w_0 as a precision matrix of the input data. The hyperparameter β_0 has a positive value, while the hyperparameter υ_0 has a value $\geq D$. These values are considered as weight for the prior distribution. Larger values result in a posterior distribution closer to the assumed prior distribution. We empirically assumed that β_0 is 1 and υ_0 is D. We conducted grid searches to determine K between 2 and 25. Although model selections for K are possible with the evidence value $L(q)$, we used different criteria to determine K. As shown in the following sections, the evidence value shows a broad peak at a large K. As geological interpretations become difficult when clustering results have too many clusters, we used a different index to determine K as follows:

$$X = \sum_{k=1}^{K} \sum_{i \in \{i; z_{i,k}=1\}} (x_i - \mu_k)\beta_k \Lambda_k (x_i - \mu_k) + \sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \left(\mu_{k_1} - \mu_{k_2}\right) \frac{\beta_{k_1}\Lambda_{k_1} + \beta_{k_2}\Lambda_{k_2}}{2} \left(\mu_{k_1} - \mu_{k_2}\right), \tag{14}$$

where, $\beta_k = \beta_0 + \sum_{i=1}^{N} E\left(z_{i,k}\right)$ and $E(\cdot)$ represents expectation.

In summary, clustering of logging data using HMM has the following explicit and quantitative assumptions: (A) a Markov process for log unit transition, (B) a normal distribution for logging data are generation probability, (C) variational approximation (Eq. 13), (D) assumed values of five hyper parameters $(\alpha, \beta_0, \upsilon_0, m_0, w_0)$, and (E) the model selection criteria (Eq. 14). Based on these assumptions, HMM clustering was conducted using data from different geological settings, as described in the next section. Statistical clustering with quantitative criteria can avoid biases due to different interpreters for different datasets and different subjective criteria due to different visuals of data from different geological settings.

Application results

In this section, we provide three examples of applications using the proposed method. HMM clustering was applied to logging data from different geological settings: a young accretionary prism at the Nankai Trough, onshore fault zones at the Kii Peninsula, Japan, and offshore coal bed in the forearc basin of the Japan Trench. The appropriate number of clusters was searched for different geological settings based on the index X (Eq. 14). The obtained results showed that HMM clustering divides logging data into the appropriate numbers of clusters, and that every cluster has a geological and/or geophysical meaning.

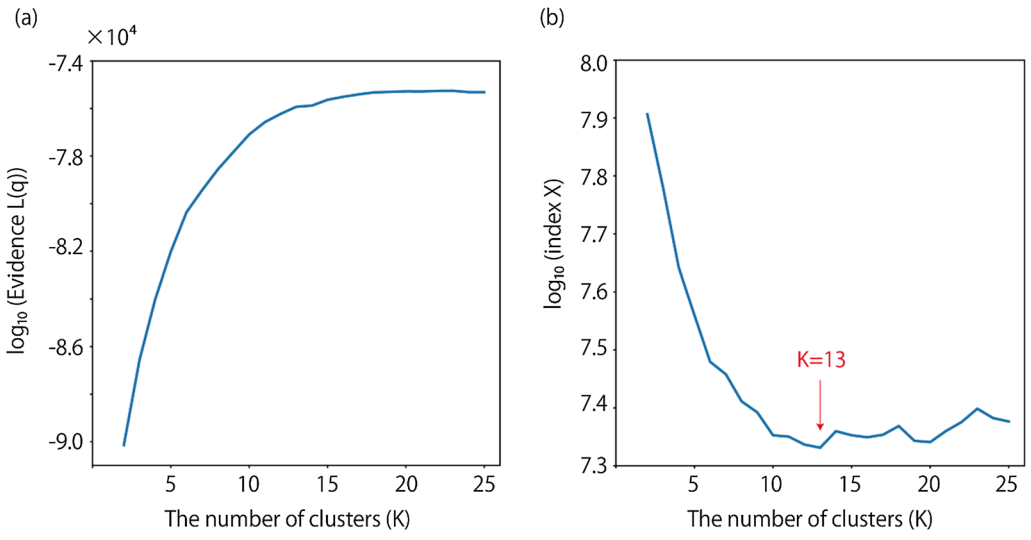


Fig. 3 Grid search results of Evidence $L(q)$ and index X for the number of clusters (K) at Site C0001

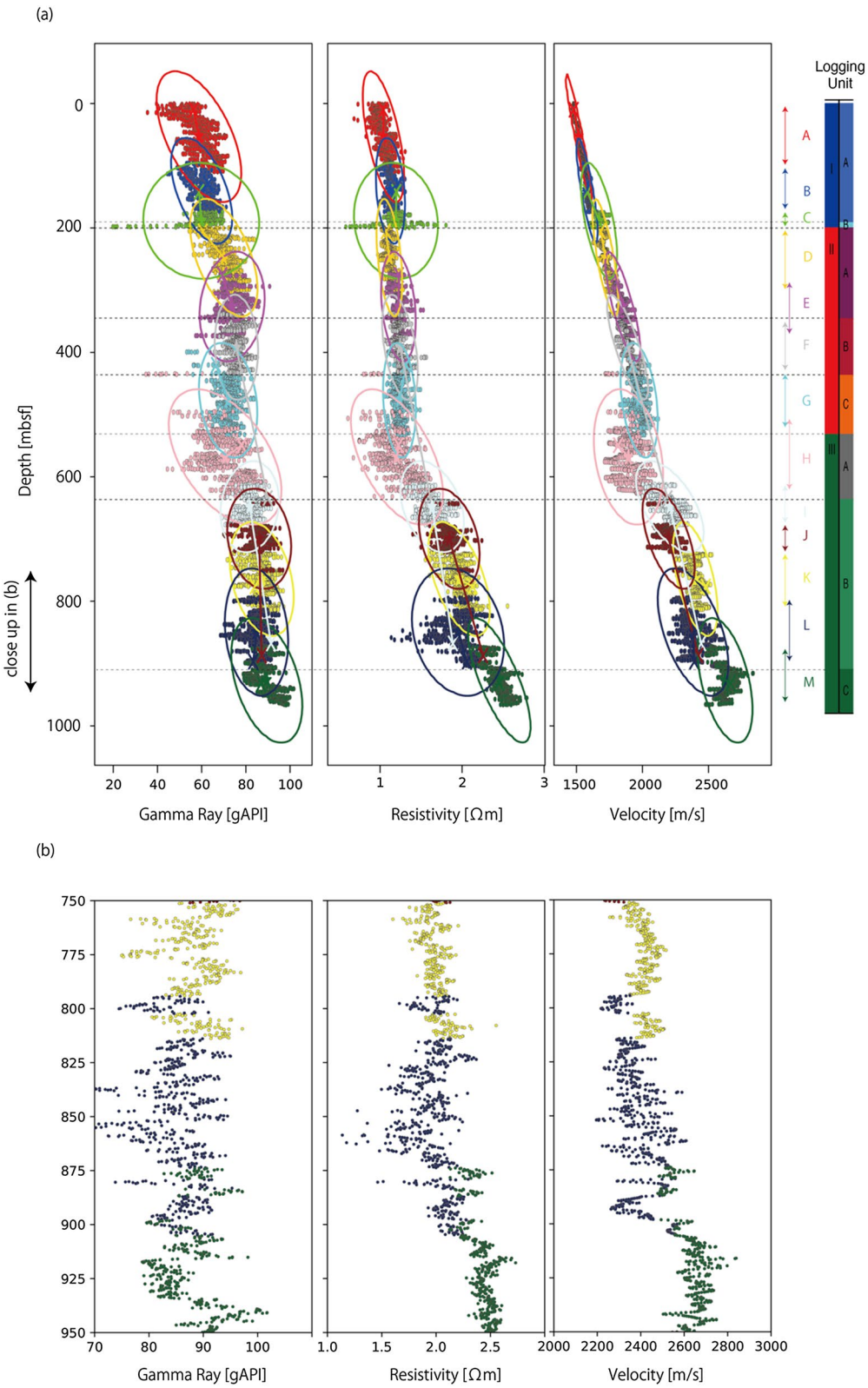


Fig. 4 Clustering results at C0001 and previously reported log units (Expedition 314 Scientists 2009b). Ellipses represent covariance of each cluster. **a** The entire section, and **b** the close-up figure of **a** where noted by a black arrow on the left

Table 1 Clustering results for Site C0001

Clusters	Depth (mbsf)	Previously reported log units	Previously reported lithological units
A	5–99	Unit 1 0–198.9	Unit I 0–207
B	105–171		
C	175–198		
D	204–299	Unit 2 198.9–529.1	Unit II 207–456
E	288–371		
F	352–430		
G	435–521		
H	507–621	Unit 3 529.1–976	Not cored
I	614–675		
J	678–721		
K	725–810		
L	799–897		
M	878–963		

Young accretionary prism at the Nankai Trough

Geological settings and data with previous interpretations The Nankai Trough is located off the coast of southwest Japan, where megathrust earthquakes have repeatedly occurred (Ando 1975). It is a typical subduction zone where an accretionary prism developed at its toe (e.g., Moore et al. 1990). Here, the Philippine Sea Plate (PSP) subducts at a rate of ~6 cm/year (Miyazaki and Heki 2001; DeMets et al. 2010; Kimura et al. 2018), and thick sediments on PSP have been accreted. Many scientific drillings have been conducted at this margin to understand the seismogenesis and evolutionary processes of the subduction zone (Tobin et al. 2020).

This study focused on Site C0001 (Fig. 2), which is located on the hanging wall of the major out-of-sequence thrust called "megasplay." This site was logged by IODP Expedition 314 (Expedition 314 Scientists 2009b) and cored by IODP Expedition 315 (Expedition 315 Scientists 2009). Although logs were obtained down to ~976 m below the sea floor (mbsf), coring was conducted down to ~456 mbsf. We used the electrical resistivity log, NGR log, and *P*-wave velocity log for the analysis. The sampling interval of logging data was 15 cm.

