Front. Earth Sci. 2019, 13(3): 628–640
https://doi.org/10.1007/s11707-019-0748-x

RESEARCH ARTICLE

An unsupervised learning approach to study synchroneity of past events in the South China Sea

Kevin C. Tse (✉)[1], Hon-Chim Chiu[2], Man-Yin Tsang[3], Yiliang Li[1], Edmund Y. Lam[4]

1 Department of Earth Sciences, The University of Hong Kong, Pokfulam, Hong Kong, China
2 Department of Geography and Centre for Geo-computation Studies, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China
3 Department of Earth Sciences, University of Toronto, Canada
4 Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong, China

Abstract Unsupervised machine learning methods were applied on multivariate geophysical and geochemical datasets of ocean floor sediment cores collected from the South China Sea. The well-preserved and continuous core samples comprising high resolution Cenozoic sediment records enable scientists to carry out paleoenvironment studies in detail. Bayesian age-depth chronological models constructed from biostratigraphic control points for the drilling sites are applied on cluster boundaries generated from two popular unsupervised learning methods: *K*-means and random forest. The unsupervised learning methods experimented have produced compact and unambiguous clusters from the datasets, indicating that previously unknown data patterns can be revealed when all variables from the datasets are taken into account simultaneously. A study of synchroneity of past events represented by the cluster boundaries across geographically separated ocean drilling sites is achieved through converting the fixed depths of cluster boundaries into chronological ranges represented by Gaussian density plots which are then compared with known past events in the region. A Gaussian density peak at around 7.2 Ma has been identified from results of all three sites and it is suggested to coincide with the initiation of the East Asian monsoon. Contrary to traditional statistical approach, a priori assumptions are not required for unsupervised learning, and the clustering results serve as a novel data-driven proxy for studying the complex and dynamic processes of the paleoenvironment surrounding the ocean sediment. This work serves as a pioneering approach to extract valuable information of regional events and opens up a systematic and objective way to study the vast global ocean sediment datasets.

Received September 25, 2018; accepted February 28, 2019

E-mail: tsekevin@connect.hku.hk

Keywords machine learning, ocean sediments, unsupervised classification

1 Introduction

Since the initiation of scientific ocean drilling projects, attempts have been made to identify global events recorded in changes in ocean sediment characteristics such as sedimentation rates (Whitman and Davies, 1979; Nakamori, 2001). Geophysical and geochemical datasets obtained from ocean sediment cores have been studied in detail to characterize paleoenvironmental conditions of the Earth and numerous attempts were made to infer regional or global event signals from such datasets (Birks, 1989; Alley et al., 1997; Bennett and Fuller, 2002). Traditionally in the study of ocean drill datasets, past events are identified through laborious analyses of a few variables such as oxygen isotopes, carbonate contents, microfossil counts and sedimentation rates (Wang et al., 2000). Based on experience and intuition, experts examine a geochronological plot for the selected few data series to determine the class membership of datapoints in order to identify potential signals of global and regional events (Penn, 2005).

A common difficulty in the process is the correlation of datasets obtained across different sites and the establishment of synchroneity of events recognised in two or more geographically separated stratigraphic records (Parnell et al., 2008). A past event is defined in this study as a single depth in the core samples corresponding to one or more spatial-temporal processes relating to the formation of the sediments (Parnell et al., 2008).

Like all other branches of the natural sciences, the availability of large datasets in earth science research has become ubiquitous (Wolfe, 2013; Philip Chen and Zhang, 2014). Williams (2011) argued that with advancement in information technology, the computation and observation

of Earth's processes will result in an exponential growth in the data generated. Traditional methods based on manual interpretation of data will not be able to cope with such massive collections of scientific data (Way et al., 2012). An increasing number of applications based on sophisticated computational algorithms has emerged in the past decade in earth sciences (Weisberg, 2005; Lary et al., 2016; Liu and Srivastava et al., 2017; Pham et al., 2016; Pham et al., 2017) to extract key features and characteristic patterns of variability from large datasets.

Machine learning (ML), also known as statistical learning, is an effective approach for classification. ML methods are further divided into two main categories, namely supervised learning consisting of knowledge-driven methodologies (Murphy, 2012) and unsupervised learning comprising data-driven approaches (Kohonen, 2001; Cracknell et al., 2014; Tse el al., 2019;). Unsupervised learning algorithms analyze the underlying nature of relationships among the data (Lary et al., 2016) in a purely statistical way and have already proven their ability in carrying out classification tasks in physical science disciplines such as astronomy (Way et al., 2012). The objective of this paper is to develop a data-driven methodology to study synchroneity of past events from scientific ocean drilling data.

2 Data

The Ocean Drilling Program (ODP) was an international scientific endeavour started in the 1970s to explore the world's oceans by drilling and collecting drill cores for scientific analyses. A vast amount of high quality geophysical and geochemical data has been extracted from hundreds of kilometers of sediment cores obtained from worldwide expeditions. The program was later succeeded by the Integrated Ocean Drilling Program (IODP) in 2003. Texas A&M University (TAMU) has been managing the ODP/IODP core database since 1984 and most multivariate datasets from the expeditions are made available for public online access at the JANUS database website. The reliability and integrity of the ODP/IODP datasets are widely acknowledged in the scientific community by the significant amount of research output derived from these data. The online availability of these data has opened up opportunities for exploratory data analysis with ML methods to reveal information difficult to be discovered with traditional techniques.

The datasets chosen in this study are extracted from three ODP sites 1143, 1146 and 1148 located in the South China Sea (SCS) belonging to expedition leg 184 drilled between February and April of 1999. The SCS sites are chosen for exploratory data analysis by unsupervised learning methods owing to their exceptional continuity of depth and age of the core samples (Exp. 349 scientists, 2014). Supervised learning methods have previously been applied on the datasets for inferring lithology of missing core sections (Benaouda et al., 1999) but unsupervised methods are only recently been applied on the datasets (Tse et al., 2019). Locations of the SCS ODP sites are shown in Table 1 summaries some of their key information.

Site 1143 lies near Nansha Islands on the southern continental slope of the SCS between the terrigenous deposits of the Mekong Rivers to the south with a high-sedimentation rate (10–30 cm/ky) and carbonate-rich region of the northernmost southern margin with low sedimentation rate (1–2 cm/ky) (Huang and Wang, 1998). Site 1146 is located at the northeastern continental slope of the SCS and Site 1148 is located on the lowermost continental slope near the continent-ocean crust boundary off southern China (Wang et al., 2000). The sedimentary sequence at site 1148 goes back to the Lower Oligocene including the initiation of sea floor spreading of the SCS (Wang et al., 2000).

As summarised in Table 2, datasets from the selected ODP sites including Moisture and Density (MAD), Gamma Ray Attenuation (GRA), Natural Gamma Radiation (NGR), Magnetic Susceptibility (MSL), Reflective Spectrophotometry and Colorimetry (RSC), Smear Slide, Carbonate, Gas Chromatography and Interstitial Water have been downloaded from the Janus online database and thirty geophysical and geochemical variables are extracted from these datasets in this study. The mcd (metre composite depth) depth scale is used consistently in this study for all age-depth relationship in the datasets for each site. The mcd scale is constructed based on stratigraphical correlation of data such as L* (lightness) associated with the recovered cores from multiple holes of the same site. The scale allows the data points from different holes in the same site to be combined seamlessly along the vertical depth scale.

