Front. Earth Sci. 2019, 13(1): 180–190
https://doi.org/10.1007/s11707-018-0704-1

RESEARCH ARTICLE

Unsupervised learning on scientific ocean drilling datasets from 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, Toronto, ON M5S 2M8, Canada
4 Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong, China

Abstract Unsupervised learning methods were applied to explore data patterns in multivariate geophysical datasets collected from ocean floor sediment core samples coming from scientific ocean drilling in the South China Sea. Compared to studies on similar datasets, but using supervised learning methods which are designed to make predictions based on sample training data, unsupervised learning methods require no *a priori* information and focus only on the input data. In this study, popular unsupervised learning methods including K-means, self-organizing maps, hierarchical clustering and random forest were coupled with different distance metrics to form exploratory data clusters. The resulting data clusters were externally validated with lithologic units and geologic time scales assigned to the datasets by conventional methods. Compact and connected data clusters displayed varying degrees of correspondence with existing classification by lithologic units and geologic time scales. K-means and self-organizing maps were observed to perform better with lithologic units while random forest corresponded best with geologic time scales. This study sets a pioneering example of how unsupervised machine learning methods can be used as an automatic processing tool for the increasingly high volume of scientific ocean drilling data.

Keywords machine learning, unsupervised learning, ODP, IODP, clustering

1 Introduction

Like all other branches of natural sciences, the study of geosciences is undergoing a major transformation with the advent of fast computers and machine learning algorithms (Longo et al., 2014). The explosive increase in data rates, data complexity and data quality of geosciences datasets (Schnase et al., 2016) means that objective and efficient methods are in high demand for geoscientists to make sense of the copious amounts of data arriving continuously. Over the past two decades, machine learning methods have been rapidly adopted in such fields of geosciences as remote sensing (Marzo et al., 2006; Lary et al., 2016), geochemical analysis (Templ et al., 2008; Xiong and Zuo, 2016), landslide mapping (Yao et al., 2008; Pham et al., 2016, 2017a, b) and scientific ocean drilling (Benaouda et al., 1999; Insua et al., 2015; Jeong and Park, 2016).

Unsupervised learning is a branch of machine learning that aims to determine hidden structures among input data, with no response variable leading the process (Romary et al., 2015). In contrast to supervised learning which requires labeled inputs and produces predictions based on training datasets, the main objective of unsupervised learning is to identify interesting patterns or features in the datasets. This is also known as data exploration (Murphy, 2012) and one of its major advantages is that no *a priori* knowledge on the data is required, and the process could be fully automated (Ripley, 1996).

The ocean drilling datasets from the South China Sea (SCS) are adopted to apply the chosen unsupervised learning methods. The SCS datasets obtained over the past decades have been widely recognized as one of the best datasets in the world to study paleoclimate and the region's geological past (Wang and Li, 2009). Only a relatively small amount of the datasets had been analyzed and published since traditional methods of data analysis often involve too much manual processing and are not efficient in processing large volumes of data. The goal of this study is to apply the latest unsupervised learning methods to a

Received September 17, 2017; accepted February 6, 2018

E-mail: tsekevin@hku.hk

small part of the datasets so that previously unknown information could be extracted from the datasets. We hope that such new information could be useful in the characterization of the tectonic and sedimentation history of the SCS.

2 Data

The Ocean Drilling Program (ODP) and Integrated Ocean Drilling Program (IODP) are long-term international scientific endeavours to explore the floors of the world's oceans by drilling and collecting drill cores for scientific analyses since the 1970s. A vast amount of high quality geophysical and geochemical data has been generated from hundreds of kilometers of sediment cores obtained from expeditions around the globe. Aggregated ODP/IODP datasets have been made accessible online, opening up opportunities to employ statistical learning techniques involving these large multivariate datasets to reveal previously hidden information.

The four SCS sites consist of ODP sites 1146 and 1148 which were drilled between February and April of 1999 during expedition leg 184. The other two sites are U1431 and U1433, drilled in expedition leg 394 between January and March of 2014. SCS is a marginal sea in the western Pacific, at the junction of the Eurasian, Pacific and Indo-Australian plates. The SCS sites are chosen for data discovery by unsupervised methods because of the exceptional continuity in length and age of core samples (Li et al., 2014). Supervised learning methods have previously been applied on the datasets to infer the lithology of missing cores (Benaouda et al., 1999). Unsupervised methods, on the other hand, have yet to be explored on these datasets. The locations of the chosen ODP/IODP drill sites are shown in Fig. 1 and summarized in Table 1.

Geophysical variables are chosen from the datasets for this study (16 out of 20 in total) due to their higher sampling frequency along the drilling cores and continuity over the entire depth range. In order to produce a synthetic dataset containing all variables, down-sampling is required as the datasets are unbalanced, meaning that the sampling data available for each variable are different. Down-sampling refers to a process in which all the variables in the datasets are binned to the same number of samples as the variable with the smallest number of samples (Hamel, 2009). The Moisture and Density (MAD) dataset is the one