Expedition reports (Expedition 314 Scientists 2009b) classified logging data into three units, which were further divided into eight subunits (Expedition 314 Scientists 2009b). Based on the logging data and seismic reflection images, unit 1 (0–198.9 mbsf) was interpreted to be the slope apron facies and divided into two subunits (Unit 1a and 1b at 0–190.5 and 190.5–198.9 mbsf, respectively) with four decametric cycles (0–54, 54–100, 100–156, and 156–191 mbsf). Unit 1b was characterized by alternating beds of conductive and resistive sediments with a negative peak of NGR at its bottom. Both Units 2 (198.9–529.1 mbsf) and 3 (529.1–976 mbsf) represented accreted sediments. Differences in the depth dependency of the electrical resistivity log divided the two units. Unit 2 was divided into three subunits (Unit 2a, 2b, and 2c at 198.9–344.0, 344.0–434.7, and 434.7–529.1 mbsf, respectively). Unit 2a had features similar to those of unit 1a. Units 2b and 2c had similar resistivity, although they had different NGR and velocity values. Unit 3 was divided into three subunits (Unit 3a, 3b, and 3c at 529.1–628.6, 628.6–904.9, and 904.9–976 mbsf, respectively). Unit 3a had transitional features from Unit 2 to Unit 3, which was interpreted as a mass transport deposit or fault zone. Faults and fractures were also reported at 650, 800, 835, and 860 mbsf based on resistivity image logs.

Expedition reports also classified the lithological descriptions of drill core samples into two units (Expedition 315 Scientists 2009). They were interpreted as slope aprons (Unit I) and accreted sediments (Unit II), with a boundary at 207 mbsf. Unit I was further divided into three subunits (Unit Ia, Ib, and Ic at 0–168.35, 168.35–196.76, and 196.76–207.17 mbsf, respectively). Unit Ia was composed of hemipelagic mud with volcanic ash; unit Ib, hemipelagic mud with silt turbidites; and unit Ic, sand and hemipelagic mud. In contrast, Unit II was a single unit without any subunits comprising hemipelagic mud. Although the overall unit classifications of logging units and lithological units were similar, there were differences in subunit classifications.

Results We conducted a grid search for the number of clusters to obtain the optimized model. Figure 3 shows the results of grid search. The approximated evidence value was high and flat at a large K range, whereas the index X showed a minimum at $K=13$. The sequence of the estimated hidden states is shown in Fig. 4. The clustering results are summarized in Table 1. Boundaries of clusters were set at depths where step changes of logging values were observed or where depth dependencies of logging values were changed. For example, NGR log showed a small step decrease at the boundary between clusters A and B. Depth dependency of resistivity log also changed at this depth, showing a slight increase in cluster A and constant in cluster B. Cluster C was a thin layer including spiky signals in NGR and resistivity. Clusters D and E were differentiated by the depth dependency of the NGR log. It showed a slight increase in cluster D, whereas it became constant in cluster E. Clusters E and F somewhat overlapped but have different features in velocity log, such that cluster F had a slower velocity than cluster E at the same depth. The boundary between clusters F and G was determined by the changes in the velocity log. The increasing trend with depth in cluster F changed to constant values in

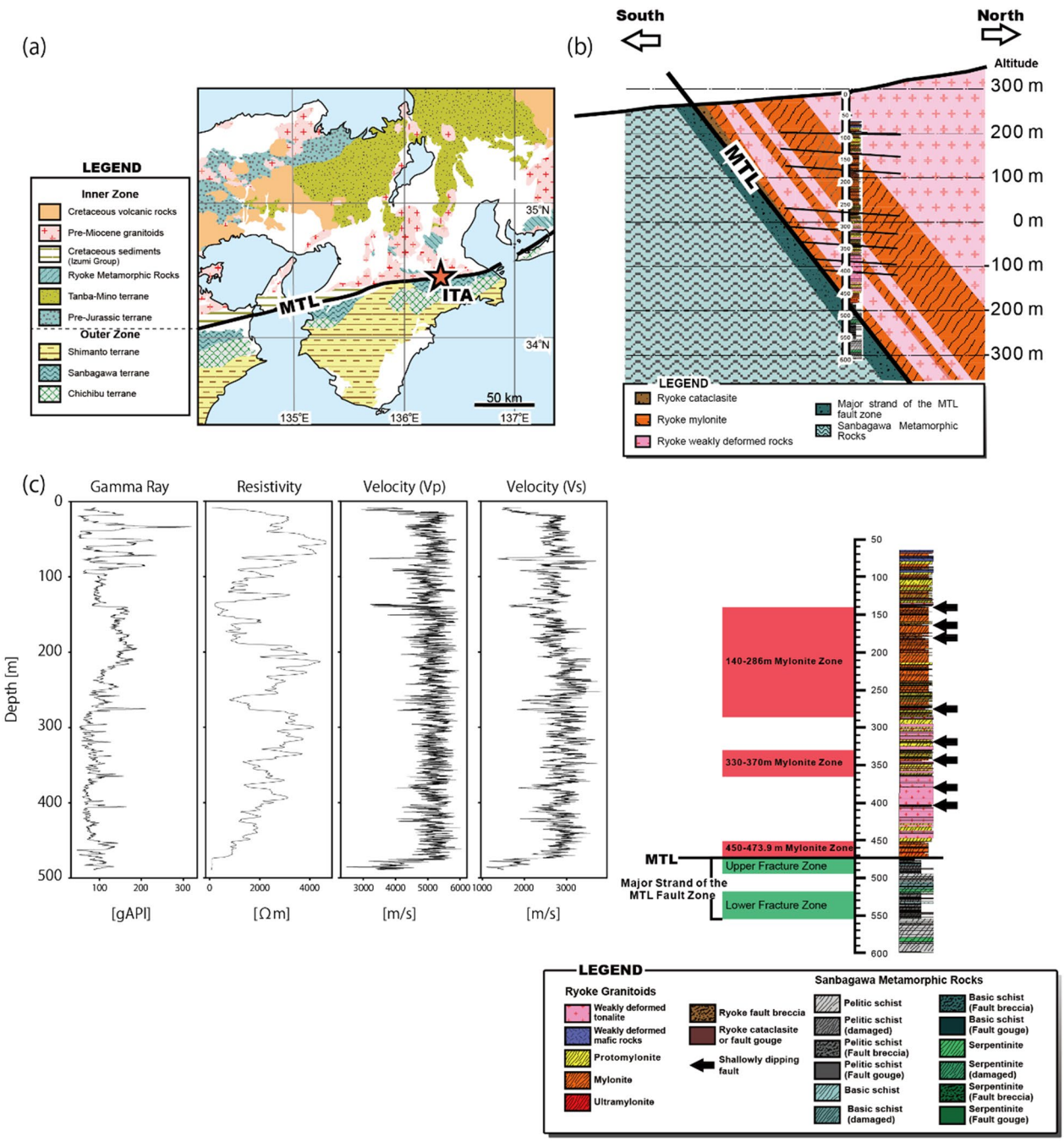


Fig. 5 Geological map and data summary at Site ITA. **a** Regional map of the Kii Peninsula showing the location of Site ITA based on Shigematsu et al. (2012). **b** Interpreted cross-section image around Site ITA after Shigematsu et al. (2012). **c** Logging data and core description previously reported by Shigematsu et al. (2012)

cluster G. Cluster H was characterized by dropping values for all logs. Cluster J was characterized by slower velocity than the surrounding clusters I and K. Cluster L was characterized by dropping values in all logs, which was similar to cluster H. Clusters L and M were differentiated by higher resistivity and velocity values for cluster M.

The clustering results obtained by HMM were in good agreement with the log unit classification of expedition scientists. The top three clusters (A–C) roughly

corresponded to Unit 1 (0–198.9 mbsf). Another four clusters (D–G) corresponded to Unit 2 (198.9–529.1 mbsf). The top depth of cluster F agreed with the boundary between the log units 2a and 2b. The boundary between clusters F and G also agreed with the boundary between log units 2b and 2c. Unit 3 (529.1–976 mbsf) consisted of six clusters (H–M). The bottom depths of clusters H and L corresponded to the boundary between log units 3a and 3b and between log units 3b and 3c, respectively.

HMM clustering divided log unit 1a into two clusters (A and B) at ~100 mbsf. These two clusters were further divided based on differences in the depth dependency of resistivity and the negative step of NGR at the boundary. The depth of the boundary corresponded to one of the boundaries of the decametric cycles in log unit 1A documented in the expedition report (Expedition 314 Scientists 2009b).

Cluster C was thicker than the corresponding log unit (1b). The top of cluster C was defined by a step increase in *P*-wave velocity. Above this step increase, the velocity log was unreliable because of the very low formation velocity close to the drilling fluid velocity (Expedition 314 Scientists 2009b). Therefore, the top depth of cluster C was considered to be artificial. The bottom depth of cluster C was defined as the depth at which the steps are observed in three logs.

HMM clustering divided log unit 3b into four clusters (I–L) at ~675, ~725, and ~800 mbsf. Cluster L was characterized by a significant drop in resistivity and velocity logs. Cluster J was characterized by a drop in the velocity log. This feature was similar to log unit 3a, which was interpreted to be a fault zone or mass transport deposit (Expedition 314 Scientists 2009b).

As explained above, clustering results by HMM were in good agreement with the log unit classification of expedition scientists. In addition, HMM clustering further suggested a finer structure in log units 1a, 2a, and 3b. On the other hand, lithological unit classification of expedition scientists did not suggest such fine structures in the accretionary prism. Hence, the different characteristics of each log unit could be mainly attributed to differences in physical properties due to the different deformation histories of each geological formation.

Onshore fault zone at Kii Peninsula, Japan

Geological settings and data with previous interpretations

The Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology constructed several integrated groundwater observatories to monitor plate-boundary motion along the Nankai Trough subduction zone in southwestern Japan, including the Kii Peninsula (Itaba et al. 2010). At one of the observatories, ITA (Fig. 5), a borehole (ITA-1) penetrated the Median Tectonic Line (MTL) (Shigematsu et al. 2012). This study focused on borehole ITA-1, which is located on the hanging wall and 300 m north of the MTL. The MTL divides the low-P/high-T Ryoke metamorphic terrain in the north from the high-P/low-T Sanbagawa terrain in the south and is the longest onshore fault in the Japan arc (Wallis and Okudaira 2016). The rocks around the observatory ITA are variably affected by the faultings along the MTL (Shigematsu et al. 2012, 2017; Mori et al. 2015; Katori et al. 2021).