3 Methods

3.1 Past events

In this study, a past event is defined as a single depth in the core sample corresponding to a single point in time. It is

Table 1 Summary of ODP Leg 184 sites selected in this study

Site (Hole)	Water depth/m	Latitude	Longitude	Penetration/mbsf	Age at Base/Ma	Reference
1143A	2772	9°21.72′N	113°17.11′W	394	16	Wang et al. (2000)
1146A	2092	19°27.401′N	116°16.363′E	607	19	Wang et al. (2000)
1148A	3292	18°50.167′N	116°33.932′E	694	32	Wang et al. (2000)

Table 2 Summary of input datasets including 30 geophysical and geochemical variables from the ODP sites

Dataset	Type	Extracted Variables
Moisture and Density	Geophysical	water content (bulk), water content (dry), bulk density, dry density, grain density, porosity, void ratio
Gamma Ray Attenuation	Geophysical	density
Natural Gamma Radiation	Geophysical	bkg corrected counts
Magnetic Susceptibility	Geophysical	Drift-corrected suscept.
Reflectance Spectrophotometry and Colorimetry	Geophysical	L*, a*, b*
Smear Slide	Geophysical	sand, silt, clay
Carbonate	Geochemical	inorganic carbon, carbonate, total carbon, organic carbon, nitrogen, sulphur
Gas Chromatography	Geochemical	methane
Interstitial Water	Geochemical	ammonia, calcium, chloride, lithium, magnesium, phosphorus, phosphate

designated to be a point of transition (not necessarily rapid) between one stable paleoenvironment to another, as reflected by the geophysical and geochemical observables on two successive samples. Events are recorded in the core samples and the major challenge is to determine precisely the depths of such event and also the ages associated with such depths. The event/depth relationship can be tackled with a Bayesian approach to address the underlying stochastic noise in the process. An event could be associated with more than one type of underlying physical or chemical process contributed by regional and global factors.

Suppose the collected geophysical and geochemical variables are constituents of an observed paleoenvironment proxy $\overline{p}$, reflecting a conceptual unobserved p rising from $p_{\min}$ to $p_{\max}$, or vice versa. A crude but simple approach is adopted in which two depths $d_{\min}$ and $d_{\max}$ above and below the event depth d_{event} are defined respectively (Parnell et al., 2008). In a Bayesian approach, the probability of the exact location of d_{event} is evenly distributed between the interval defined by $d_{\min}$ and $d_{\max}$. The precision of the d_{event} depends on the sampling interval, core sizes, rapidity of the event and noise in the proxy $\overline{p}$.

3.2 Age-depth model

Two major uncertainties arise when determining the exact age of a past event. The first is the uncertainty in deciding on the depth position of the event and the second is the uncertainty in the age associated with this particular depth. The Bayesian age-depth model has become increasingly popular as a statistical tool to tackle the uncertainties and establish the synchroneity of two past events observed in two geographically separated sites (Parnell et al., 2008).

The key feature in building any age-depth chronology is monotonicity which is based on the principle that older sediments always lie beneath more recent ones. Traditional age-depth relationships are established through specific markers obtained in biostratigraphic and magnetostratigraphic studies (Isabella et al., 2006). Uncertainties arise in the calibration of these markers and in the stochastic interpolation between these markers when no age markers are present. A Bayesian age-depth chronology is a continuous random function $\theta(d)$ consistent with all available age data. The generated age-depth model is able to accommodate flat or very steep sections while maintaining monotonicity and continuity at all parts of the curve (Haslett and Parnell, 2008).

Suppose three core samples with depths d_1, d_2, d_3 are associated with age markers with ages t_1, t_2, t_3. The ages are modelled as normal distributions N(t_1, s_1), N(t_2, s_2), N(t_3, s_3) where s_1, s_2, s_3 are standard deviations for these ages. A simple Monte Carlo process can greatly reduce the uncertainties of t_1, t_2, t_3 in the age-depth chronology by adopting a stochastic linear interpolation to go through all possible depth ranges of dated core samples and express all the information in the form of confidence levels (Parnell et al., 2008). The Bayesian way of treating the age-depth profile is regarded as an example of a partially deterministic Markov process (Davis, 1984). In this process, the observed data is treated as a basis to generate random sample paths with likelihood computations (Haslett and Parnell, 2008). A robust method based on Poisson-gamma (CPG) and Treedie distributions (Jorgensen, 1987) is adopted in this study to establish the Bayesian age-depth chronology.

Figures 1–3 illustrate the Bayesian age-depth chronology for sites 1143, 1146, and 1148, respectively. Instead of a simple polynomial regression of the age control points, the Bayesian age chronology produced a 95% highest posterior density region. They are established with the biostratigraphic control markers listed in Tables A1, A2, and A3. The different thickness of core samples associated with each marker is also incorporated in the generation of the age-depth model.

3.3 Unsupervised classification

Unsupervised learning, also known as clustering, is a branch of machine learning which has been widely used in exploratory data analysis. The method identifies homo-

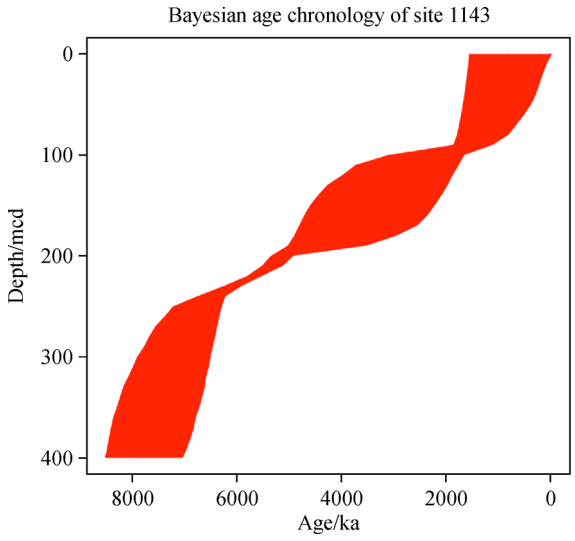

Fig. 1 Bayesian Age Chronology Plot of Site 1143. Red shaded area represents 95% highest poterior density regions (HDR) (Parnell et al., 2008), indicating the uncertainty of the ages between the dated depths of the control markers.