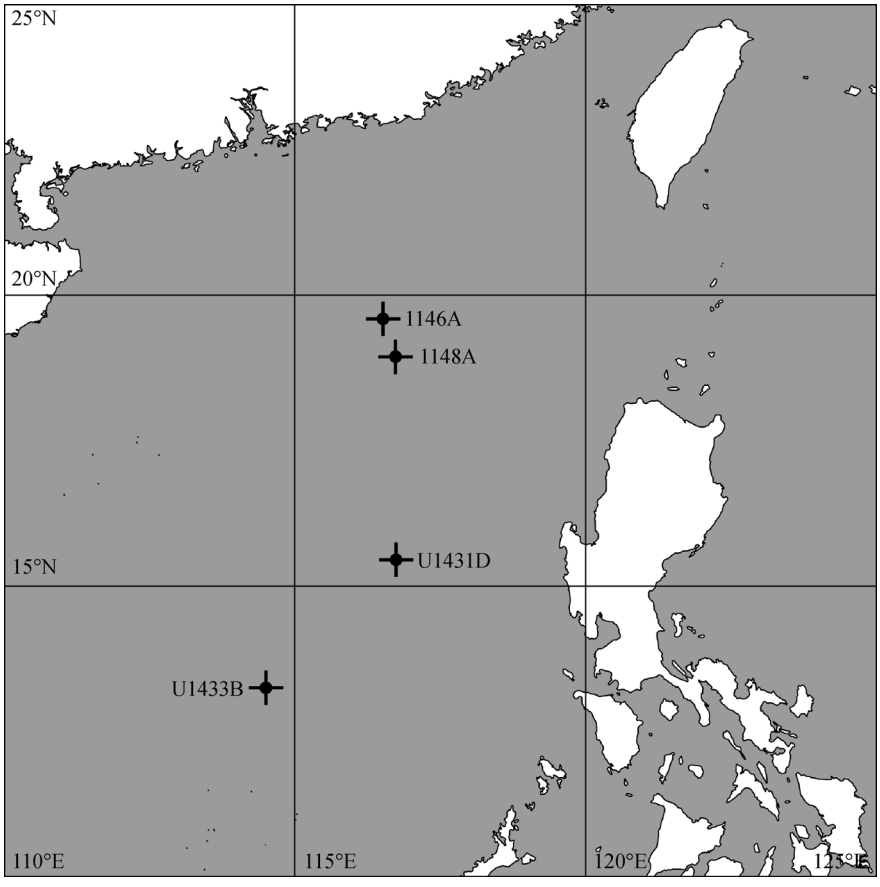

Fig. 1 Map showing locations of the four ocean drill sites used in this study.

Table 1 Summary of the datasets of the four ODP/IODP sites

Date	Leg	Site	Latitude	Longitude	Drilled depth/m	Water depth/m	Core recovery	Age at base/Ma	Ref.
Feb–Apr 2000	146	1146	19°27.4′N	116°16.37′E	603.5	2091.7	95%	15	Moore et al., 2001; Wang and Li, 2009
Feb–Apr 2000	146	1148	18°50.2′N	116°33.93′E	853.2	3291.8	98%	13	Moore et al., 2001; Wang and Li, 2009
Jan–Mar 2014	349	U1431	15°22.5′N	117°00.00′E	617.0	4240.5	100%	16	Wang and Li, 2009; Li et al., 2014
Jan–Mar 2014	349	U1433	12°55.1′N	115°02.85′E	858.5	4379.3	96%	26	Wang and Li, 2009; Li et al., 2014

with the least number of sample data points in the selected geophysical datasets. Table 2 summarizes the geophysical variables extracted from the four SCS drill sites.

3 Methods

3.1 Unsupervised learning

Being a branch of machine learning, unsupervised learning, also known as clustering, is an exploratory data analysis technique used for identifying similar groups (clusters) that satisfy a pre-defined criteria for similarity in the datasets of interest (Romary et al., 2015). For the purpose of this study, four different unsupervised learning methods, K-means, self-organizing maps, hierarchical clustering, and random forest will be used as a "black box" that takes the input dataset X and finds a function f: $\mathrm{R}^N \rightarrow \mathrm{R}^K$ that maps an input vector $x^{(i)}$ to a new feature vector of K clusters. Unlike in supervised learning, there are no predicted values Y to be considered. This study uses three common distance measures for assessing similarity between data points. Euclidean distance, d_{euc} is defined as $d_{\mathrm{euc}}(x,y) = \sqrt{\sum_{i=1} (x_i - y_i)^2}$. The second one is Manhattan distance (also known as taxicab metric) (Krause, 1987) is defined as $d_{\mathrm{man}}(x,y) = \sum_{i=1} |x_i - y_i|$. The third distance measured is Chebyshev distance (also known as maximum metric) (Cantrell, 2000) defined as $d_{\mathrm{che}}(x,y) = \max|x_i - y_i|$.

Before being input into the unsupervised learning methods, the selected datasets are processed to remove any missing or incorrect values. Outliers are kept to preserve the completeness of the datasets and three normalization methods are tested for the pre-processing of the data to transform the raw data values into a comparable format. The three methods are statistical standard score ($x' = \frac{x - \overline{x}}{\sigma_x}$), unity based scaling ($x' = \frac{x - \min(x)}{\max(x) - \min(x)}$), and log transformation (x' = ln (x– min(x) + 1)), which are commonly used for data pre-processing in machine learning pipelines (Way et al., 2012).

Although it may seem desirable to perform cluster analysis with all available observations and variables (Templ et al., 2008), including any irrelevant variables may adversely impact the desired clustering results (Templ et al., 2008). Clustering will be performed on individual datasets in addition to the synthetic dataset containing all the datasets in order to reveal the effects of different selections of variables on the clustering results.