Wireline logging data are acquired during the drilling process (Kiguchi et al. 2014). The drilling of ITA-1 was

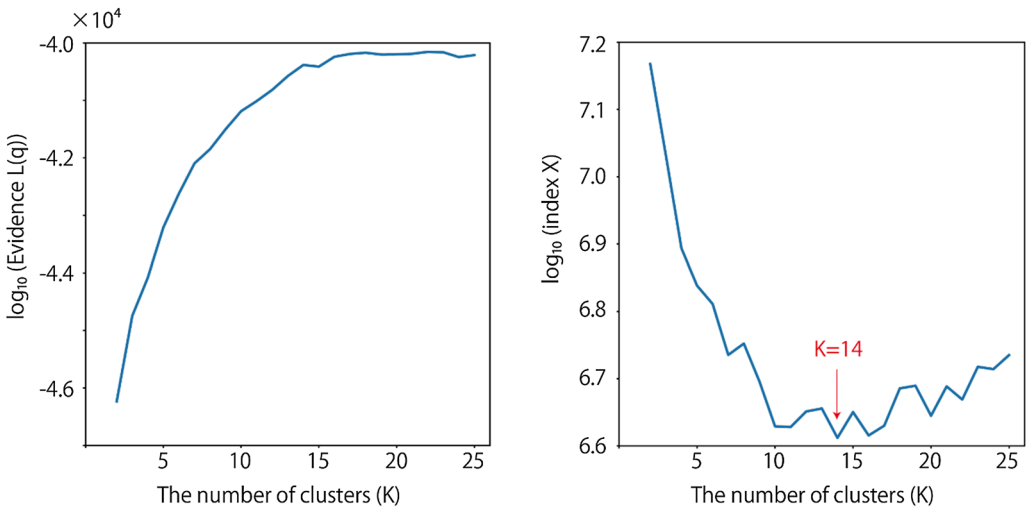


Fig. 6 Grid search results of Evidence *L*(*q*) and index *X* for the number of clusters (*K*) at Site ITA

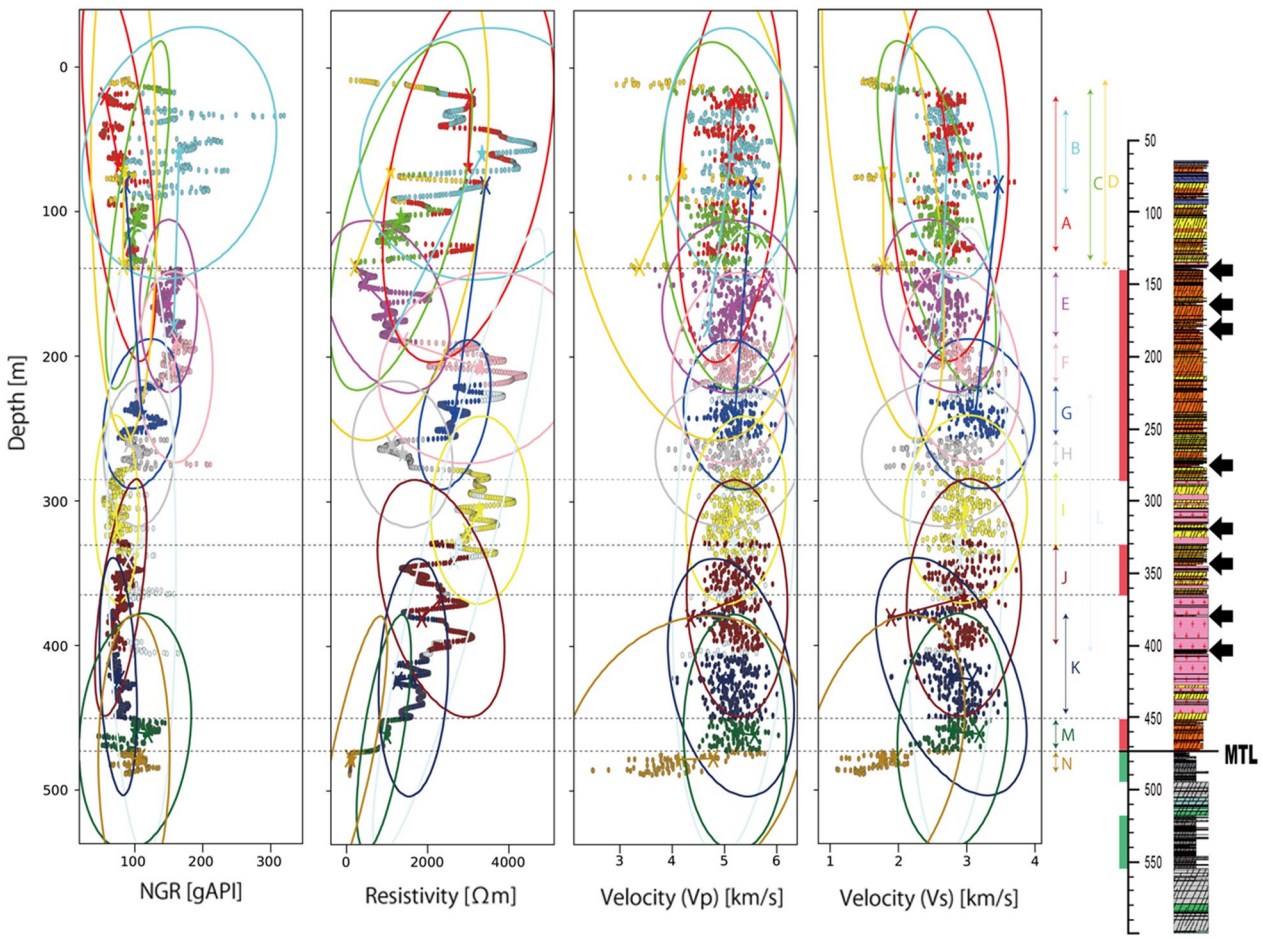


Fig. 7 Clustering results at ITA and core descriptions (Shigematsu et al. 2012). Ellipses represent covariance of each cluster

conducted in two steps. The borehole was first drilled with a diameter of 96 mm to acquire drill core samples. Resistivity and NGR logs were acquired for this initial borehole. Then, the borehole was widened to a diameter of 10–5/8 inches. P- and S-wave velocity logs were recorded for a widened borehole. We used four logs (electrical resistivity, NGR, and P- and S-wave velocity logs) for the analysis (Fig. 5). Although the borehole was drilled down to ~600 m below the land surface (mbls), complete sets of logging data with good quality were obtained until ~490 mbls owing to the unstable borehole wall below 473.9 mbls. The sampling interval of logging data was 10 cm.

Although log unit classification has not been published for logging data, lithologies and structures have been described for drill core samples (Shigematsu et al. 2012, 2014; Mori et al. 2015) (Fig. 5). The borehole ITA-1 penetrated the MTL at 473.9 mbls with significant lithological changes from the Ryoke granitoids to the Sanbagawa metamorphic rocks, resulting in the large drop in resistivity log. Below the MTL, the rocks were intensively fractured at 473.9–495.0 and 520–555 mbls. In the hanging wall of the MTL, three mylonite zones were recognized at 140–286, 330–370, and 450–473.9 mbls; in this study, we referred to them as the top, middle, and bottom mylonite zones, respectively. Shallow dipping faults were also identified at 135, 170, 260, 350, 380, and 410 mbls (Shigematsu et al. 2012).

Results We conducted a grid search for the number of clusters to obtain the optimized model. Figure 6 shows the results of the grid search. The approximated evidence value was high and flat at a large K range, whereas the index X showed a minimum at $K=14$. The sequence of the estimated hidden states is shown in Fig. 7. The clustering results are summarized in Table 2. Similar to the previous example, the boundaries of clusters were set at depths where step changes of logging values were observed or where depth dependencies of logging values were changed. At depths shallower than ~140 mbsf, four clusters (A–D) overlapped. Clusters E–M had simi-

Table 2 Clustering results for site ITA

Clusters	Depth (mbls)	Previously reported core description
A	20–128	Not described– ~65
B	30–88	
C	15–134	
D	9–139	
E	142–187	Mylonite zone 140–286
F	191–219	
G	221–255	
H	258–277	
I	281–334	
J	331–400	Mylonite zone 330–370
K	378–448	
L	226–405	
M	451–472	Mylonite zone 450–473.9
N	474–488	Footwall of MTL 473.9

lar velocities, although they had different sets of resistivity and NGR values. Cluster N, which appears at depths below 473.9 m, was characterized by clear drops in P- and S-wave velocity logs and resistivity logs.

We observed a good correlation between the clustering results and the three mylonite zones. Above the top mylonite zone, four clusters were identified (A–D). The top mylonite zone could be compared with four clusters (E–H). Cluster I covered the depth range between the top and middle mylonite zones. Cluster J could be compared to the middle mylonite zone. Cluster K covered the depth range between the middle and bottom mylonite zones. Two spikes of the NGR log at 360–410 mbls were categorized into cluster L. The bottom mylonite zone corresponded to cluster M. The footwall of the MTL corresponded to cluster N.

The shallowest part in cluster D was considered to represent the effect of the surface because the velocity logs showed steep increases. Clusters A, B, and C had similar P- and S-wave velocities but different NGR and resistivity values. Cluster A was characterized by low NGR log values, whereas cluster B had high NGR log values corresponding to mafic rocks that often appear in surrounding Ryoke granitoids (Hayama et al. 1982; Shigematsu et al. 2012). Cluster C had lower resistivity log values than clusters A and B.