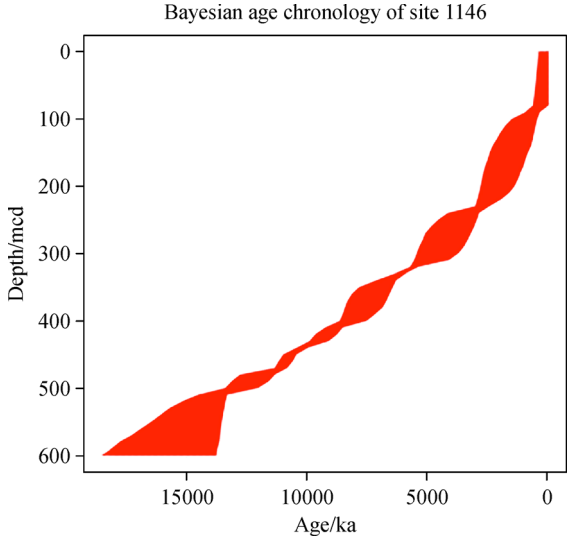

Fig. 2 Bayesian Age Chronology Plot of Site 1146. Red shaded area represents 95% highest poterior density regions (HDR) (Parnell et al., 2008), indicating the uncertainty of the ages between the dated depths of the control markers.

geneous groups (clusters) of data by applying a pre-defined criteria of similarity in the form of a distance metrics (Murphy, 2012; Romary et al., 2015). In this study, two well-tested unsupervised learning methods, *K*-means and random forest, are used to find a function $f: R^{N} \rightarrow R^{K}$ that maps an input vector x^{i} to a new feature vector of K clusters. Euclidean distance, $d = \sqrt{\sum_{i=1}(x^{i} - x^{i+1})^{2}}$, a common distance metric for measuring similarity (Sam-

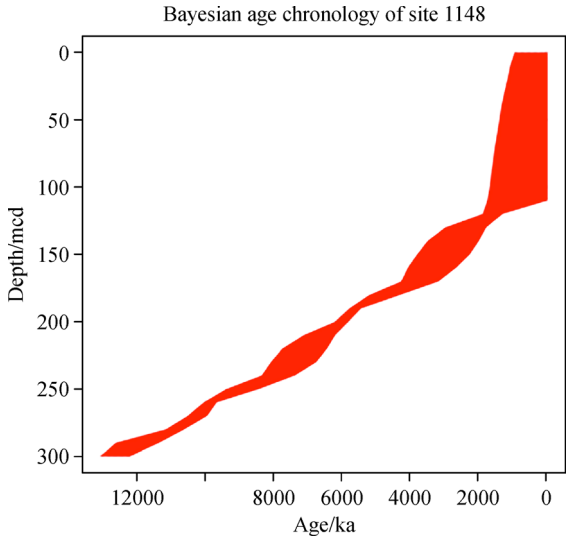

Fig. 3 Bayesian Age Chronology Plot of Site 1148. Red shaded area represents 95% highest poterior density regions (HDR) (Parnell et al., 2008), indicating the uncertainty of the ages between the dated depths of the control markers.

mon, 1969; Kohonen, 2001; Singh et al., 2013) in unsupervised learning, is used in this study. Unlike supervised learning, there is no a priori information assumed in the input data and no predicted values are to be produced.

Unsupervised classification has been successfully applied in exploratory data analysis for discovering "interesting patterns" in the data itself, and the results can potentially lead to new research questions (Jain, 2010; Hennig, 2016). For example, in astronomy, a new type of star was discovered based purely on clustering astrophysical measurements (Cheeseman et al., 1988). Clustering methods aim at identifying data structure in either labeled or unlabeled datasets by organizing data objectively into homogeneous groups in which the within-group similarity is minimized and between group dissimilarity is maximized (Liao, 2005). One advantage of this approach over manual classification is that an objective statistical significance is established within large datasets while overall coherence of the resulting classes is maintained (Romary et al., 2015).

3.3.1 *K*-means

The *K*-means clustering algorithm is one of the most popular unsupervised learning method to solve clustering problems (MacQueen, 1967; Chauhan et al., 2016). The *K*-means is a subset of Self Organizing Map with particular choice of parameters (Liu and Weisberg, 2011). As a partitioning method, the number of clusters, k, is predetermined and data is decomposed into a set of k non-overlapping clusters by initializing k centroids and then

refining the cluster centers by iteration (Chauhan et al., 2016). The clusters formed are also known as disjoint clusters, meaning that each data point is in exactly one subset cluster (MacQueen, 1967). The method is capable of handling large datasets with continuous data provided that no non-convex clusters are present (Kabacoff, 2015).

Given an integer k, as a set X of n points (where $n \geqslant k$) in an m-dimensional Euclidean space, where

$X = [x_i = (x_{i_1},...x_{i_m})^{\mathrm{T}} \in R^m, i = 1,...,n]$, the objective is to assign the n points into k disjoint clusters $C = (C_1, \cdots, C_k)$, $C_k \cap C_{k'} = \phi$ centered at cluster means μ_j for $j = 1,...,k$, based on the initial conditions. K-means seeks a clustering result where within-cluster variation $W(C_k)$ is a minimum in Eq. (1). A common choice of $W(C_k)$ is the Euclidean distance:

$$W(C_k) = \sum_{i,i' \in C_k} \sum_{j=1}^{k} (x_{i,j} - \mu_{i'j})^2. \quad (1)$$

K-means has the advantage of being easy to implement and relatively simple to visualize the result. The number of clusters is required to be specified beforehand. A common method to determine the right number of clusters is by plotting the sum of within-cluster variance against number of clusters as shown in Fig. 4. The initial choice of cluster centers will also influence the clustering results and their subsequent convergence into local minima.

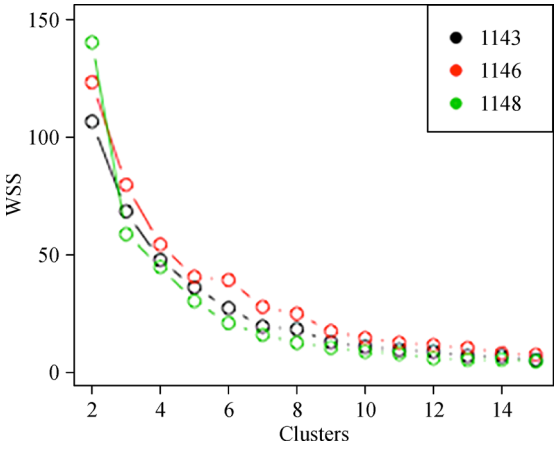

Fig. 4 Total within-groups sums of squares against the number of clusters.

3.3.2 Random Forest

Random Forest (RF) is an ensemble method that utilizes a majority vote to predict classes based on the partition of data from multiple decision trees. Multiple trees are grown by randomly subsetting a number of variables to split at each node of the decision trees and by bagging (Breiman, 2001). RF implements the Gini index obtained from Eq. (2) to determine the "best split" threshold of input values (p_i stands for the probability of class i at node n_c) for given classes.

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i). \quad (2)$$

The Gini index returns a measure of class heterogeneity within child nodes as compared to the parent node (Breiman, 1984). Instead of using the Euclidean distance metric used in this study, a distance measure based on the proximity of the RF algorithm is used (Breiman, 2001). Successful applications of RF in different fields of geosciences have been demonstrated (Cracknell et al., 2014; Goetz et al., 2015; Insua et al., 2015).

3.4 Data pipeline

A data pipeline is constructed to process the datasets for the unsupervised learning study as shown in Fig. 5. Values of thirty geophysical and geochemical variables are first extracted from each of the three chosen OPD sites and combined as a single datacube for each site. No outliers are discarded in the process and the resolution of the data is chosen to be no smaller than the length of a core sample. The datacube for each site is then processed by the two chosen machine learning algorithms to form data clusters for each site. The number of clusters chosen was based on the within-group variance shown on Fig. 4 and a balance is struck between minimizing total variance and the ease of visualization for the clustering results. Six clusters are generated for each of the unsupervised learning methods on the three datacubes, resulting in a minimum of five cluster boundaries for each site. The cluster boundaries for each datacube are then converted into Bayesian chronological ranges based on the site-specific Bayesian age-depth profiles illustrated in Figs. 1–3. A total of six sets of these cluster boundaries (3 from K-means and 3 from RF) are converted into depth ranges which are statistically bounded, allowing a study of their synchroneity across the three different ODP sites to be made. A mixed Gaussian density plot for these depth ranges representing the underlying cluster boundaries is produced for each site for both unsupervised learning methods.