3.1.1 K-means

The K-means clustering algorithm proposed by MacQueen (1967) is one of the simplest unsupervised learning algorithms commonly used to solve clustering problems

Table 2 Summary of input datasets in this study

Dataset	Extracted Variables	Sampling Frequency
MAD (Moisture andDensity)	Water content (bulk), water content (dry), bulk density (g/cc), dry density (g/cc), grain density (g/cc), porosity (%)	1.5 m
RSC (Reflectance Spectrophotometry and Colorimetry)	Reflectance values at 400, 450, 500, 550, 600, 650, 700 nm in % intensity (For IODP sites, L', a', b', tristimulus (X,Y,Z) are used instead)	4 cm
MSL (Magnetic Sus-ceptibility)	Drift-corrected suscept. (inst. units)	5 cm
NGR (Natural Gamma Radiation)	Bkg-corrected counts (cps)	5 cm
GRA (Gamma Ray Attenuation)	Density (g/cc)	5 cm

(Chauhan et al., 2016). As a partitioning method, the number of resulting clusters, k, is pre-determined. Data are decomposed into a set of non-overlapping k clusters by initializing k centroid centers and then refining the cluster centers by iteration. The method is capable of handling large datasets with continuous data and the absence of non-convex clusters (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 $C = C_1,...,C_K$, where $C_k \cap C_{k'} = \varnothing$, centered at cluster means μ_j for $j = 1,...,k$, based on the initial conditions. K-means seeks a clustering results where within-cluster variation $W(C_k)$ is a minimum:

$$W(C_k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2,$$

$$\text{where } \mu_j = \frac{\sum_{x_i \in C_j} x_i}{|C_j|}. \tag{1}$$

3.1.2 Hierarchical clustering

Hierarchical clustering (HC) is a classical unsupervised learning method. Unlike K-means which requires the number of clusters as the input in the algorithm, HC does not. There are a number of different types of HC including complete linkage, single linkage, mean linkage, and centroid linkage. In this study, the average linkage method is adopted, in which the mean linkage clustering which finds all possible pairwise distances for points belonging to two different clusters is calculated:

$$\text{Distance between clusters} = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x,y). \tag{2}$$

The process starts with a single cluster containing all data points and with each clustering step, the distance between clusters A and B given by Eq. (2) is evaluated (Murphy, 2012). The "top-down" approach generates splitting nodes recursively as the hierarchy is moved downwards.

3.1.3 Self-organizing maps (SOMs)

Self-organizing maps (SOMs), proposed by Kohonen (1982) are widely adopted as a versatile unsupervised learning algorithm based on neural networks (Kohonen, 1982; Ripley, 1996; Augustijn and Zurita-Milla, 2013). In a sense, SOMs (Kohonen, 1982, 2001) can be thought of as a spatially constrained form of K-means clustering (Ripley, 1996; Wehrens and Buydens, 2007), they have been shown to be useful for identifying, visualizing, and analyzing coherent groups within multivariate geoscience data (Penn, 2005; Peeters et al., 2007; Bierlein et al., 2008; Bedini, 2009, 2012).

A self-organizing map (SOM) first arranges the neurons in a grid topology, then uses a distance metric to determine the positions of the neurons in the topology (Chauhan et al., 2016). A winner node, or best matching unit (BMU) will emerge as the competitive learning process is performed iteratively. All the neurons in a defined neighborhood around the winner node are defined as a cluster using the Kohonen rule (Kohonen, 2001):

$$|x - m_c| = \min|x - m_i|. \tag{3}$$

SOM nodes are trained from randomly sampled reference vectors m_i of equal length to n, via an iterative two-stage process. Seed-factors are shown to the network and compared to any $x_n \in R^n$ that falls within a distance metric. Depending upon whether a neuron i is within a certain spatial neighborhood $N_t(l)$ around t, its weight is updated according to Warren Liao (2005), where

$$w_i(l+1) = \begin{cases} w_i(l) + \alpha(l)[x(l) - w_i(l)] & \text{if} \quad i \in N_t(l) \\ w_i(l) & \text{if} \quad i \notin N_t(l) \end{cases}. \tag{4}$$

3.1.4 Random forest

Random forests (RF) is an ensemble method that utilizes a majority vote to predict classes based on the partition of data from multiple decision trees. In a random forest, 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 to determine the "best split" threshold of input values (p_i stands for the probability of class i at node n_c) for a given class:

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i). \tag{5}$$

The Gini index returns a measure of class heterogeneity within child nodes as compared to the parent node (Breiman, 1984). Instead of using one of the three distance metrics 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; Insua et al., 2015; Goetz et al., 2015).

3.2 Cluster validation

The performance of unsupervised learning can be validated externally using some known ground truth on the datasets (Halkidi et al., 2002). In this study, the clustering results

are validated by measuring the correspondence between the clustering partition and the classification assigned by lithologic units and geologic time scales for ocean drilling cores. In this study, two forms of Rand Index (RI) (Rand, 1971) will be adopted. Given two partitions on set $S = \{O_1,...,O_n\}$ containing n objects, $U = \{u_1,...,u_R\}$ and $V = \{v_1,...,v_C\}$, where $u_i \cap u_{i'} = v_j \cap v_{j'} = \varnothing$, $1 \leqslant i \neq i' \leqslant R$ and $1 \leqslant j \neq j' \leqslant C$. The information on class overlap between two partitions U and V can be expressed in the form of a contingency table (Table 3) where n_{ij} denotes the number of objects that are common to classes u_i and v_j, while $n_{i\cdot}$ and $n_{\cdot j}$ denote the sum of each row or column.

Table 3 Notation for comparing two partitions (Hubert and Arabie, 1985), also referred as contingency table

Class	v_1	v_2	$\cdots$	vC	Sums
u_1	$n11$	$n12$	$\cdots$	$n1C$	$n1\cdot$
u_2	$n21$	$n22$	$\cdots$	$n2C$	$n2\cdot$
			$\ddots$		
u_R	$nR1$	$nR2$	$\cdots$	nRC	$nR\cdot$
Sums	$n\cdot 1$	$n\cdot 2$	$\cdots$	$n\cdot C$	$n_{\cdot\cdot} = n$

RI_1, the unadjusted Rand Index, is defined as $\frac{A}{A+D}$, where $A = \binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2\right)$ is the total number agreements (e.g., objects from S are placed in the same class in U and V) and $D = \frac{1}{2}\left(\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2\right) - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2$ is the total number of disagreements (Rand, 1971). Two partitions that are similar produce relatively large values of A and small values of D. In other words, an RI_1 value close to 1 implies a relatively higher similarity between the two partitions. In order to correct the values of A and D for chance, it can be shown that an adjusted RI, denoted as RI_2 can be defined as $\frac{I-EI}{MI-EI}$, where $I = \sum_{i,j=1}^{R,C} \binom{n_{ij}}{2}$ is the calculated index, $EI = \sum_{i=1}^{R} \binom{n_{i\cdot}}{2} \sum_{j=1}^{C} \binom{n_{\cdot j}}{2} / \binom{n}{2}$ is the expected index and $MI = \frac{1}{2}\left(\sum_{i=1}^{R} \binom{n_{i\cdot}}{2} + \sum_{j=1}^{C} \binom{n_{\cdot i}}{2}\right)$ is maximum index (Hubert and Arabie, 1985). Since EI can be larger than I in some cases, the value of RI_2 ranges from −1 to 1.