The top mylonite zone corresponded to four clusters (E–H). The bottom cluster H was characterized by a low NGR. Clusters E and F had similar NGR, although they had low and high resistivities, respectively. Fault breccias of shallow dipping faults have been documented in the depth range corresponding to cluster E, including the fault located at the screened depth of ITA-2 (Matsumoto and Shigematsu 2018), which is interpreted to represent faults. Large differences in NGR logs for clusters F (191–219 mbls) and G (221–255 mbsf) implied that there are lithological differences between the two clusters. X-ray diffraction (XRD) analyses by Tanaka et al. (2012) suggest that the formation at 169–222 mbls is K-feldspar originating from granite, whereas the formation at 222–275 mbls is chlorite originating from tonalite.

Clusters J and K may reflect differences in the alteration minerals. XRD analyses by Tanaka et al. (2012) reported the presence of laumontite at 275–408 mbls, which is a hydrothermal alteration mineral, while it is not abundant below 408 mbls. The bottom of cluster J was defined at 400 mbls, which may correspond to such lithological differences. Such differences could be attributed to the differences in the CO_2 partial pressure of the fluid causing hydrothermal alternation. The fluid electrical conductivity log for drilling mud suggests that there is a fluid pathway at approximately 408 mbls (Kiguchi and Kuwahara 2020). Fault breccias have also been documented at 410 mbls (Shigematsu et al. 2012). Inflow of fluid at this depth may change the characteristics of the fluid responsible for hydrothermal alternation.

In this example, the clustering results showed a good correlation with the lithological description. This is because onshore geological formations were well lithified, and the depth dependence of elastic properties was small. For example, the P- and S-wave velocity logs were almost constant in Station ITA, except for the top of the hole. Therefore, lithological differences could be the main factor in clustering logging data. On the other hand, in the previous example, logging data showed strong depth dependency due to differences in consolidation state (i.e., physical properties), resulting in less contribution of lithological differences in clustering.

Coal bed at forearc basin of the Japan Trench

Geological settings and data with previous interpretations IODP Expedition 337 conducted scientific drilling at Site C0020 (Fig. 8) located in the forearc basin of the Japan Trench, where a deep coal bed was confirmed at ~2000 mbsf (Expedition 337 Scientists 2013a, 2013b). As this drilling used the riser drilling technique (Expedition 337 Scientists 2013a), cutting samples were acquired through 636.5–2466 mbsf. Coring was skipped for the top half of the hole and was conducted for 1256.5–2466 mbsf. Logging data were acquired using wire-line logging. Only the NGR log was acquired for the shallow hole. The full dataset is available below 1256.5 mbs.

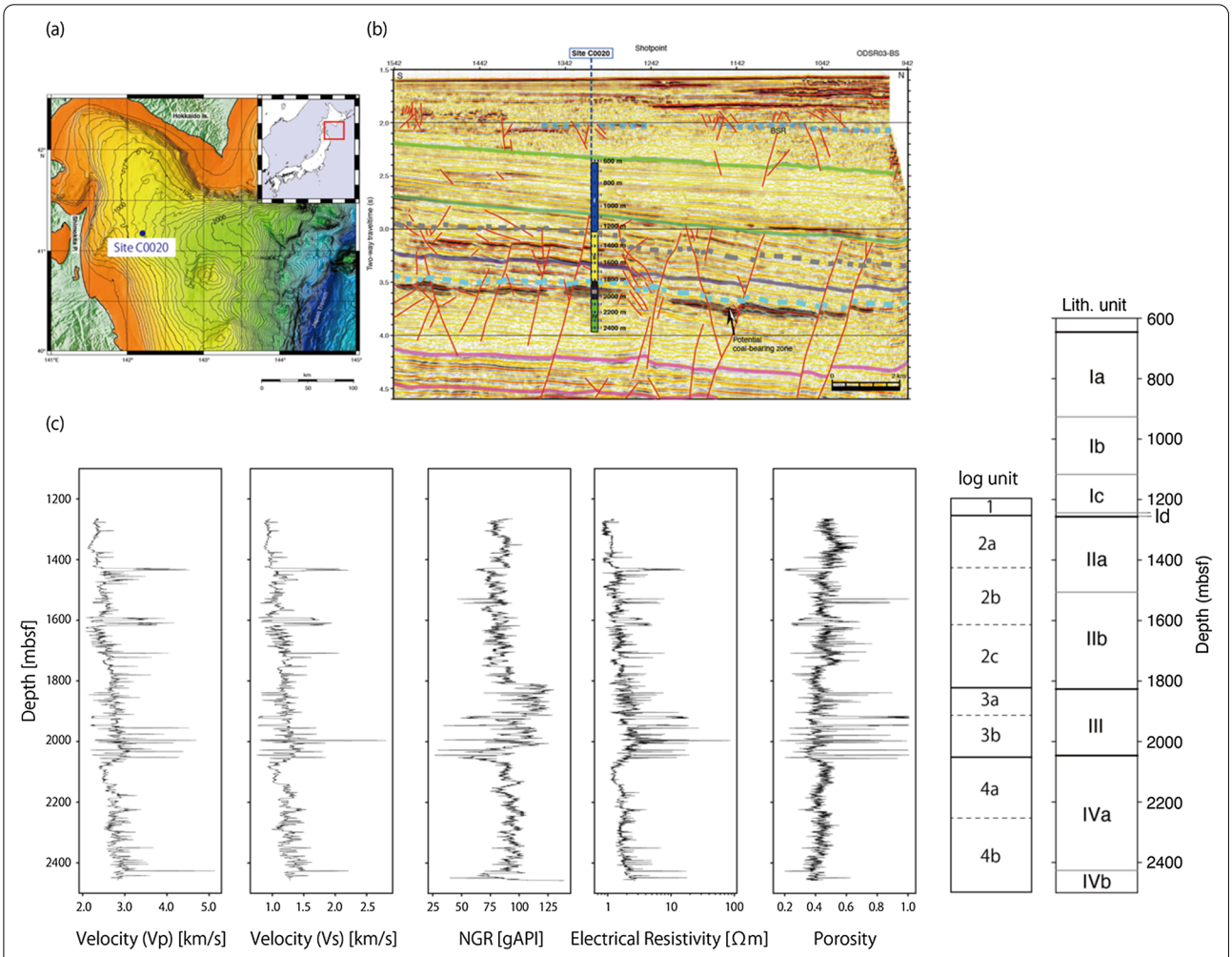


Fig. 8 Geological settings and data summary at Site C0020. **a** Regional map at the Japan Trench showing the location of Site C0020, after Expedition 337 Scientists (2013a). **b** Local seismic reflection image after Expedition 337 Scientists (2013a). **c** Logging data and previously reported log- and litho-units (Expedition 337 Scientists 2013b)

Logging data were classified into four units (Expedition 337 Scientists 2013b). Unit 1 was at 647–1256.5 mbsf, where only the NGR log was available. This unit was not included in our analysis because of the lack of complete datasets. Unit 2 was at 1256.5–1825.5 mbsf and was interpreted to have alternated relatively thick massive sandstones and siltstones. It was further divided into three subunits: unit 2a was located at 1256.5–1429.1 mbsf, which consisted of sandstone and siltstone of 60–70 m thickness; unit 2b at 1429.1–1617.4 mbsf was a highly permeable sandy sequence with several thick conglomerate layers at its base (1592.5–1617.4 mbsf); unit 2c at 1617.4–1825.5 mbsf gradually changed from silty layers at the top to sandy layers at the bottom. Unit 3 at 1825.5–2055.0 mbsf was characterized by frequent coal layers and was further divided into two subunits at 1916.2 mbsf. The shallow Unit 3a was silty with frequent sandy layers, whereas the deeper Unit 3b was sandier. Lastly, unit 4 at 2055.0–2466 mbsf was also divided into two subunits, with the shallow Unit 4a (2055.0–2255.0 mbsf) as thick massive shale and the deeper Unit 4b as an alternation of sandstone and shale.

According to a previous report (Expedition 337 Scientists 2013b), lithological units are also divided into four units, although subunit classification is different from log units. Unit I at 647–1256.5 mbsf primarily consists of diatom-bearing silty clay. Unit II at 1256.5–1826.5 mbsf mostly consists of silty shale. A rapid decrease in sand content characterized the boundary between Units I and II. Unit II was further divided into two subunits, with unit IIb having more abundant organic materials in shale than in unit IIa. The boundary between units IIa and IIb was documented at 1506.5 mbsf. Unit III at 1826.5–2046.5 mbsf is characterized by coal layers. The thick coal

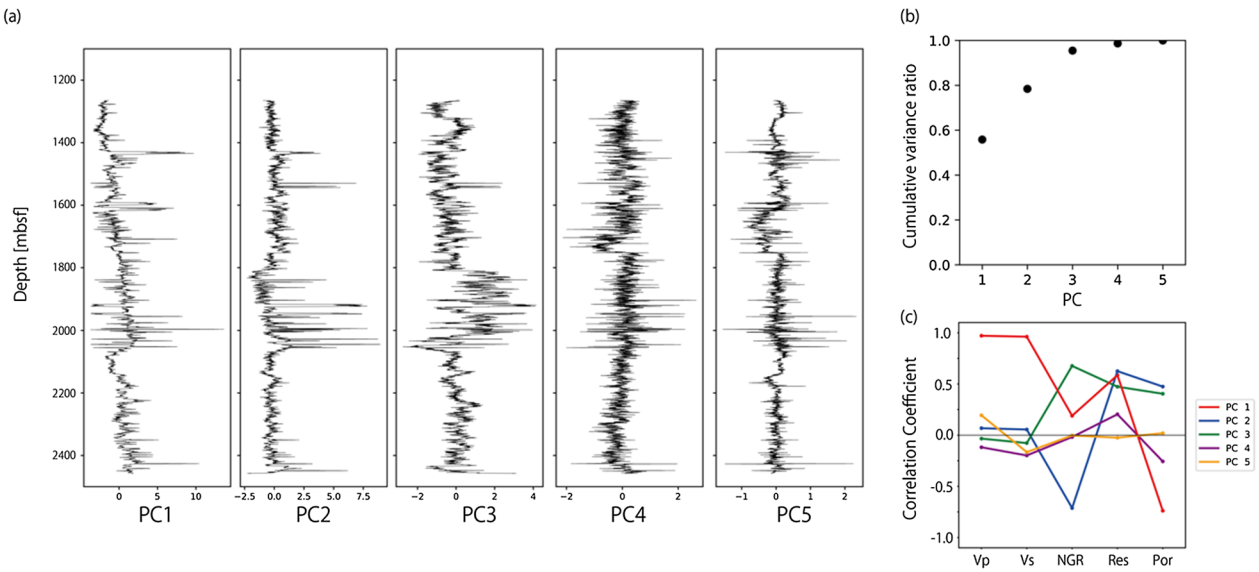


Fig. 9 Principal component analysis results at Site C0020. **a** Depth profiles of logging data converted into principal component domain. **b** Cumulative variance ratio of principal components. **c** Loadings of principal components