4 Results

Unsupervised clustering results are shown in Figs. 6–8 for K-means and RF on the combined datacube of thirty variables from the three selected ODP sites in the SCS. Data points are plotted with their depths in mcd (meters composite depth) against their assigned cluster class. Each data point corresponds to a multi-variate observation in the synthetic dataset and the compact and connected clusters are observed for the results obtained by both methods. Unsupervised learning algorithms have successfully partitioned the data into unambiguous clusters segments and the overlapping of clusters is insignificant.

Results from both methods for site 1143 are shown in Fig. 6 and the clusters are observed to be distributed quite

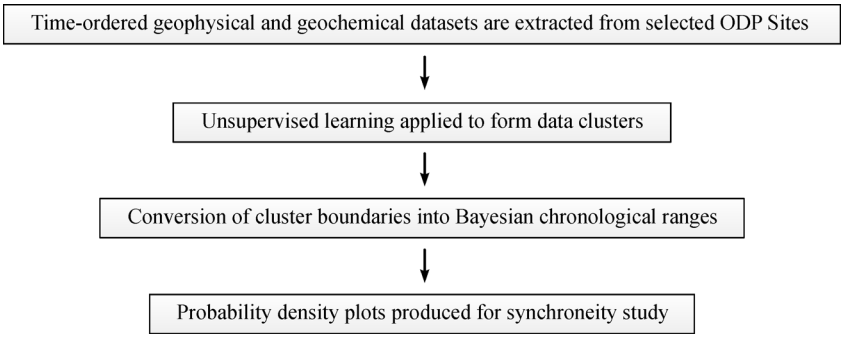

Fig. 5 Unsupervised machine learning pipeline for ODP datasets.

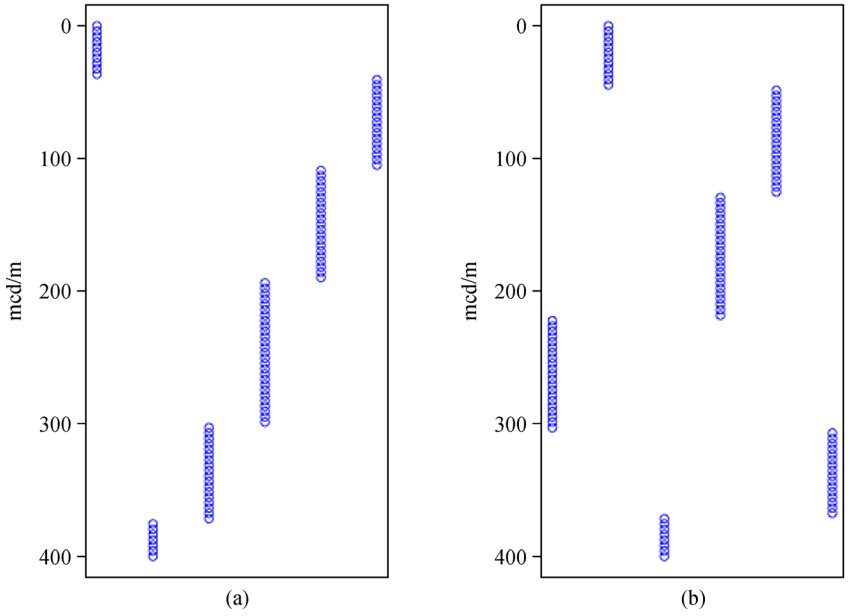

Fig. 6 Data clusters assigned by class numbers 1 to 6 by *K*-means (left) and RF (right) for ODP site 1143A. Each data point is only assigned one class and the class number is arbitrary for the two adopted unsupervised classification methods. (a) 1143A *K*-means (b) 1143A RF.

evenly along the depth scale. It should be noted that the assignment of class number is arbitrary for the different machine learning methods. The same data point at a particular depth may not be assigned the same class number by *K*-means and RF. Only the boundaries between different classes or clusters are the focus of this study.

Significant agreement by both methods is observed on cluster boundaries at 299.0–307.1 (mcd) and 367.7–375.8 (mcd). The cluster boundary is expressed as a depth range rather than a single point in depth since each core sample has a finite length. Tables 3 and 4 are shown for the corresponding chronological age ranges converted from the cluster boundaries obtained by the two unsupervised learning methods. The Bayesian chronological ranges are obtained by applying the Bayesian age chronology of Site 1143 shown on Fig. 1.

Figure 7 displays the results for site 1146 and it is observed that both methods agree quite well with each other in the classification experiment except for cluster boundary at 78.8–88.8 (mcd) obtained by *K*-means and cluster boundary at 454.5–460.6 (mcd) by RF. Tables 5 and 6 list the chronological age ranges obtained by *K*-means and RF for the cluster boundaries. The corresponding

Table 3 *K*-means class boundaries and Bayesian chronological ranges for site 1143 (Li et al., 2004)

Cluster boundary (mcd)	Bayesian chronological range/Ma
36.4–40.4	0.3–1.6
105.1–109.1	1.5–2.1
189.9–193.9	3.1–5.0
299.0–303.0	6.5–7.7
371.7–375.8	7.0–8.2

Table 4 RF class boundaries and Bayesian chronological ranges for site 1143 (Li et al., 2004)

Cluster boundary (mcd)	Bayesian chronological range/Ma
44.4–48.5	0.3–1.6
125.3–129.3	2.0–4.0
218.2–222.2	5.5–5.8
303.0–307.1	6.6–7.8
367.7–371.8	7.0–8.2

Bayesian chronological ranges are obtained by applying the Bayesian age chronology of site 1146 shown on Fig. 2.

Compact and connected clusters are produced from both methods with no overlapping of clusters as before for results from site 1148 as shown in Fig. 8. Agreement of cluster boundaries by both methods is still very good, noticeably at 55.7–58.6 (mcd), 93.7–99.6 (mcd) and 216.8–219.7 (mcd). Tables 7 and 8 are shown for the chronological age ranges obtained by *K*-means and RF respectively. The corresponding Bayesian chronological ranges are obtained by applying the Bayesian age chronology of Site 1148 shown on Fig. 3.

Figure 9 shows the density plot of Bayesian chronological ranges obtained by the two methods on the cluster boundaries for site 1143. Both plots display a total of three major peaks although the RF plot has a minor peak located before 6 Ma. The peaks of the density plot indicate a higher probability of the cluster boundary being located at the particular age. A sharper peak is resulted from a narrower portion of the underlying Bayesian age chronology while a broader peak is produced from a less certain part of the Bayesian age-depth profile. The first two peaks obtained by both methods are observed to be occurring before 5 Ma although *K*-means provides more spaced peaks and the peaks for RF are closer. The peak at around 7.5 Ma is very sharp for both method at site 1143.

More peaks are observed in Fig. 10 for both methods for 1146. Both *K*-means and RF yield three peaks before 8 Ma. *K*-means shows a peak between 12–14 Ma while RF yields two peaks at around 11 Ma and 14 Ma. The general pattern of peaks indicates a good agreement between the two methods.

