Report: Using unsupervised learning to predict lithologies based on geophysical well log data

王驭捷 Yujie Wang

1700012781

Introduction

Lithology prediction has always been the focus of geoscience researchers, as it is one of the significant problems in reservoir characterization. Usually, well log data are used to predict lithology. Well-logs are measurement data recorded in a well and contain the physical properties of the subsurface, such as moisture, density, gamma radiation and etc. These physical properties are great lithology indicators. In this work, unsupervised learning method, K-means, is applied to geophysical well log data in order to predict lithology. Different from other work using supervised learning methods, the unsupervised model in this work does not require any priori information and only needs input features. The clustering result of the K-means model is validated by lithologic units and geologic time scales that are given by conventional methods. The experiment shows that the predicted clusters given by the model highly correspond to the existing result of lithologic units and geologic time scales, thus the model reaches great performance. Such powerful unsupervised model has significant practical meaning as it gives reliable lithology prediction based on well log data, and helps humans to understand the underground structure in a better way. Besides, it may benefit some industry, such as oil extraction.

Methods

Dataset

In this work, well log data collected from scientific ocean drilling in the South China Sea is used. These well-logs are publicized on Integrated Ocean Drilling Program (IODP) website [1]. Specifically, the following datasets from the well-logs of site U1431 from the expedition 349 are used in this work: Gamma Ray Attenuation (GRA), Magnetic Susceptibility (Pass-through) (MS), Moisture and Density (MAD), Natural Gamma Radiation (NGR), Reflectance Spectroscopy and Colorimetry (RSC).

Among the above five datasets, the following 15 features are selected, just as Tse et al. did in [2]: Water content (bulk), water content (dry), bulk density, dry density, grain density, porosity from MAD dataset; L’ , a’ , b’ , tristimulus (X,Y,Z) from RSC dataset; Drift-corrected susceptibility from MSL dataset; Bkg-corrected counts from NGR; Density from GRA.

Down-sampling is applied to these five datasets to create a synthetic dataset that contain all variables. The MAD dataset is the one with the least number of datapoints, so all other four

datasets are down-sampled to the same size as MAD. Down-sampling process is done by pandas function merge_asof with direction=' nearest' to fill missing values with nearest neighbor. All data are sorted by the common depth scale 'Depth CSF-B(m)' and the items with missing values are deleted. Then, the outliers are eliminated from the dataset and the features are normalized by standard scaler, and the size of final input features is [370, 15].

The python package pandas is used to preprocess the data and prepare the input features.

Algorithm

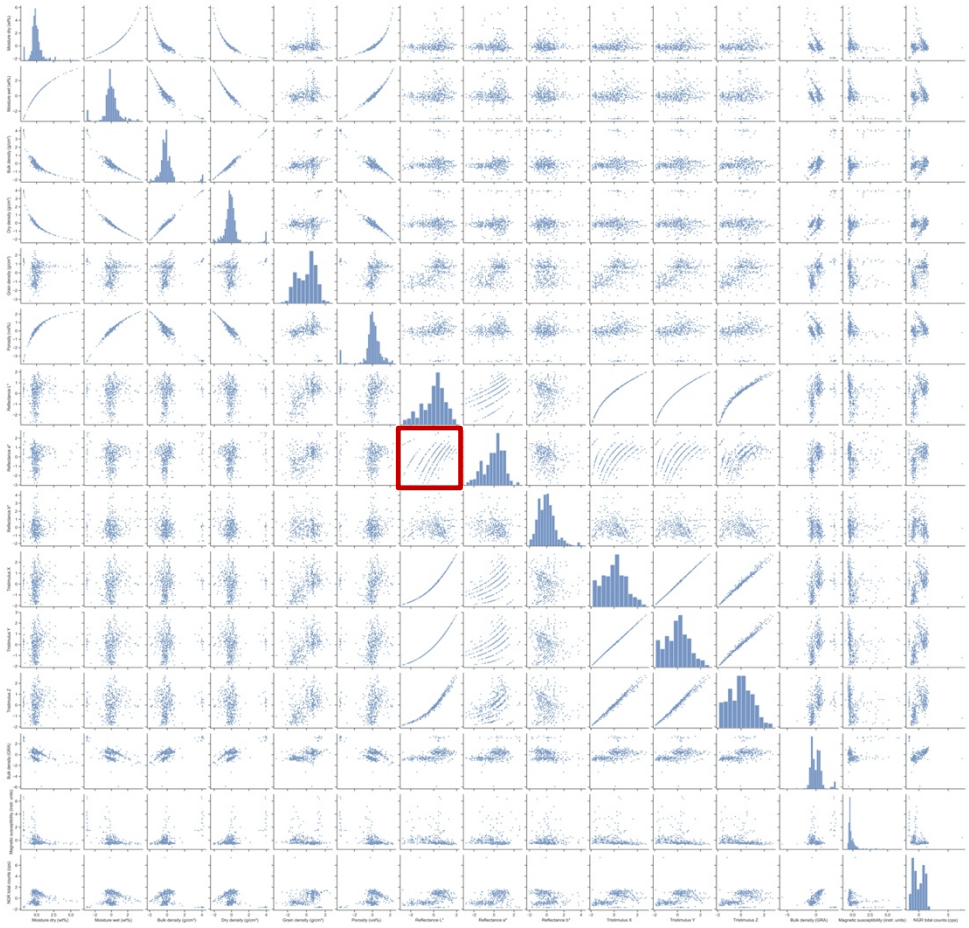
Unsupervised machine learning method K-means is applied to the well-log data in this work. K-means is a common unsupervised learning method which partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid) [3]. The result of the K-means algorithm is the predicted clusters of the input data. In this work, the outcome clusters of K-means are validated by lithologic units and geologic time scales that are given by conventional methods. The hyper parameter of K-means model includes n_clusters, namely k. The selection of k will also be explored in this work.