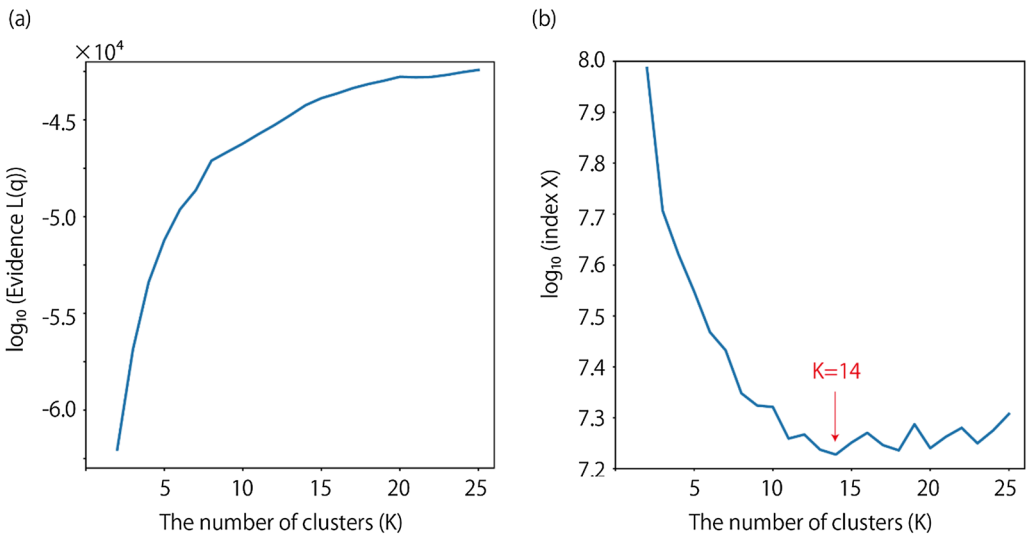


Fig. 10 Grid search results of Evidence $L(q)$ and index X for the number of clusters (K) at Site C0020

layers disappeared in Unit IV at 2046.5–2466 mbsf. Unit IV was further divided into two subunits (IVa and IVb). Unit IVa mainly comprises shale and sandstone, whereas Unit IVb comprises clay and silt. The boundary between the two subunits was documented at 2426.5 mbsf.

Here, we used five types of logging data (electrical resistivity, NGR, P- and S-wave velocity, and porosity) for HMM clustering (Fig. 8). The sampling interval for NGR and porosity log was 2.5 cm; resistivity log, 5 cm; and P- and S-wave velocity, 15 cm. To align the logging data, we set the depth grid at every 20 cm and calculated the average values of the logging data within 20 cm from each grid. As the data dimension increased, the average density in the data space decreased, and the calculation cost increased. However, data are expected to be distributed in some subspaces in the data space because logging data are constrained by the physical properties of geological formations. Hence, dimensionality reduction helped

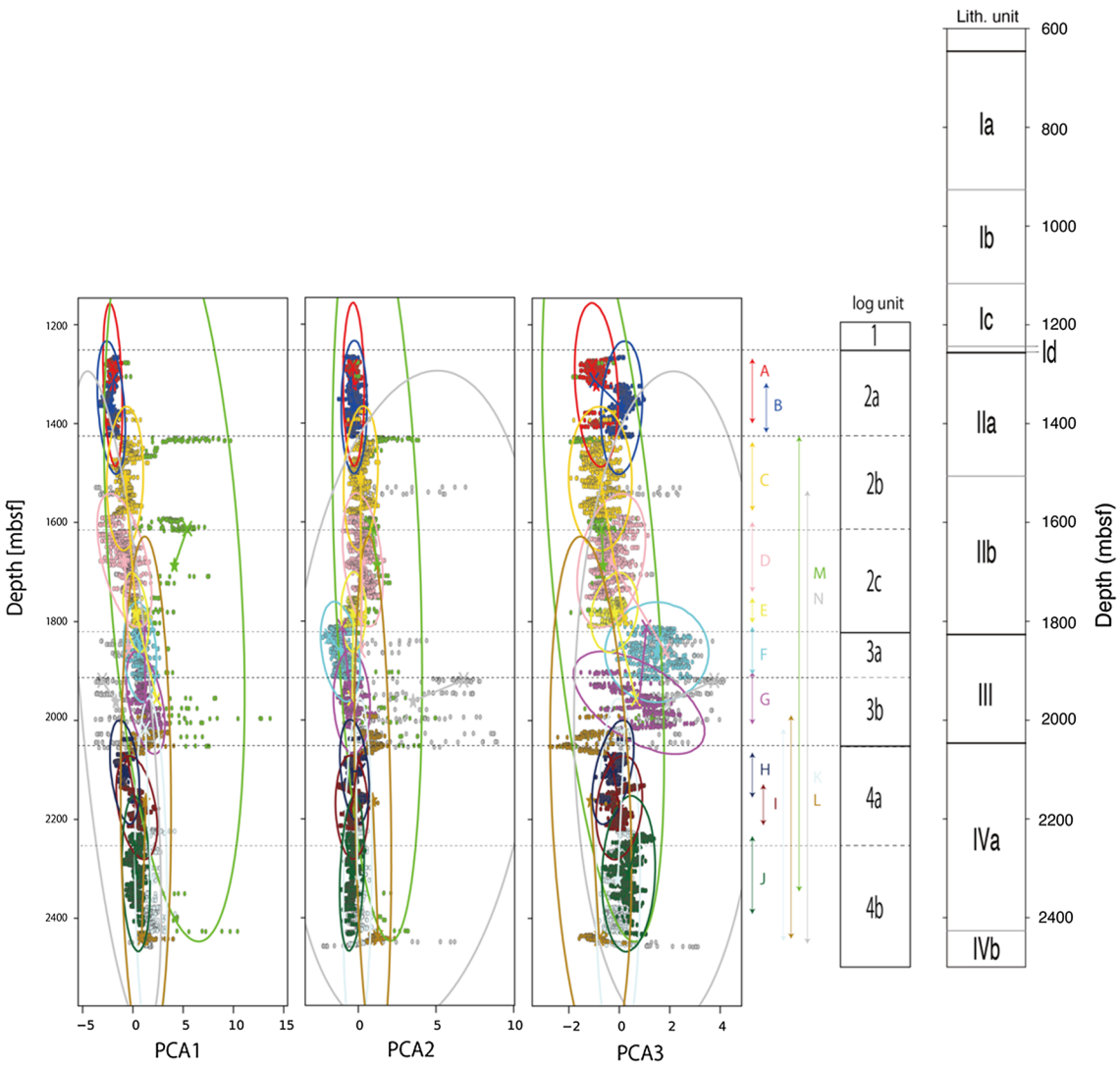


Fig. 11 Clustering results at C0020 in principal component domain and previously reported log units (Expedition 337 Scientists 2013b). Ellipses represent covariance of each cluster

in reducing the calculation cost and in conducting clustering more efficiently. We used the principal component analysis to reduce dimensionality (Fig. 9). As the resistivity varied significantly in this example, we converted the resistivity log to a logarithmic scale. The cumulative variance ratio for the first to third principal components exceeded 0.95. We then conducted HMM clustering with the first to third principal components.

Results We conducted a grid search for the number of clusters to obtain the optimized model. Figure 10 shows the results of the grid search. The approximated evidence value continued to increase at a large K range, whereas the index X showed the minimum at $K = 14$. The sequence of the estimated hidden states in the principal component domain and the original data domain are shown in Figs. 11 and 12, respectively. The clustering results are summarized in Table 3. Similar to the previous examples, the boundaries of clusters were set at depths where step changes of logging values are observed or where depth dependencies of logging values are changed. In addition, spiky data points were assigned to different clusters in this example. Clusters A and B were separated by a step increase in resistivity and NGR logs. Although several data points were classified as cluster A at ~1400 mbsf, the main body of cluster A was located above cluster B. Clusters B and C were separated by a step decrease in NGR log and step increase in P- and S-wave velocity logs.