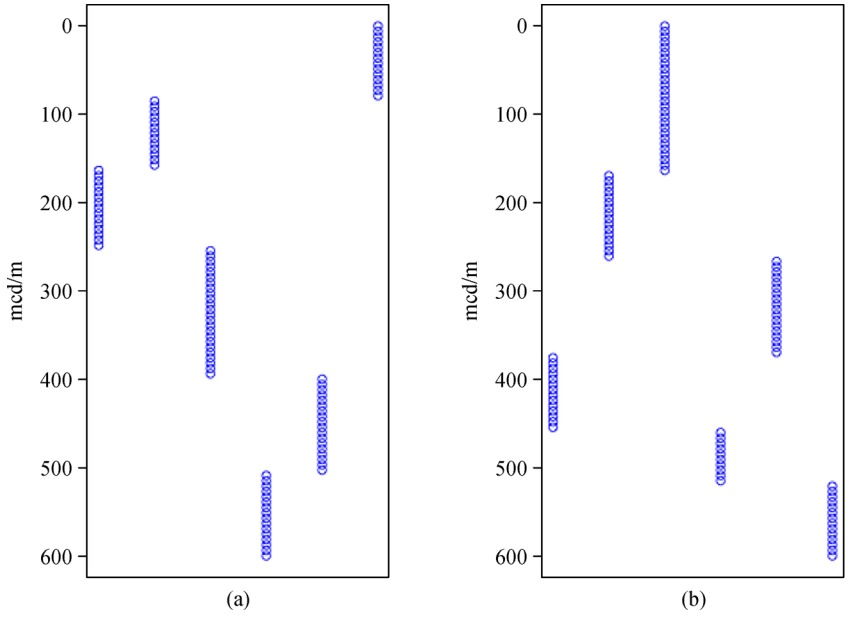

Fig. 7 Data clusters assigned by class numbers 1 to 6 by *K*-means (left) and RF (right) for ODP site 1146A. Each data point is only assigned one class and the class number is arbitrary for the two adopted unsupervised classification methods. (a) 1146A *K*-means; (b) 1146A RF.

Table 5 *K*-means class boundaries and Bayesian chronological ranges for site 1146 (Li et al., 2004)

Cluster boundary (mcd)	Bayesian chronological range/Ma
78.8–88.8	0.2–0.9
157.6–163.6	1.8–3.0
248.5–254.5	2.9–4.5
339.4–345.5	6.8–7.6
503.0–509.1	12.0–14.0

Table 6 RF class boundaries and Bayesian chronological ranges for site 1146 (Li et al., 2004)

Cluster boundary (mcd)	Bayesian chronological range/Ma
163.6–169.7	1.5–2.6
260.6–266.7	3.0–4.9
369.7–375.8	7.2–7.8
454.5–460.6	10.9–11.8
515.2–521.2	13.0–14.5

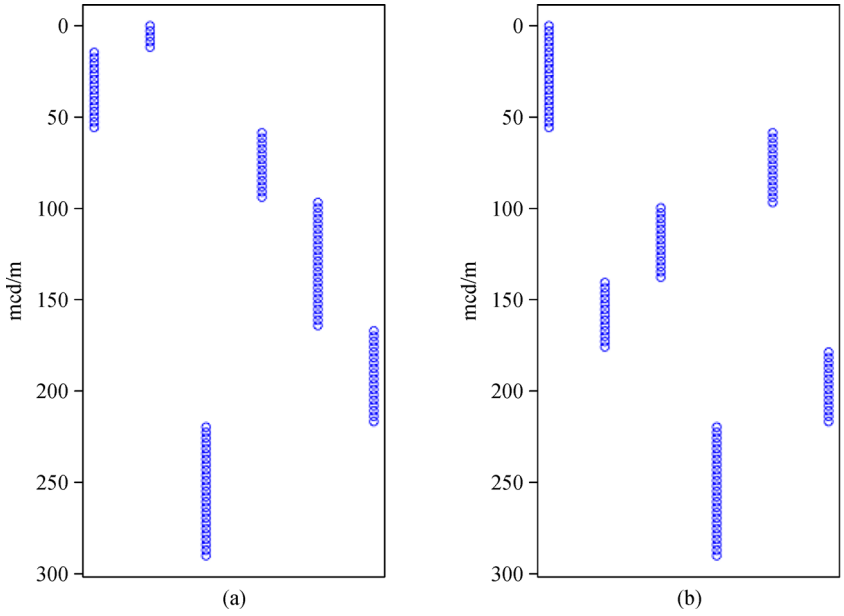

Fig. 8 Data clusters assigned by class numbers 1 to 6 by *K*-means (left) and RF (right) for ODP site 118A. Each data point is only assigned one class and the class number is arbitrary for the two adopted unsupervised classification methods. (a) 1148A *K*-means; (b) 1148A RF.

Table 7 *K*-means class boundaries and Bayesian chronological ranges for site 1148, max depth is limited to 290 to match 1146 and 1143 (Li and Li, 2004)

Cluster boundary (mcd)	Bayesian chronological range/Ma
11.7–14.6	0.1–1.6
55.7–58.6	0.2–1.7
93.7–96.7	0.3–1.9
164.0–167.0	2.7–4.0
216.8–219.7	6.5–7.5

Table 8 RF class boundaries and Bayesian chronological ranges for site 1148, max depth is limited to 290 to match 1146 and 1143 (Li and Li, 2004)

Cluster boundary (mcd)	Bayesian chronological range/Ma
55.7–58.6	0.2–1.7
96.7–99.6	0.3–1.8
137.7–140.6	2.1–3.1
175.8–178.7	4.0–4.8
216.8–219.7	6.1–6.3

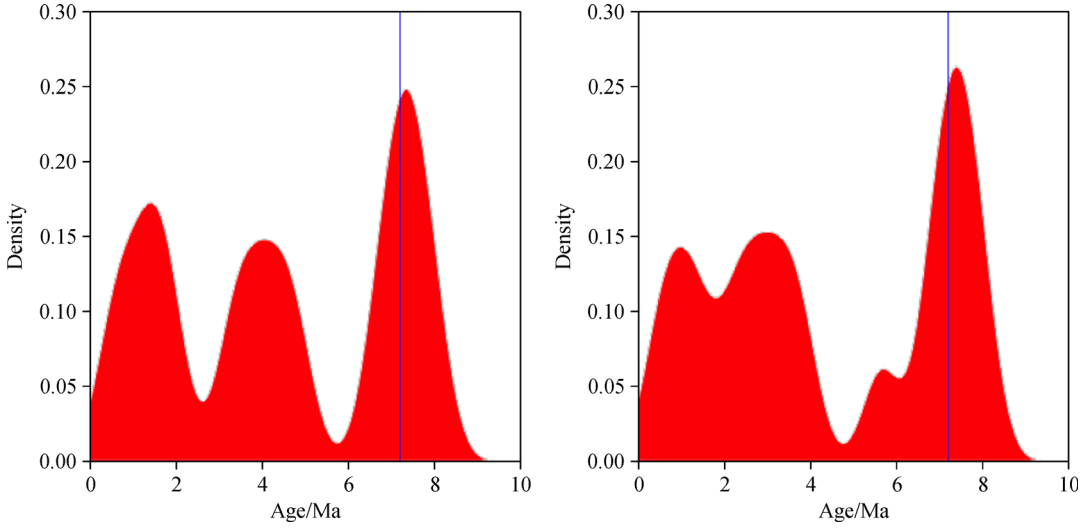

Fig. 9 Mixed Gaussian density plots for the cluster boundaries produced from *K*-means (left) and RF (right) at site 1143. The blue line indicates the commencement of east Asian monsoon at 7.2 Ma (An, 2000).