In this study, the number of clusters used in the unsupervised learning was specified to be equal to the number of lithologic units or geologic time scales assigned to the drill core samples. The determination of the number of clusters has been a subject of debate for unsupervised learning (Hennig, 2015) and the number of clusters used in the study has been checked with their within-cluster sum of squares (WSS) ($\sigma^2 = \frac{1}{N}(x-\bar{x})^2$). WSS calculated for different numbers of clusters are checked so that a higher number of clusters do not decrease the cluster variance measure significantly.

Internal validity metrics commonly used for finding an optimal number of clusters (Baarsch and Celebi, 2012), namely the Davies-Bouldin (DB) index and Silhouette Index (SI) are not used since the number of clusters are already fixed with reference to the external validation in this study.

4 Results

The clustering results consist of two similar sets by comparing their correspondence with lithologic units and geologic time scales assigned on the datasets. Each set was produced by applying the selected unsupervised learning methods coupled with three different distance metrics (except RF for which its own proximity distance measure was used) on the ODP and IODP datasets.

Table 4 lists the values of RI_1 and RI_2 indicating the correspondence between the cluster results and assigned lithologic units on the ocean drilling cores. The highest RI_1 and RI_2 values calculated for the four SCS sites studied are 0.832/0.584, 0.869/0.503, 0.839/0.357, and 0.697/0.282 for sites 1146, 1148, U1431, and U1433, respectively. K-means and the SOM appear to be better at producing higher RI values than the other two methods when predicting the lithologic units. Site U1431 is an exception with the RF performing better.

Table 5 shows RI_1 and RI_2 for comparison with assigned geologic time scale to the datasets. The values are generally lower than those in Table 4. The highest RI_1 and RI_2 values recorded are 0.836/0.543, 0.861/0.435, 0.731/0.425, and 0.706/0.254 for sites 1146, 1148, U1431, and U1433, respectively. Out of the four unsupervised methods attempted, RF appears to fare better than the other three methods in producing the closest correspondence with assigned geologic time scales.

Tables 6 and 7 show the results of unsupervised clustering for individual datasets, obtained from K-means with the Euclidean distance metric. A common result displayed for the four sites is that MAD produces the highest RI values among all datasets, while NGR produces the lowest results. It is worth noting that RI values obtained by combining different datasets are mostly higher than the values obtained from any single dataset.

Another important observation made in the study is that

Table 4 Validation of clustering results for the four SCS sites with lithological units

Method	Index	1146			1148			U1431			U1433		
		d_{euc}	d_{man}	d_{che}	d_{euc}	d_{man}	d_{che}	d_{euc}	d_{man}	d_{che}	d_{euc}	d_{man}	d_{che}
K-means	RI_1	0.817	0.832	0.782	0.859	0.837	0.860	0.820	0.824	0.800	0.636	0.696	0.634
	RI_2	0.551	0.584	0.447	0.455	0.382	0.452	0.300	0.298	0.192	0.214	0.274	0.213
H.C.	RI_1	0.727	0.832	0.340	0.686	0.683	0.649	0.312	0.410	0.419	0.369	0.369	0.369
	RI_2	0.464	0.583	0.027	0.234	0.222	0.272	0.046	0.072	0.049	0.071	0.071	0.071
SOMs	RI_1	0.817	0.832	0.719	0.868	0.798	0.869	0.815	0.817	0.776	0.635	0.697	0.633
	RI_2	0.551	0.584	0.339	0.497	0.306	0.503	0.312	0.307	0.194	0.211	0.282	0.215
RF	RI_1	0.832	-	-	0.811	-	-	0.839	-	-	0.694	-	-
	RI_2	0.577			0.246	-	-	0.357	-	-	0.179	-	-

Table 5 Validation of clustering results for the four SCS sites with geological time scales

Method	Index	1146			1148			U1431			U1433		
		d_{euc}	d_{man}	d_{che}	d_{euc}	d_{man}	d_{che}	d_{euc}	d_{man}	d_{che}	d_{euc}	d_{man}	d_{che}
K-means	RI_1	0.835	0.840	0.829	0.835	0.813	0.831	0.616	0.707	0.580	0.593	0.586	0.599
	RI_2	0.381	0.554	0.520	0.381	0.320	0.368	0.188	0.381	0.098	0.189	0.163	0.214
H.C.	RI_1	0.708	0.768	0.544	0.447	0.444	0.600	0.430	0.432	0.433	0.375	0.375	0.375
	RI_2	0.405	0.406	0.166	0.081	0.073	0.169	0.050	0.044	0.016	0.074	0.074	0.074
SOMs	RI_1	0.773	0.791	0.722	0.838	0.816	0.833	0.573	0.578	0.461	0.591	0.588	0.599
	RI_2	0.386	0.447	0.309	0.389	0.345	0.375	0.118	0.130	0.003	0.190	0.163	0.217
RF	RI_1	0.836	-	-	0.861	-	-	0.731	-	-	0.706	-	-
	RI_2	0.543	-	-	0.435	-	-	0.425	-	-	0.254	-	-

Table 6 Clustering results for individual datasets for lithological units

Dataset	Index	1146	1148	U1431	U1433
MAD	RI_1	0.74	0.813	0.787	0.682
	RI_2	0.363	0.28	0.199	0.269
RSC	RI_1	0.66	0.778	0.769	0.643
	RI_2	0.243	0.145	0.12	0.103
MSL	RI_1	0.622	0.796	0.697	0.593
	RI_2	0.117	0.233	0.091	0.174
GRA	RI_1	0.679	0.748	0.78	0.59
	RI_2	0.231	0.102	0.154	0.049
NGR	RI_1	0.593	0.688	0.808	0.633
	RI_2	0.122	0.001	0.28	0.125

All *RI* are calculated with K-means and Euclidean distance.