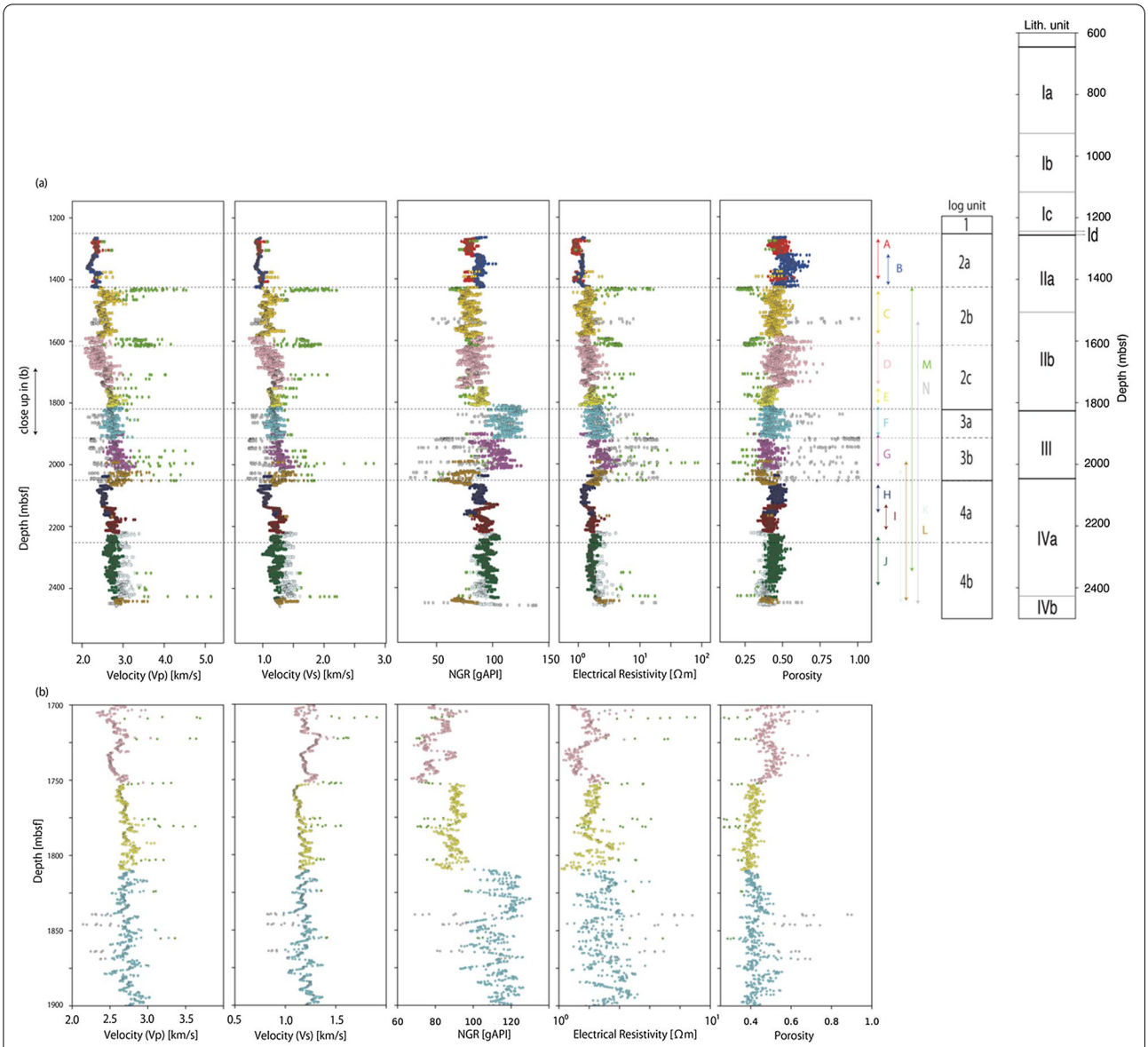


Fig. 12 Clustering results at C0020 in original logging data domain and previously reported log units (Expedition 337 Scientists 2013b). **a** The entire section, and **b** the close-up figure of **a** where noted by a black arrow on the left

Clusters C and D were separated by a step decrease in P- and S-wave velocity logs. The boundary between clusters D and E was marked by a step decrease in the porosity logs. Cluster F had the highest NGR log values among all clusters. Characterized by the lowest NGR log values among all clusters, Cluster L was composed of two parts: one was sandwiched between clusters G and H, and the other was at the bottom of the hole. Clusters H and I were differentiated using P- and S-wave velocity logs. The boundary between clusters I and J was characterized by a step increase in the electrical resistivity and porosity logs. Cluster K corresponded to spiky data points with high P- and S-wave velocities and electrical resistivity logs and low NGR and porosity logs. Cluster M had characteristics similar to cluster K, although it was more dominant at shallow depths. Cluster N corresponded to spiky data points with high electrical resistivity and porosity logs and low P- and S-wave velocities and NGR logs.

We observed a good correlation between clustering results and log units described by the expedition report (Expedition 337 Scientists 2013b), whereas subunits of lithological units did not correlate well with clustering results. Unit 2a was divided into two clusters (A and B). Unit 2b was mainly composed of cluster C. The

Table 3 Clustering results for Site C0020

Clusters	Depth (mbsf)	Previously reported log units	Previously reported lithological units
–	–	Unit 1 647–1256.5	Unit I 647–1256.5
A	1272.2–1404.0	Unit 2 1256.5–1825.5	Unit II 1256.5–1826.5
B	1321.6–1422.8		
C	1440.8–1581.0		
D	1602.6–1744.6		
E	1755.6–1807.0		
F	1814.4–1911.4	Unit 3 1825.5–2055.0	Unit III 1826.5–2046.5
G	1905.2–2012.2		
H	2068.6–2160.4	Unit 4 2055.0–2466.0	Unit IV 2046.5–2466
I	2132.6–2216.2		
J	2237.6–2395.8.6		
K	2020.6–2450.4		
L	1993.0–2445.8		
M	1429.4–2350.4	Units 2–4	Units 2–4
N	1539.6–2456.0		

Table 4 Documented coal layers at Site C0020

Depth of reported coal layers thicker than 30 cm (mbsf)	Log unit	Cluster N
1529.5–1530.1	Unit 2b	1529.0–1530.2
1539.4–1539.7		1539.0–1539.8
1543.4–1544.0		1542.2–1544.2
		1704.0–1704.4
1839.1–1840.0	Unit 3a	1839.0–1840.2
1846.3–1846.9		1845.4–1846.6
1863.7–1864.6		1863.4–1864.4
		1868.6–1869.0
1916.2–1923.5	Unit 3b	1916.2–1923.8
1944.4–1947.9		1944.2–1949.0
1958.4–1958.7		1958.2–1958.8
1978.8–1979.3		1978.6–1979.4
1993.7–1994.8		1993.4–1994.6
1997.5–1998.8		1997.2–1998.8
		2001.0–2001.2
2002.3–2003.0		2002.2–2003.2
2027.0–2028.1		2026.8–2028.4
2043.9–2045.3		2043.4–2046.2
2054.7–2055.0	Unit 4b	2054.6–2055.6
2448.4–2449.3		2448.2–2449.6
		2455.0–2458.0

bottom depth of cluster C was consistent with the top depth of the thick conglomerate layer in unit 2b. The bottom of unit 2b, where the thick conglomerate layer was observed, and Unit 2c was composed of two clusters (clusters D and E). Units 3a and 3b corresponded to clusters F and G, respectively. Unit 4a was composed of two clusters (clusters M and N). Unit 4b was mainly composed of cluster J with spiky data points classified as cluster K.

Unit 2b was documented as a thick sandy layer with at least three coal layers and frequent cemented resistive zones (Expedition 337 Scientists 2013b). Hence, cluster C corresponded to a thick sandy layer. Cluster M showed high resistivity with low porosity and high velocity. These characteristics of cluster M were consistent with those of cemented sandstone layers (Expedition 337 Scientists 2013b).

Unit 3 was the main target of the drilling site. A total of 13 coal layers thicker than 30 cm and a number of thin coal layers were observed in Unit 3. Depths of coal layers documented in the expedition report (Expedition 337 Scientists 2013b) showed an excellent match with the data points in Cluster N (Table 4). Hence, cluster N corresponded to coal layers. This was also the case for Units 2b and 4b. The characteristics of cluster N (high resistivity and low NGR) were consistent with those of the coal layers in logging data (Expedition 337 Scientists 2013b). Cluster F was a major component of unit 3a, whereas cluster G was a major component of 3b. An expedition report (Expedition 337 Scientists 2013b) described that Unit 3a was silty with sand interbeds, whereas Unit 3b was sandier. Hence, clusters F and G represented silty and sandy layers, respectively.

The clustering results in this example showed less correlation with lithological descriptions, as in the first example. This could be attributed to the slight depth dependency of the datasets, which was also observed in the first example. However, our method could classify lithological characteristics observed as peaky data points, such as coal bed layers and cemented layers, into individual clusters (clusters M and N, respectively). These results demonstrated that the proposed method is applicable for the detection and classification of irregular geological formations, such as layers, including those of natural resources.

Discussion

Some clusters defined by HMM overlapped in correlation plots (Additional file 1: Figures S1–S4); however, they were well separated in the depth plots (Figs. 4, 7, and 12). As depth trend of logging data is an important feature when log units are defined visually, capturing such features in statistical models should be crucial to conduct good clustering. In the developed method, we include depth information in observables to explicitly represent depth trend of logging data.

We searched for the optimum number of clusters by grid search for K, assuming five other hyperparameters. Among the assumed hyperparameters, α represented uniform distribution of the prior probability. m_0 and w_0 were determined based on the input data. However, β_0 and υ_0 were assumed empirically. Additional file 1: Figures S5–S7 show the clustering results with different β_0 and υ_0 and with the same K of the optimum model. The clustering results were not strongly dependent on these two hyperparameters. Boundaries of clusters, which correspond to previously reported log-unit boundaries, are usually consistent among clustering results with different hyperparameters. On the other hand, some clusters, which are smaller than subunits previously reported, could vary according to the assumed hyperparameters. With increasing β_0, covariance ellipses tend to become larger because larger β_0 put more emphasis on prior distribution, which is assumed to be covariance matrix calculated for the entire input data. Based on the results of the three cases in this study, we consider that $\beta_0 = 10$ may be too large, and we select $\beta_0 = 1$ as a reference parameter. On the other hand, υ_0 may influence clustering results less significantly than β_0, at least in the parameter range studied in Additional file 1: Figures S5–S7. As the reference parameter set $(\beta_0, \upsilon_0) = (1, D)$ works good in datasets from completely different geological settings, it is also expected to be good in other datasets. However, future analysts can adjust these hyperparameters for their own datasets.

At Site C0020, logging data were preprocessed for alignment because different logs were acquired at different depth. To investigate the effects of preprocessing on the clustering results, we conducted clustering for logging data collected at different depth intervals (40 cm instead of 20 cm in the original dataset) using the reference set of hyperparameters. According to the grid search for K with index X, the optimum number of clusters was defined to be 18. The clustering results (Additional file 1: Figure S8) showed that the boundaries of clusters, which correspond to log-unit boundaries previously reported, are consistent between the results of the two depth intervals. Therefore, preprocessing of logging data did not significantly affect clustering results.