It is worth noting that by comparing Fig. 9 and Fig. 10, the peak at around 7 Ma is present for both sites. A regime of peaks before around 4 Ma is also present for both sites.

Figure 11 shows the density plots for site 1148. The shape of the plot is more skewed than the plots of other two sites. The highest peak is located at around 1 Ma for both methods and a much lower peak is displayed before 5 Ma. A peak at around 7 Ma is shown by *K*-means and at around 6 Ma by RF.

5 Discussion

A novel method for studying past events based on scientific ocean drilling datasets is explored in this study with the simultaneous application of unsupervised learning methods and Bayesian age-depth chronologies. The two-pronged approach opened up a possibility of data-driven study of event synchroneity across different drilling sites.

A key assumption in this study is the definition of a past event and in this study a past event is defined as the cluster boundaries interpreted by unsupervised learning methods without any a priori information. Unlike human interpretation which is often limited to a few variables, unsupervised learning methods take a larger number of variables and tens of thousands of observations into account at the same time to extract hidden patterns within the datasets. By imposing a similarity metric on the data, data points with

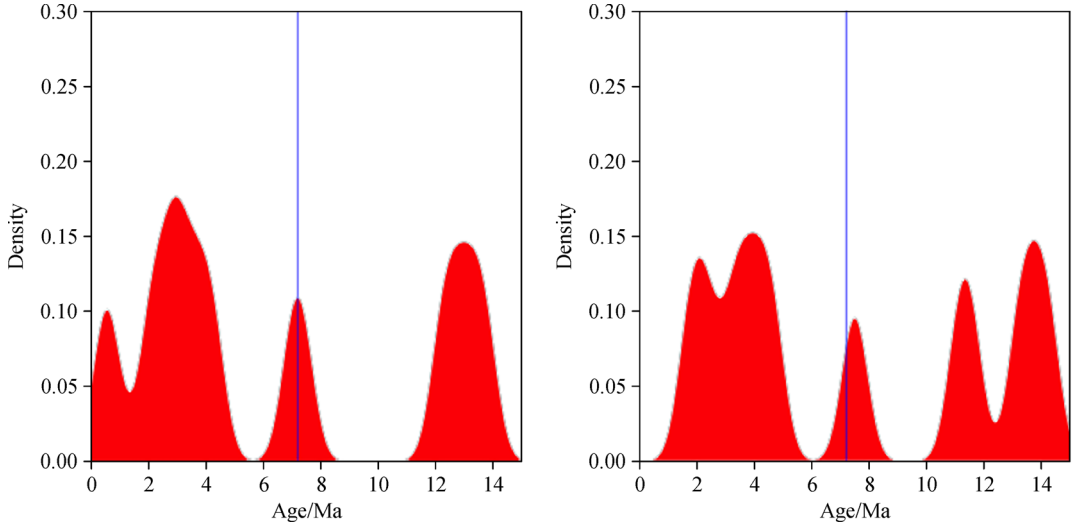

Fig. 10 Mixed Gaussian density plots for the cluster boundaries produced from *K*-means (left) and RF (right) at site 1146. The blue line indicates the commencement of east Asian monsoon at 7.2Ma (An, 2000).

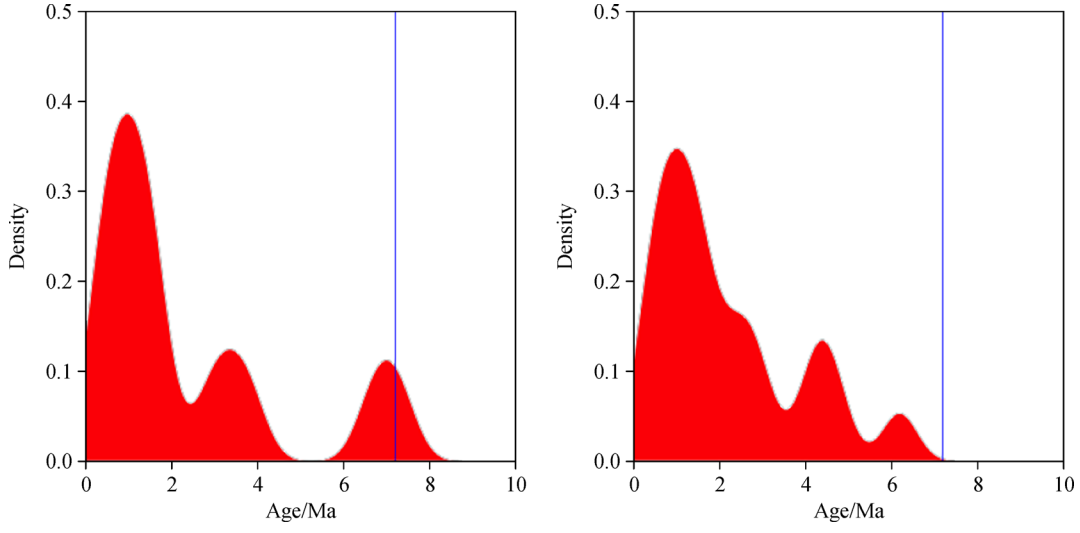

Fig. 11 Mixed Gaussian density plots for the cluster boundaries produced from *K*-means (left) and RF (right) at site 1148. The blue line indicates the commencement of east Asian monsoon at 7.2 Ma (An, 2000).

the shortest measured distances among themselves are assigned to the same cluster by the computer algorithm.

The principles behind the two popular unsupervised learning methods, *K*-means and RF in this study are very different. *K*-means clustering produces results based on linear decision hyperplanes while RF utilizes decision trees more sensitive to non-linear relationship between the variables (Tse et al., 2019). Similar class boundaries observed from results generated from the two unsupervised learning methods suggest that the results could carry meaningful information.

The clusters generated from the unsupervised methods are surprisingly continuous in that each resulting cluster corresponds to a section of depth range without interruption by other clusters assigned within it. Each cluster is interpreted as representing a certain relatively stable stage of an evolving paleoenvironment including the generation, transport and deposition of the ocean sediments. Based on this interpretation, the cluster boundaries would contain a proxy signal of certain hidden changes in the paleoenvironment. In reality these hidden changes could be corresponding to anything ranging from short-lived excursions to complete transitions of state in the paleoenvironment concerned. The hidden changes may involve multiple geophysical or geochemical processes in the paleoenvironment and could be either local or regional across spatial and temporal dimensions.

The mixed Gaussian density plots of data events for the three sites provide a good visual demonstration of the potential strength of this approach in the study of event synthroneity. The location and shape of density peaks of site 1143 and site 1146 show a striking resemblance even though their age-depth models are very different and the sites are separated by hundreds of kilometers in geographical distance. Both site 1143 and site 1146 display a group of two to three peaks before around 5 Ma and a very good agreement of a peak at around 7 Ma is observed from both methods. The agreement of the location of the peaks could be attributed to the fact that both sites are located over a relative thick layer of ocean sediment and that the cores contain continuous sediment samples of excellent quality. The past events could be regional instead of local as the sediment sources and related paleoenvironment must be quite different with such a large geographical distance apart.