Table 7 Clustering results for individual datasets for geological time scales

Dataset	Index	1146	1148	U1431	U1433
MAD	RI_1	0.818	0.808	0.612	0.673
	RI_2	0.53	0.295	0.197	0.313
RSC	RI_1	0.659	0.762	0.544	0.624
	RI_2	0.222	0.123	0.032	0.108
MSL	RI_1	0.685	0.763	0.445	0.558
	RI_2	0.233	0.163	-0.086	0.188
GRA	RI_1	0.702	0.767	0.584	0.475
	RI_2	0.253	0.156	0.135	0.027
NGR	RI_1	0.601	0.683	0.652	0.607
	RI_2	0.122	0.007	0.258	0.114

All *RI*s are calculated with K-means and Euclidean distance.

results obtained from dataset values before and after normalization differ significantly, as indicated in Table 8. In fact, results from all three distance metrics studied display the same trend. For instance, while RI_1 for unnormalized datasets for site 1146 is 0.650, the resulting RI_1 values for the three experimented normalization methods are 0.829, 0.828, and 0.817. For consistency and ease of comparison, log transformation was chosen as the normalization method for the input datasets in this study.

In Figs. 2 and 3, clustering results with the highest RI_1 and RI_2 are plotted against depth measured in meters below the sea floor (mbsf) as the *y*-axis. Each data point represents a down-sampled multi-variate observation in the synthetic dataset and the red lines represent the

Table 8 Clustering results for different normalization methods on lithologic unit

Normalization	Index	1146	1148	U1431	U1433
Original	RI_1	0.65	0.769	0.668	0.525
	RI_2	0.233	0.21	0.098	0.097
$x' = \frac{x-\overline{x}}{\sigma_x}$	RI_1	0.829	0.808	0.826	0.697
	RI_2	0.568	0.272	0.329	0.236
$x' = \frac{x-\min(x)}{\max(x)-\min(x)}$	RI_1	0.828	0.846	0.82	0.622
	RI_2	0.567	0.403	0.3	0.168
$x' = \ln(x-\min(x)+1)$	RI_1	0.817	0.859	0.815	0.636
	RI_2	0.551	0.455	0.267	0.214

All *RI*s are calculated with K-means and Euclidean distance.

boundaries of different lithologic or geologic time scale units assigned to the ODP/IODP ocean floor sediment cores. Compact and connected clusters are observed for clustering results on ODP site 1146, in Figs. 2(a) and 2(b). The overlapping of clusters is minimal compared with other sites (i.e., different depth ranges are assigned with different clusters), indicating that the unsupervised clustering has successfully sorted the data into unambiguous cluster segments. The last cluster segment in Fig. 2(a) is terminated almost exactly on the boundary between Unit I and Unit IIA (222.68 mbsf) while the last cluster segment shown in Fig. 2(b) starts at the boundary between Pliocene and late Miocene (~300 mbsf). Figures 2(c) and 2(d) present clustering results for ODP site 1148, with larger numbers of lithological units and geological time scales compared with ODP site 1146. The cluster segments are less compact and connected than those of 1146, but some of the starting and terminating positions of the cluster segments show a remarkable agreement with the various boundaries assigned by scientists. In Fig. 2(c), two cluster segments are observed to overlap with lithologic Unit II (from 181.8 to 316.6 mbsf) and Unit IV (from 348 to 400 mbsf). In Fig. 2(d), one cluster segment is seen to overlap with the epoch of early Miocene (from 350 to 460 mbsf).

Figures 3(a) and 3(b) present clustering results for the IODP site 1431. The number of lithologic units is the highest among all four SCS sites studied, and the drill cores are more geological complex than other sites. In Fig. 3(a), two cluster segments are observed to stretch the ranges of lithologic Unit VI to Unit VIII (from 603.42 to 885.25 mbsf) while other cluster segments do not demonstrate a very well-defined agreement with the lithologic unit boundaries. In Fig. 3(b), a cluster segment terminates around the Pliocene/late Miocene boundary (~300 mbsf). The results of this site are the only one for which the RI values for geologic time scales are higher than those of lithologic units. Figures 3(c) and 3(d) show clustering results for IODP site 1433. In Figure 3(c), three cluster segments run through multiple lithologic units while one cluster segment is observed to overlap perfectly on Unit IIB (from 551.32 to 747.93 mbsf). In Fig. 3(d), two lengthy cluster segments terminate approximately at the Pliocene/late Miocene boundary (~750 mbsf). In this site, more than one cluster is observed to be assigned to the same depth range.

5 Discussion

The objective of the study is to apply unsupervised machine learning methods onto the SCS ocean drilling datasets which are widely regarded as some of the most comprehensive in the world. While traditional analysis methods always involve substantial manual and expert judgements, machine learning methods can process entire datasets automatically, and possibly extract new information from existing datasets for new insights.