Clustering results are not necessarily consistent with lithological (or geological) classifications, especially when logging data show systematic changes with depth, as observed in Sites C0001 and C0020. This was because we mainly used data that are more sensitive to physical properties (such as resistivity and velocity logs) rather than lithology (such as mineral composition data). When logging data do not show a systematic trend with depth, as observed in Site ITA, lithological differences should be more considered in HMM clustering by focusing on the differences in the NGR log, which is more sensitive to lithology.

The clustering results obtained by the proposed method were usually consistent with previously reported log unit classifications. In addition, HMM clustering sometimes divided logging data in a more detailed manner than that of previous interpretations. In Sites C0001 and C0020, where logging units were reported by expedition reports, some log subunits corresponded to single clusters classified by HMM clustering (log units 1b, 2b, 2c, 3a, and 3c at Site C0001 and log units 2b, 3a, 3b, and 4b at Site C0020), whereas some log subunits were composed of several clusters by HMM clustering (log units 1a, 2a, and 3b at Site C0001 and log units 2a, 2c, and 4a at Site C0020). Such differences between the results of subjective clustering and our quantitative clustering offers a good opportunity to discuss interpretations of the geological meaning of data. These differences might imply that the resolutions of subjective clustering based on previous interpretations were not consistent for the entire depth range. Otherwise, our quantitative clustering method might emphasize the differences on acoustic properties within a single geological unit. Quantitative clustering results using our proposed method assist scientists in making geological interpretations and determining unit classifications efficiently.

The detailed division of logging data by the proposed method provides us new insights on geological interpretations of the logging data. At Site C0001, log unit 3b was divided into four clusters. Of the four clusters, Cluster L is characterized by a clear drop in electrical resistivity and P-wave velocity (Fig. 4b). This cluster was stably detected in results with any sets of hyperparameters tested in Additional file 1: Figure S5. The characteristics of Cluster L are similar to those of Cluster H (log unit 3a), which is interpreted as the fault zone or the mass transport deposit (Expedition 314 Scientists 2009b). According to the expedition report (Expedition 314 Scientists 2009b), the hole was intersecting faults and fractures at depths of 800, 835, and 860 mbsf, which correspond to the depth range of Cluster L. Therefore, it is highly possible that Cluster L corresponds to the fault zone in log unit 3b. At Site C0020, log unit 2c was divided into two clusters (Clusters D and E), the boundary of which at ~1750 mbsf is characterized by a clear drop in porosity and S-wave velocity and a clear increase in NGR log. This boundary was stably detected in results with any sets of hyperparameters tested in Additional file 1: Figure S7, except for the one set of $(\beta_0, \upsilon_0) = (0.1, 10)$. According to the site report of Site C0020 (Expedition 337 Scientists 2013a), a minor fault intersects with the hole at approximately 1700–1750 mbsf based on the reflection image in this region (red lines in Figure F2 of Expedition 337 Scientists

2013a). The clear boundary at ~1750 mbsf documented in our clustering results may correspond to the fault in the reflection image. As demonstrated by these examples, clustering results using our proposed model is useful to interpret geological structures from logging data.

The proposed method is expected to have a wide applicability and extensibility. In this study, it provided reasonable results at offshore accretionary prism, onshore fault zone, and offshore buried coal layers. The last example showed that spiky data points related to natural resources are efficiently extracted from other data points at the baseline, which helped specify layers with natural resources. As tephra layers in marine sediments were also identified using spiky data points in logging data (Mahony et al. 2016), this method could also be applied for the detection of tephra layers in marine sediments to study large magnitude explosive volcanism. Although this study used only basic logging data (electrical resistivity, NGR, porosity, and velocities), other types of data can be used as inputs. For example, clustering the orientations of bedding, fractures, and faults identified from resistivity image logs or borehole televiewers may be useful for structural geological interpretations. When using such periodical data, von Mises distributions should be used as the data generation probability instead of the normal distribution assumed in this study. Geochemical data from drill core samples are also a candidate. Although the proposed method cannot be used directly for such types of data not constantly acquired in depth, minor modifications (such as marginalization of probability at data-missing depth) could be used to treat them. In addition, the statistical framework for logging data we proposed can serve as the basis of quantitative core-log-seismic integration for the efficient interpretation of subsurface structures.

This study proposed a statistical method using unsupervised HMM for quantitative logging data clustering. Our proposed method assists our interpretations of logging data, providing a quantitative basis and validity to define log unit boundaries, which has been thus far done subjectively. We applied our model on three different geological settings and showed that clustering results agree with the log unit classification previously conducted using manual inspection. The proposed statistical model is expected to have wide applicability and extensibility to incorporate various types of data other than those used in this study.

Abbreviations

IODP: International Ocean Discovery Program; NGR: Natural gamma ray; HMM: Hidden Markov model; MTL: Median tectonic line; XRD: X-ray diffraction; mbsf: meter below the sea floor; mbls: meter below the land surface.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40623-022-01651-0.

Additional file 1: Figure S1. Clustering results for Site C0001. Three panels represent correlation plots among observable parameters other than depth. Ellipses represent covariance of each cluster. **Figure S2.** Clustering results for Site ITA. Six panels represent correlation plots among observable parameters other than depth. Ellipses represent covariance of each cluster. **Figure S3.** Clustering results for Site C0020. Three panels represent correlation plots among observable parameters in principal component domain other than depth. Ellipses represent covariance of each cluster. **Figure S4.** Clustering results for Site C0020. Three panels represent correlation plots among observable parameters in original logging data domain other than depth. **Figure S5.** Depth plot of electrical resistivity and clustering results for Site C0001 with different hyper parameters ($\boldsymbol{\beta}_0, \boldsymbol{\upsilon}_0$) at $K=13$. Ellipses represent covariance of each cluster. Log units previously reported (Expedition 314 Scientists 2009b) are shown on the right and dotted black lines. The original set of hyper parameters shown in the main text (Fig. 4) is the left panel at the central raw. **Figure S6.** Depth plot of P-wave velocity and clustering results for Site ITA with different hyper parameters ($\boldsymbol{\beta}_0, \boldsymbol{\upsilon}_0$) at $K=14$. Ellipses represent covariance of each cluster. Core descriptions previously reported (Expedition 314 Scientists 2009b) are shown on the right and dotted black lines. The original set of hyper parameters shown in the main text (Fig. 7) is the left panel at the central raw. **Figure S7.** Depth plot of porosity and clustering results for Site C0020 with different hyper parameters ($\boldsymbol{\beta}_0, \boldsymbol{\upsilon}_0$) at $K=14$. Log units previously reported (Expedition 337 Scientists 2013b) are shown on the right and dotted black lines. The original set of hyper parameters shown in the main text (Fig. 12) is the left panel at the central raw. **Figure S8.** Depth plot of porosity and clustering results for Site C0020 with different depth intervals. The left panel shows the results at a depth interval of 20 cm and $K=14$, which is identical to the figure shown in Fig. 12. The right panel shows the results at a depth interval of 40 cm and $K=18$. Log units previously reported (Expedition 337 Scientists 2013b) are shown on the right and dotted black lines.

Acknowledgements

Some figures were produced using the GMT software package (Wessel et al. 2013). We appreciate two reviewers for constructive comments to improve our manuscript.

Author contributions

SY, YH, RF, and KU contributed toward the conception of this study. SY wrote python codes for analysis and SN helped constructing methodology. NS and TK curated data and helped geological interpretations for Site ITA. YH, RF, and KU helped geological interpretations for Site C0001 and C0020. SY analyzed data and drafted the manuscript. All authors read and approved the final manuscript.

Funding

This research was supported by the management expenses of the Japan Agency for Marine-Earth Science and Technology and the National Institute of Advanced Industrial Science and Technology.

Availability of data and materials

The logging data at sites C0001 and C0020 are available at http://sio7.jamstec.go.jp and the logging data at ITA are available at https://doi.org/10.5281/zenodo.5546788. The Python code used is at https://doi.org/10.5281/zenodo.5546841.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology, Tsukuba Central 7, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8567, Japan. [2]Institute for Extra-Cutting-Edge Science and Technology Avant-Garde Research, Kochi Institute for Core Sample Research, Japan Agency for Marine-Earth Science and Technology, 200 Monobe Otsu, Nankoku-shi, Kochi 783-8502, Japan. [3]Naruto University of Education, 748, Nakajima, Takashima, Naruto-cho, Naruto-shi 772-8502, Japan. [4]Graduate School of Accountancy, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan. [5]Research Institute for Marine Geodynamics, Japan Agency for Marine-Earth Science and Technology, 2-15, Natsushima-cho, Yokosuka, Kanagawa 237-0061, Japan.