This argument also explains why the density plots of site 1148 differ a lot from those of the other two sites even though it is relatively close to one of the two sites. Site 1148 is located at the bottom part of the continental shelf in the vicinity of the continental/ocean boundary. The constituents of sediment, physical and chemical processes could be quite different from that of sites 1143 and 1146 where they are located on the upper part of continental shelf where a thick sediment layer has been accumulated. It is important, however, to observe that a small but visible peak is present at around 7 Ma at site 1148. This further suggest that the peak represents a proxy signal of a regional event affecting all of the three sites in this study.

A possible candidate for such a regional event is the commencement of the East Asian monsoon at around 7.2 Ma (An, 2000). The commencement of the monsoon entails a pervasive environmental change across whole catchments feeding into the SCS. Such a regional paleoenvironmental event could have profound effects on the formation, transportation and deposition of the sediments. If such a signal exists in the sediment record, it will most likely be presenting itself across all cores taken from different sites in the region instead of locally. According to Wang and Li (2009), other candidates for the major SCS structural and tectonic events occurring at around 7 Ma include the beginning of the uplift of Taiwan Island, end of uplift of Western Cordillera and the beginning of formation of the Philippine trench. Although the identification of events interpreted from the unsupervised methods cannot be confirmed, the study has demonstrated the potential power of a statistical approach in the identification of synchronous regional events.

It should be emphasized that, while the data clusters are real in their own sense, there is no guarantee that they correspond to any real physical event. Data clustering methods are sometimes coined as trying to "find a needle in a haystack when it is not sure whether there is one inside" (Hennig, 2016; Pavlidou et al., 2016). Any clustering results will be questioned for their relevance to the real world (Wagstaff, 2012). Nonetheless, previously unrecognized patterns and correlations will emerge from a logical integration and evaluation of reliable datasets (Hazen, 2014). It is important for any ML studies to be able to communicate back to the domain problem with applicable results to improve on current methods. In other words, any revealed data pattern will be useful in further investigations of the datasets.

6 Conclusions

The study demonstrated that unsupervised machine learning method could be utilized in exploratory data analysis on scientific ocean drilling datasets. The methods have not been applied on the SCS datasets before and the results from this study indicated that the methods have unveiled meaningful information otherwise difficult to be uncovered by human experts. Further studies could be conducted with more sophisticated classification methods such as Self Organizing Map (Liu and Weisberg, 2011) on more sites and datasets.

Appendix

Table A1 Astrochronologically tuned Planktonic foraminifer biostratigraphic zonal boundaries, Site 1143 (~45 to ~93 ky resolution). FO and LO represents the bioevents of first occurrence and last occurrence of species respectively into biozones (Nathan and Leckie, 2003)

Depth (mcd) Top/Bottom	Thickness	Age/Ma	Datum	Stratigraphic position	Ref.
93.5/94.29	0.8	1.69	FO *medium Gephyrocapsa* spp.	-	Wang et al., 2000
190.8/200.6	9.8	4.99	LO *C. acutus*	-	Wang et al., 2000
216.6/219.6	3	5.54	FO *Sphaeroidinella dehiscens*	N18 / N19	Nathan and Leckie, 2003
224.07/232.52	8.5	5.82	FO *Globorotalia tumida*	N17b / N18	Nathan and Leckie, 2003
238.52/241.05	3.5	6.4	FO *Pulleniatina primalis*	N17a / N17b	Nathan and Leckie, 2003
453.06/456.06	3	8.58	FO *Globorotalia plesiotumida*	N16 / N17a	Nathan and Leckie, 2003

Table A2 Astrochronologically tuned Planktonic foraminifer biostratigraphic zonal boundaries, Site 1146. FO and LO represents the bioevents of first occurrence and last occurrence of species respectively into biozones (Nathan and Leckie, 2003)

Depth (mcd) Top/Bottom	Thickness	Age/Ma	Datum	Stratigraphic position	Ref.
83.9/93.4	9.5	0.46	LO *P. Iacunosa*	-	Wang et al., 2000
226.7/237.1	10.4	2.83	LO *D. tamalis*	-	Wang et al., 2000
318.36/321.41	3.1	5.54	FO *Sphaeroidinella dehiscens*	N18 / N19	Nathan and Leckie, 2003
321.41/324.36	3.0	5.82	FO *Globorotalia tumida*	N17b / N18	Nathan and Leckie, 2003
337.56/338.52	1.0	6.4	FO *Pulleniatina primalis*	N17a / N17b	Nathan and Leckie, 2003
406.63/409.63	3	8.58	FO *Globorotalia plesiotumida*	N16 / N17a	Nathan and Leckie, 2003
432.38/433.88	1.5	9.82	FO *Neogloboquadrina acostaensis*	N15 / N16	Nathan and Leckie, 2003
443.83/ 445.33	1.5	10.49	LO *Paragloborotalia mayeri*	N14 / N15	Nathan and Leckie, 2003
471.92/ 473.42	1,5	11.19	FO *Globoturborotalita nepenthes*	N13 / N14	Nathan and Leckie, 2003
504.1/505.6	1.5	13.42	FO *Globorotalia fohsi s.l.*	N11 / N12	Nathan and Leckie, 2003

Table A3 Age-Depth Model based on planktonic foraminifer datums, Site 1148. FO and LO represents the bioevents of first occurrence and last occurrence of species respectively into biozones (Li et al., 2004)

Depth (mcd) Top/Bottom	Thickness	Age/Ma	Datum	Ref.
115.34/125.76	10.42	1.69	FO *medium Gephyrocapsa* spp.	Wang et al., 2000
176.83/176.98	0.15	4.20	LO *Globoturborotalita nepenthes*	Li et al., 2004
188.23/189.08	0.85	5.54	FO *Sphaeroidinella dehiscens*	Li et al., 2004
195.98/196.18	0.2	5.82	FO *Globorotalia tumida*	Li et al., 2004
206.71/206.88	0.17	6.2	FO *Globigerinoides conglobatus*	Li et al., 2004
244.18/244.33	0.15	8.3	FO *Globigerinoides extremus*	Li et al., 2004
257.23/257.08	0.15	9.5-9.8	LO *Globoquadrina dehiscens*	Li et al., 2004
259.63/259.78	0.15	9.82	FO *Neogloboquadrina acostaensis*	Li et al., 2004
275.21/275.23	0.02	10.49	LO *Paragloborotalia mayeri*	Li et al., 2004
283.63/283.93	0.3	11.19	FO *Globoturborotalita nepenthes*	Li et al., 2004
301.03/301.00	0.03	13.00	LO *Globorotalia fohsi*	Li et al., 2004
303.43/303.13	0.3	13.42	FO *Globorotalia fohsi*	Li et al., 2004

References

Alley R B, Mayewski P A, Sowers T, Stuiver M, Taylor K C, Clark P U (1997). Holocene climatic instability: a prominent, widespread event 8200 yr ago. Geology, 25(6): 483–507

An Z (2000). The history and variability of the East Asian paleomonsoon climate. Quat Sci Rev, 19(1): 171–187