Results from the K-means and SOM methods demonstrate higher degrees of correspondence with lithological unit boundaries, while the UL method showing the highest degree of correspondence with geologic time scales is random forest. The higher performance of K-means and SOM for classifying lithological units may be due to the fact that the K-means clustering produces Voronoi diagrams (Murphy, 2012) which consists of linear decision boundaries or hyperplanes in the geophysical multivariate space. This is in line with the understanding that the lithological units are highly correlated to the geophysical variables selected in this study. On the other hand, since geologic time units are determined by more factors and some of them are not directly correlated to the geophysical variables, the underlying relationship is non-linear and hence the classification is better handled by the decision trees of the random forest. Insua et al. (2015) also demonstrated that non-linear methods such as RF are able to establish non-linear relations among measured variables in predicting lithologies in carbonate sediments.

For results from all UL methods, RI_1 and RI_2 values for lithologic units are generally higher than those of geologic time scales. This can be explained by the fact that determination of lithologic units based on mineralogy and other physical features is more unambiguous. For instance, for site 1146, lithologic unit I is a bioturbated clay with plenty of microfossils, while unit II is a calcite rich layer much whiter in color. The boundary between the two is physically, if not visually well defined. The case for geologic time scales is much more different. The determination of the geologic age for different strata is based on many proxy measures including magnetostratigraphy, biostratigraphy and radiostratigraphy, each one with its own shortcoming and margin of errors. Experts of different domains may have different opinions on the exact age of a strata and an agreement on an age boundary is sometimes hard to be reached.

RI values for site U1433 are only as low as half of the highest ones from the other three sites. U1433 is located at

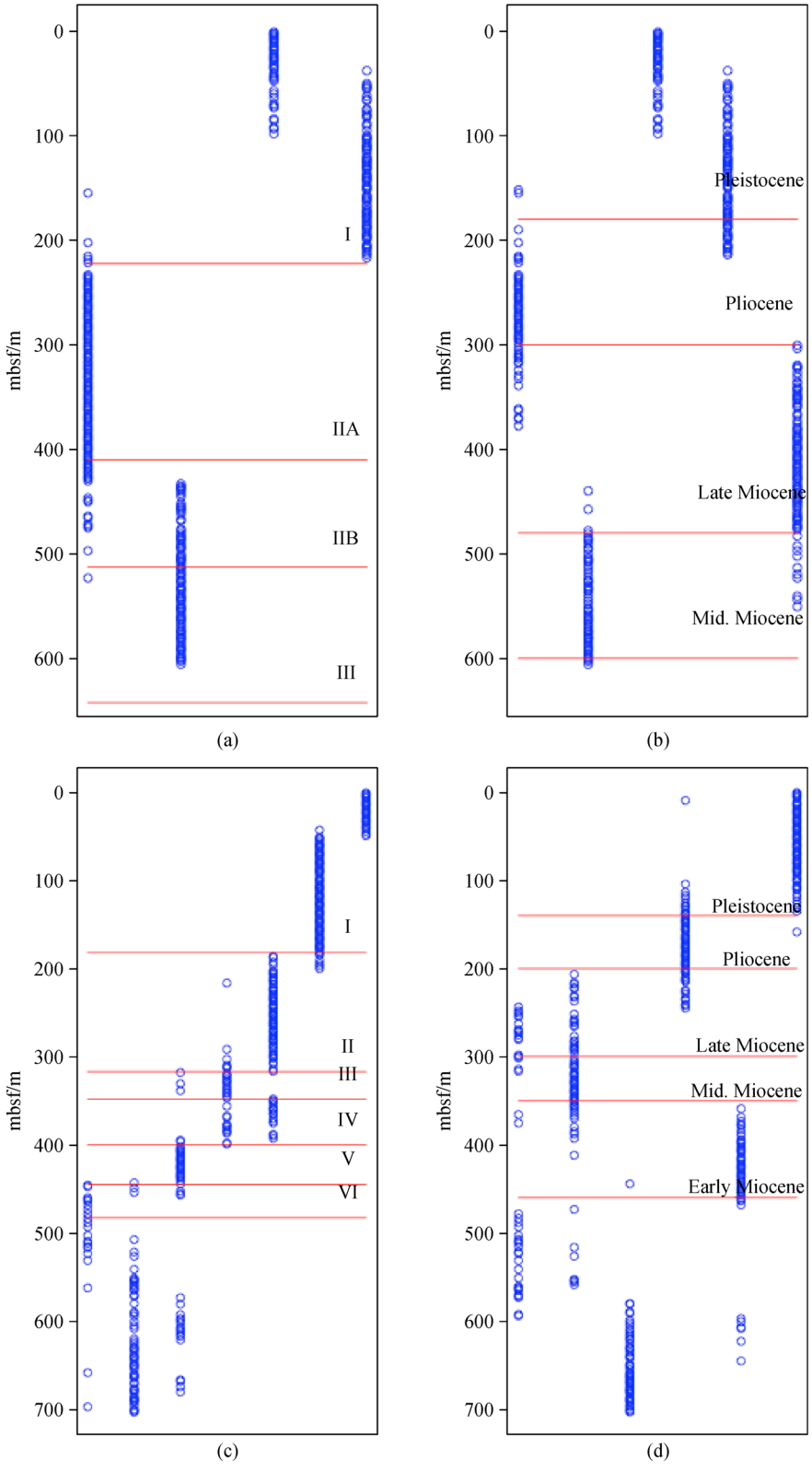

Fig. 2 Unsupervised clustering results compared with lithological units and geological time scales for the ODP sites 1146 and 1148. In calculating the RI, the number of clusters are set to equal to the number of lithological units or number of geological time scales. (a) 1146 K-means lith. units (cluster = 4, RI_1 = 0.832, RI_2 = 0.584); (b) 1146 K-means geo. units (cluster = 5, RI_1 = 0.840, RI_2 = 0.554); (c) 1148 SOMs lith. units (cluster = 7, RI_1 = 0.869, RI_2 = 0.503); (d) 1148 RF geo. units (cluster = 6, RI_1 = 0.861, RI_2 = 0.435).