Received: 18 November 2021 Accepted: 24 May 2022
Published online: 13 June 2022

References

Ando M (1975) Source mechanisms and tectonic significance of historical earthquakes along the nankai trough, Japan. Tectonophysics 27:119–140
Arthur D, Vassilvitskii S (2007) k-means++: the advantages of careful seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, Philadelphia, pp 1027–1035
Baldi P, Brunak S, Bach F (2001) Bioinformatics: the machine learning approach. MIT Press, Cambridge
Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8
Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
Cerchiari A, Fukuchi R, Gao B, Hsiung KH, Jaeger D, Kaneki S et al (2018) IODP workshop IODP workshop: Core-Log Seismic Investigation at Sea—integrating legacy data to address outstanding research questions in the Nankai Trough Seismogenic Zone Experiment. Sci Drill 24:93–107
DeMets C, Gordon RG, Argus DF (2010) Geologically current plate motions. Geophys J Int 181:1–80
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39:1–22
Eidsvik J, Avseth P, Omre H, Mukerji T, Mavko G (2004) Stochastic reservoir characterization using prestack seismic data. Geophysics 69:978–993
Expedition 314 methods (2009) Proceedings of the IODP. In: Kinoshita M, Tobin H, Ashi J, Kimura G, Lallemant S, Screaton EJ, Curewitz D, Masago H, Moe KT, The Expedition 314/315/316 Scientists (eds) Proceedings the Expedition of the Integrated Ocean Drilling Program. Integrated Ocean Drilling Program Management International Inc., Washington
Expedition 314 Site C0001 (2009) 314, Site C. In: Kinoshita M, Tobin H, Ashi J, Kimura G, Lallement S, Screaton EJ, Curewitz D, Masago H, Moe KT, The Expedition 314/315/316 Scientists (eds) Proceedings of the IODP Proceedings of the Integrated Ocean Drilling Program. Integrated Ocean Drilling Program Management International Inc., Washington
Expedition 315 Site C0001 (2009) 315, Site C. In: Kinoshita M, Tobin H, Ashi J, Kimura G, Lallemant S, Screaton EJ, Curewitz D, Masago H, Moe KT, The Expedition 314/315/316 Scientists (eds) Proceedings of the IODP Proceedings of the Integrated Ocean Drilling Program. Integrated Ocean Drilling Program Management International Inc., Washington
Feng R, Luthi SM, Gisolf D, Angerer E (2018) Reservoir lithology determination by hidden markov random fields based on a Gaussian mixture model. IEEE Trans Geosci Remote Sens 56:6663–6673
Hammer H, Kolbjørnsen O, Tjelmeland H, Buland A (2012) Lithology and fluid prediction from prestack seismic data using a Bayesian model with Markov process prior. Geophys Prospect 60:500–515
Harding T (1974) Petroleum traps associated with wrench faults. AAPG Bull 58:1290–1304
Hayama Y, Yamada T, Ito M, Kutsukake T, Masaoka K, Miyakawa K et al (1982) Geology of the Ryoke Belt in the eastern Kinki District, Japan: the phase-divisions and the mutual relations of the granitic rocks. J Geol Soc Jpn 88:451–466 (**in Japanese with English abstract**)
Inagaki F, Hinrichs KU, Kubo Y, The Expedition 337 Scientists (2013) Expedition 337 summary. Proceedings of the Integrated Ocean Drilling Program, 337. Integrated Ocean Drilling Program Management International, Inc., Tokyo
Itaba S, Koizumi N, Matsumoto N, Ohtani R (2010) Continuous observation of groundwater and crustal deformation for forecasting Tonankai and Nankai earthquakes in Japan. Pure Appl Geophys 167:1105–1114
Jaakkola T (2001) Tutorial on variational approximation methods. In: Saad D, Opper M (eds) Advanced mean field methods: theory and practice. MIT Press, Cambridge
Jelinek F (1997) Statistical methods for speech recognition. MIT Press, Cambridge
Jeong J, Park E, Han WS, Kim KY (2014) A novel data assimilation methodology for predicting lithology based on sequence labeling algorithms. J Geophys Res 119:7503–7520
Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1998) An introduction to variational methods for graphical models. Learning in graphical models. Springer, Dordrecht
Katori T, Shigematsu N, Kameda J, Miyakawa A, Matsumura R (2021) 3D fault-zone architecture across the brittle–plastic transition along the Median Tectonic Line, SW Japan: fault-rock characterization. J Struct Geol 153:104446
Kiguchi T, Kuwahara Y (2020) Controlling factors of orientations of sub-surface permeable fractures: borehole data analyses of 16 AIST observation stations in Aichi, Kii Peninsula and Shikoku regions, southwestern Japan. Ann Rep Active Fault Paleoearthquake Res 20:1–78
Kiguchi T, Kuwahara Y, Koizumi N, Tsukamoto H, Sato T, Satoh T et al (2014) Geophysical loggings at AIST integrated groundwater observation stations for forecasting megathrust earthquakes in the Nankai Trough. GSJ Openfile report, no. 598. Geological Survey, Japan
Kimura G, Kitamura Y, Hashimoto Y, Yamaguchi A, Shibata T, Ujiie K, Okamoto S (2007) Transition of accretionary wedge structures around the up-dip limit of the seismogenic subduction zone. Earth Planet Sci Lett 255:471–484
Kimura G, Koge H, Tsuji T (2018) Punctuated growth of an accretionary prism and the onset of a seismogenic megathrust in the Nankai Trough. Prog Earth Planet Sci 5:78
Larsen AL, Ulvmoen M, Omre H, Buland A (2006) Bayesian lithology/fluid prediction and simulation on the basis of a Markov-chain prior model. Geophysics 71:R69–R78
Lindberg DV, Omre H (2014) Blind categorical deconvolution in two-level hidden Markov models. IEEE Trans Geosci Remote Sens 52:7435–7447
Lockner DA, Morrow C, Moore D, Hickman S (2011) Low strength of deep San Andreas fault gouge from SAFOD core. Nature 472:82–85
Mahony SH, Sparks RSJ, Wallace LM, Engwell SL, Scourse EM, Barnard NH et al (2016) Increased rates of large-magnitude explosive eruptions in Japan in the Late Neogene and Quaternary. Geochem Geophys Geosyst 17:2467–2479
Matsumoto N, Shigematsu N (2018) In-situ permeability of fault zones estimated by hydraulic tests and continuous groundwater-pressure observations. Earth Planets Space 70:13
McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New Jersey
Miyazaki S, Heki K (2001) Crustal velocity field of southwest Japan: subduction and arc–arc collision. J Geophys Res 106:4305–4326
Moore JC, Saffer D (2001) Updip limit of the seismogenic zone beneath the accretionary prism of southwest Japan: an effect of diagenetic to low-grade metamorphic processes and increasing effective stress. Geology 29:183–186
Moore GF, Shipley TH, Stoffa PL, Karig DE, Taira A, Kuramoto S et al (1990) Structure of the Nankai Trough Accretionary Zone from multichannel seismic reflection data. J Geophys Res 95:8753–8765
Mori H, Wallis S, Fujimoto K, Shigematsu N (2015) Recognition of shear heating on a long-lived major fault using Raman carbonaceous material thermometry: implications for strength and displacement history of the MTL, SW Japan. Island Arc 24:425–446
Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286

Saffer DM (2007) Pore pressure within underthrust sediment in subduction zones. In: Moore JC (ed) The seismogenic zone of subduction thrust faults Dixon TH. Columbia University Press, New York
Schumann A (2002) Hidden Markov models for lithological well log classification. Terra Nostra 4:373–378
Shigematsu N, Fujimoto K, Tanaka N, Furuya N, Mori H, Wallis S (2012) Internal structure of the Median Tectonic Line fault zone, SW Japan, revealed by borehole analysis. Tectonophysics 532–535:103–118
Shigematsu N, Otsubo M, Fujimoto K, Tanaka N (2014) Orienting drill core using borehole-wall image correlation analysis. J Struct Geol 67:293–299
Shigematsu N, Kametaka M, Inada N, Miyawaki M, Miyakawa A, Kameda J et al (2017) Evolution of the Median Tectonic Line fault zone, SW Japan, during exhumation. Tectonophysics 696–697:52–69
Site C0020 (2013) Expedition 337 scientists. In: Inagaki F, Hinrichs KU, Kubo Y, The Expedition 337 Scientists (eds) Proceedings of the Integrated Ocean Drilling Program, 337. Integrated Ocean Drilling Program Management International Inc., Tokyo
Strasser M, Moore GF, Kimura G, Kitamura Y, Kopf AJ, Lallemant S et al (2009) Origin and evolution of a splay fault in the Nankai accretionary wedge. Nat Geosci 2:648–652
Sutherland R, Townend J, Toy V, Upton P, Coussens J, Allen M et al (2017) Extreme hydrothermal conditions at an active plate-bounding fault. Nature 546:137–140
Tanaka H, Fujimoto K, Ohtani T, Ito H (2001) Structural and chemical characterization of shear zones in the freshly activated Nojima fault, Awaji island, southwest Japan. J Geophys Res 106:8789–8810
Tanaka N, Furuya N, Fujimoto K, Shigematsu N (2012) X-ray diffraction analysis on the samples of the borehole core penetrating the median tectonic line, the eastern Kii Peninsula, SW Japan. Bull Tokyo Gakugei Univ Div Nat Sci 64:77–128
Tian M, Omre H, Xu H (2021) Inversion of well logs into lithology classes accounting for spatial dependencies by using hidden markov models and recurrent neural networks. J Petrol Sci Eng 196:107598
Tobin H, Hirose T, Ikari M, Kanagawa K, Kimura G, Kinoshita M et al (2020) Expedition 358 summary. In: Tobin H, Hirose T, Ikari M, Kanagawa K, Kimura G, Kinoshita M, Kitajima H, Saffer D, Yamaguchi A, Eguchi N, Maeda L, Toczko S, Kanamatsu T, The Expedition 358 Scientists (eds) NanTroSEIZE Plate Boundary Deep Riser 4: Seismogenic Nankai Slow slip megathrust Proceedings the International Ocean Discovery Program, 358. International Ocean Discovery Program, College Station
Townend J, Sutherland R, Toy VG, Eccles JD, Boulton C, Cox SC, McNamara D (2013) Late-interseismic state of a continental plate-bounding fault: petrophysical results from DFDP-1 wireline logging and core analysis, Alpine Fault, New Zealand. Geochem Geophys Geosyst 14:3801–3820
Townend J, Sutherland R, Toy VG, Doan M-L, Célérier B, Massiot C et al (2017) Petrophysical, geochemical, and hydrological evidence for extensive fracture-mediated fluid and heat transport in the Alpine Fault's hanging-wall damage zone. Geochem Geophys Geosyst 18:4709–4732
Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 13:260–269
Wallis SR, Okudaira T (2016) Paired metamorphic belts of SW Japan: the geology of the Sanbagawa and Ryoke metamorphic belts and the Median Tectonic Line. In: Moreno T, Wallis S, Kojima T, Gibbons W (eds) The geology of Japan. Geological Society of London, London
Wessel P, Smith WHF, Scharroo R, Luis J, Wobbe F (2013) Generic mapping tools: improved version released. EOS Trans Am Geophys Union 94:409–410

Publisher's Note