Benaouda D, Wadge G, Whitmarsh R B, Rothwell R G, MacLeod C

(1999). Inferring the lithology of borehole rocks by applying neural network classifiers to downhole logs: an example from the ocean drilling program. Geophys J Int, 136(2): 477–491

Bennett K D, Fuller J L (2002). Determining the age of the Mid-Holocene *Tsuga canadensis* (hemlock) decline, eastern North America. Holocene, 12(4): 421–429

Birks H J B (1989). Holocene isochrone maps and patterns of tree-spreading in the British isles. J Biogeogr, 16(6): 503–540

Breiman L (1984). Classification and Regression Trees. New York: Chapman & Hall

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

Chauhan S, Ruhaak W, Khan F, Enzmann F, Mielke P, Kersten M, Sass I. (2016). Processing of rock core microtomogrpahy images: using seven different machine learning algorithms. Comput Geosci, 86: 120–128

Cheeseman P, Self M, Kelly J, Taylor W, Freeman D, Stutz J (1988). Bayesian classification. In: Proceedings of the Seventh AAAI National Conference on Artificial Intelligence. AAAI'88. New York: AAAI Press, 607–611

Cracknell M J, Reading A M, McNeill A W (2014). Mapping geology and volcanic hosted massive sulfide alteration in the Hellyer-Mt Charter region, Tasmania, using random forest and self-organising maps. Aust J Earth Sci, 61: 287–304

Davis M H A (1984). Piecewise-deterministic markov processes: a general class of non-diffusion stochastic models (with discussion). J R Stat Soc B, 46: 353–388

Exp. 349 scientists. (2014). IODP expedition 349 preliminary report, South China Sea tectonics- opening of the South China Sea and its implications for southeast asian tectonics, climates and deep mantle processes since the late mesozoic. Initial reports. New York: IODP

Goetz J N, Brenning A, Petschko H, Leopold P (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput Geosci, 81: 1–11

Haslett J, Parnell A (2008). A simple monotone process with application to radiocarbon-dated depth chronologies. J R Stat Soc Ser C Appl Stat, 57(4): 399–418

Hazen R (2014). Data-driven abductive discovery in mineralogy. Am Mineral, 99: 2165–2170

Hennig C (2016). What are the true clusters? Pattern Recognit Lett, 64: 53–62

Insua T L, Hamel L, Moran K, Anderson L M, Webster J M (2015). Advanced classification of carbonate sediments based on physical properties. Sedimentology, 62: 590–606

Isabella R, Backman J, Fornaciari E. (2006). A review of calcareous nannofossil astrobiochronology encompassing the past 25 million years. Quat Sci Rev, 25: 3113–3137

Jain A K (2010). Data clustering: 50 years beyond *k*-means. Pattern Recognit Lett, 31: 651–666

Jorgensen B (1987). Exponential dispersion models. J R Stat Soc B, 49: 127–162

Kabacoff R I (2015). R in Action- Data analysis and graphics with R. San Jose: Manning

Kohonen T (2001). Self-Organizing Maps. New York: Springer-Vertag

Lary D J, Alavi A H, Gandomi A H, Walker L W. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7: 3–10

Li Q, Jian Z, Li B (2004). Oligocene-miocene planktonic foraminiferal biostratigraphy, site 1148, northern South China Sea. In: Proceedings of ODP Sci. Results. New York: IODP, 184 (1): 1–26

Liao T W (2005). Clustering of time series data—a survey. Pattern Recognit, 38: 1857–1874

Liu Y, Weisberg R H (2005). Patterns of ocean current variability on the west florida shelf using the selforganizing map. J Geophys Res Oceans, 110(C6): 0148–0227

Liu Y, Weisberg R H (2011). A review of self-organizing map applications in meteorology and oceanography. In: Mwasiagi J I, ed. Self-Organizing Maps—Applications and Novel Algorithm Design. Rijeka, Croatia: Intech, 253–272

MacQueen J (1967). Some methods for classification and analysis of multivariate observations. In: Le Cam L M, Neyman J, eds. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. San Francisco: University of California, 281–297

Murphy K P (2012). Machine Learning A Probabilistic Perspective. New York: The MIT Press

Nakamori T (2001). Global carbonate accumulation rates from cretaceous to present and their implications for the carbon cycle model. Isl Arc, 10(1): 1–8

Nathan S, Leckie R (2003). Miocene planktonic foraminiferal biostratigraphy of sites 1143 and 1146, ODP leg 184, South China Sea. Proc ODP, Sci Results, 184 (1): 1–43

Parnell A, Haslett J, Allen J, Buck C, Huntley B (2008). A flexible approach to assessing synchroneity of past events using bayesian reconstructions of sedimentation history. Quat Sci Rev, 27(19): 1872–1885

Pavlidou E, van der Meijde M, van der Werff H, Hecker C (2016). Finding a needle by removing the haystack: a spatio-temporal normalization method for geophysical data. Comput Geosci, 90: 78–86

Penn B S (2005). Using self-organizing maps to visualize high-dimensional data. Comput Geosci, 31(5): 531–544

Pham B T, Bui D T, Prakash I (2017). Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and J48 decision trees methods: a comparative study. London. Geotech Geol Eng, 35(6): 2597–2611

Pham B T, Tien Bui D, Pham H V, Le H Q, Prakash I, Dholakia M B (2016). Landslide hazard assessment using random subspace fuzzy rules based classifier ensemble and probability analysis of rainfall data: a case study at Mu Cang Chai District, Yen Bai Province (Viet Nam). J In Soc of Remote Sensing, 45(4): 673–683

Philip Chen C L, Zhang C Y (2014). Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf Sci, 275: 314–347

Romary T, Rivoirard J, Deraisme J (2015). Unsupervised classification of multivariate geostatistical data: two algorithms. Comput Geosci, 85: 96–103

Sammon J W (1969). A nonlinear mapping for data structure analysis. IEEE Trans Comput, 18: 401–409

Singh A, Yadav A, Rana A (2013). *K*-means with three different distance metrics. Int J Comput Appl, 67(10): 13–17

Srivastava A, Nemani R, Steinhaeuser K (2017). Large-Scale Machine

Learning in the Earth Sciences. New York: Chapman and Hall/CRC

Tse K C, Chiu H C, Tsang M Y, Li Y, Lam E Y (2019). Unsupervised learning on scientific ocean drilling datasets from the South China Sea. Front Earth Sci, 13(1): 180–190

Wagstaff K L (2012). Proceedings of the 29th international conference on machine learning. San Francisco: California Institute of Technology

Wang P, Blum P, et al. (2000). 2000 Proceedings of the Ocean Drilling Program, Initial Reports, Vol. 184. Initial Reports. New York: ODP Press

Wang P, Li Q (2009). The South China Sea–paleoceanography and sedimentology. In: The South China Sea–Paleoceanography and Sedimentology. Berlin: Springer

Way M J, Scargle J D, Ali K M, Srivastava A N (2012). Advances in Machine Learning and Data Mining for Astronomy. New York: CRC Press

Whitman J M, Davies T A (1979). Cenozoic oceanic sedimentation rates: How good are the data? Mar Geol, 30(34): 269–284

Williams R (2011). Earth Science: New Methods and Studies. London: Apple Academic Press

Wolfe P J (2013). Making sense of big data. Proc Natl Acad Sci USA, 110(45): 18031–18032