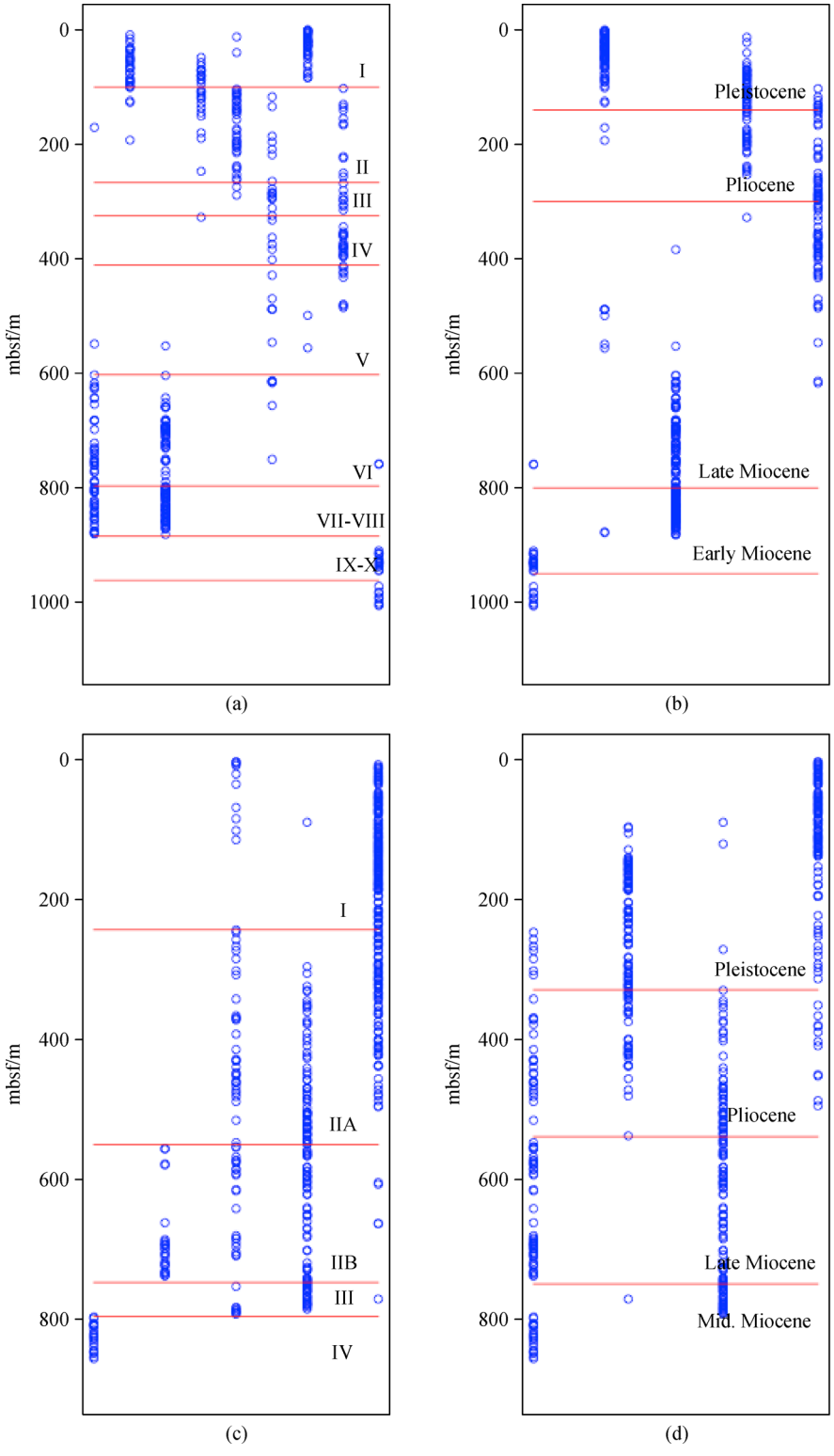

Fig. 3 Unsupervised clustering results compared with lithological units and geological time scales for the IODP sites U1431 and U1433. In calculating the RI, the number of clusters is set to equal the number of lithological units or number of geological time scales. (a) U1431 RF lith. units (cluster = 9, RI_1 = 0.839, RI_2 = 0.357); (b) U1431 RF geo. units (cluster = 5, RI_1 = 0.731, RI_2 = 0.425); (c) U1433 SOMs lith. units (cluster = 5, RI_1 = 0.697, RI_2 = 0.282); (d) U1433 RF geo. units (cluster = 4, RI_1 = 0.706, RI_2 = 0.254).

a relic spreading center and the geological complexity may be a factor affecting the clustering results. Other variables such as geochemical or mineralogical datasets might be needed to produce better clustering results.

A characteristic of unsupervised learning is that the results cannot be validated directly by training data (Romary et al., 2015). The UL method is useful in revealing the underlying structure of the datasets but not all of these data patterns may be of scientific interest since the structures may not indicate thematic representations (e.g., lithologic units or geologic time scales). A high value of RI_1 and RI_2 indicate that the correspondence between the unsupervised clustering results and the lithologic classification or geological time scale classification schemes is significant, but there is no scientific measurement of whether the clustering results are "correct" or not, since in unsupervised learning there is no real "ground truth".

Nonetheless, the results are still remarkable as under completely *a priori* information, these algorithms are able to sort the data into homogeneous groups with varying degree of resemblance to classification schemes carried out by conventional methods, usually involving a large amount of manual work and expert interpretation. The results are more than statistical coincidence and the clustering results do reveal some fundamental structure of the datasets not directly visible to human perception using traditional manual data interpretation methods.

6 Conclusions

In this study, four popular unsupervised machine learning methods, namely K-means, self-organizing maps, hierarchical clustering and random forest have been applied on scientific ocean drilling data from the SCS. The objective of demonstrating that these machine learning methods are able to produce classification results comparable to results obtained by traditional methods has been achieved. Compact and connected exploratory data clusters formed using the machine learning methods have shown varying degrees of correspondence with existing classification by lithologic units and geologic time scales. Results by K-means and SOM performed well with lithologic units and RF corresponded best with geologic time scales. The results have demonstrated that such methods are capable of auto-processing vast amounts of data and uncover interesting information externally validated by traditional classification methods. Further studies should be conducted with other learning methods such as deep learning and include datasets from more sites involving more variables.

References

Augustijn E W, Zurita-Milla R (2013). Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. Int J Health Geogr, 12(1): 60

Baarsch J, Celebi M (2012). Investigation of internal validity measures for k-means clustering. In: Proceedings of the International Multi Conference of Engineers and Computer Scientists

Bedini E (2009). Mapping lithology of the Sarfartoq carbonatite complex, southern West Greenland, using HyMap imaging spectrometer data. Remote Sens Environ, 113(6): 1208–1219

Bedini E (2012). Mapping alteration minerals at Malmbjerg molybdenum deposit, central East Greenland, by Kohonen self-organizing maps and matched filter analysis of HyMap data. Int J Remote Sens, 33(4): 939–961

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

Bierlein F P, Fraser S J, Brown W, Lees T (2008). Advanced methodologies for the analysis of databases of mineral deposits and major faults. Aust J Earth Sci, 55(1): 79–99

Breiman L (1984). Classification and Regression Trees. New York: Chapman & Hall, 87–91

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

Cantrell C D (2000). Modern Mathematical Methods for Physicists and Engineers. Cambridge University Press

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

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(2): 287–304

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

Halkidi M, Batistakis Y, Vazirgiannis M (2002). Clustering validity checking methods: part II. ACM SIGMOD Rec, 31(3): 19–27

Hamel L (2009). Knowledge Discovery with Support Vector Machines. New York: John Wiley and Sons, 89–132

Hennig C (2015). What are the true clusters? Pattern Recognit Lett, 64: 53–62

Hubert L, Arabie P (1985). Comparing partitions. J Classif, 2(1): 193–218

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(2): 590–606

Jeong J, Park E (2016). Comparative Application of Various Machine Learning Techniques for Lithology Predictions. J Soil Groundw Environ, 21(3): 21–34

Kabacoff R I (2015). R in Action- Data analysis and graphics with R. Greenwich, CT: Manning, 102–112

Kohonen T (1982). Self-organized formation of topologically correct feature maps. Biol Cybern, 43(1): 59–69

Kohonen T (2001). Self-Organizing Maps (3rd ed). New York: Springer, 132–154

Krause E F (1987). Taxicab Geometry- An Adventure in Non-Euclidean

Geometry. Stroud, UK: Dover, 120–132

Lary D J, Alavi A H, Gandomi A H, Walker A L (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7 (1): 3–10

Li C F, Lin J, Kulhanek D K (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. Technical report

Longo G, Brescia M, Djorgovski S, Cavuoti S, Donalek C (2014). Data driven discovery in astrophysics. Proceedings of ESA-ESRIN Conference: Big Data from Space 2014, Frascati, Italy

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. University of California, 281–297

Marzo G A, Roush T L, Blanco A, Fonti S, Orofino V (2006). Cluster analysis of planetary remote sensing spectral data. Journal of Geophysical Research, 111: E03002

Moore G, Taira A, Klaus A, Becker K, Saffer M, Screaton E (2001). Proc. ODP, Init. Repts., 190. College Station, TX (Ocean Drilling Program)

Murphy K P (2012). Machine Learning A Probabilistic Perspective. Cambridge: The MIT Press, 578–490

Peeters L, Bação F, Lobo V, Dassargues A (2007). Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen's self-organizing map. Hydrol Earth Syst Sci, 11(4): 1309–1321

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

Pham B T, Khosravi K, Prakash I (2017b). Application and comparison of decision tree-based machine learning methods in landside susceptibility assessment at Pauri Garhwal area, Uttarakhand, India. Environmental Processes, 2017, 4(3): 711–730

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). Journal of the Indian Society of Remote Sensing, 45: 673–683

Rand W M (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336): 846–850

Ripley B D (1996). Pattern Recognition and Neural Networks. Cambridge University Press, 248–290

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

Schnase J L, Lee T J, Mattmann C A, Lynnes C S, Cinquini L, Ramirez P M, Hart A F, Williams D N, Waliser D, Rinsland P, Webster W P, Duffy D Q, McInerney M A, Tamkin G S, Potter G L, Carriere L (2016). Big data challenges in climate science. IEEE Geosciences and Remote Sensing, 4(3): 10–22

Templ M, Filzmoser P, Reimann C (2008). Cluster analysis applied to regional geochemical data: problems and possibilities. Appl Geochem, 23(8): 2198–2213

Wang P X, Li Q Y (2009). The South China Sea Paleoceanography and Sedimentology. New York: Springer, 388–421

Warren Liao T (2005). Clustering of time series data- a survey. Pattern Recognit, 38(11): 1857–1874

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, 240–312

Wehrens R, Buydens L M C (2007). Self- and super-organising maps in R: the Kohonen package. Journal of Statistical Software, 21(5):1–19

Xiong Y, Zuo R (2016). Recognition of geochemical anomalies using a deep autoencoder network. Comput Geosci, 86: 75–82

Yao X, Tham L G, Dai F C (2008). Landslide susceptibility mapping based on Support Vector Machine: a case study on natural slopes of Hong Kong, China. Geomorphology, 101(4): 572–582