Besides, PCA (Principal Components Analysis) is used in this work to extract important features.

The python package sklearn is used to train the K-means model and conduct PCA. Besides, matplotlib package is used to visualize the results.

Results

Data investigation



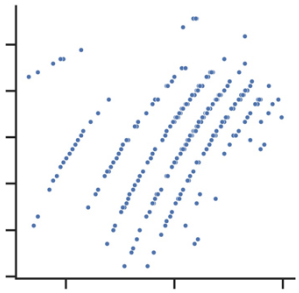


Figure 1: Pair plot of the features (the red box is enlarged)

Figure 1 gives the pair plot of the features. The contents in the red box are enlarged. We can see that in the red box, there are apparent clusters. There seems to be several lines in the red box, and different lines may be different clusters. Similar situation can be also found in other images. Therefore, from the pair plot, we can find apparent clusters.

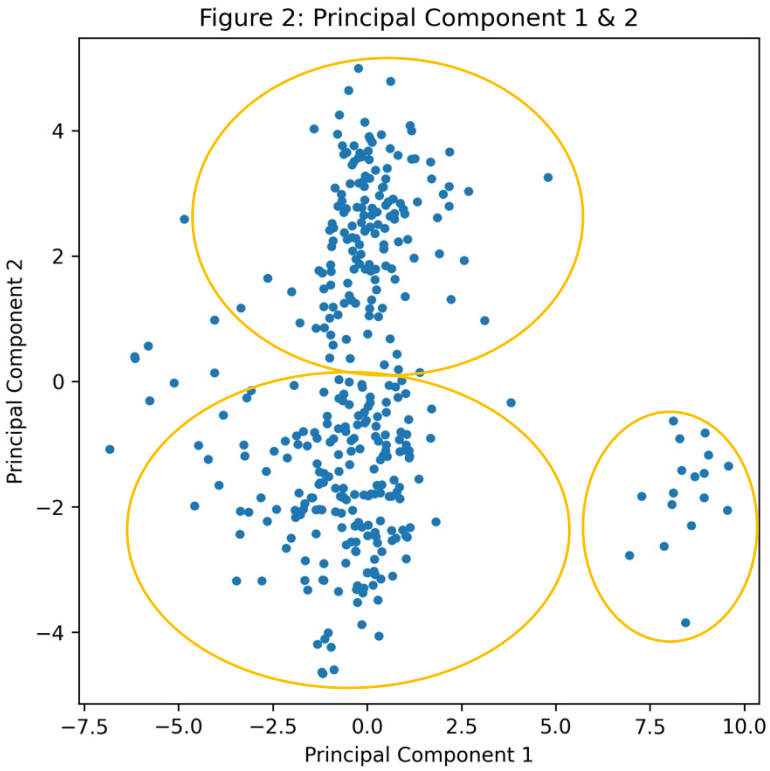


Figure 2 shows the principal component 1 and principal component 2 obtained by PCA. There seems to be at least three clusters in Figure 2, one above, one below, one on right (marked by the yellow circle). Of course, these three clusters recognized by eyes may also be subdivided.

Figure 1 and Figure 2 shows that the data has apparent clusters that can be recognized by human eyes.

K-means clustering results

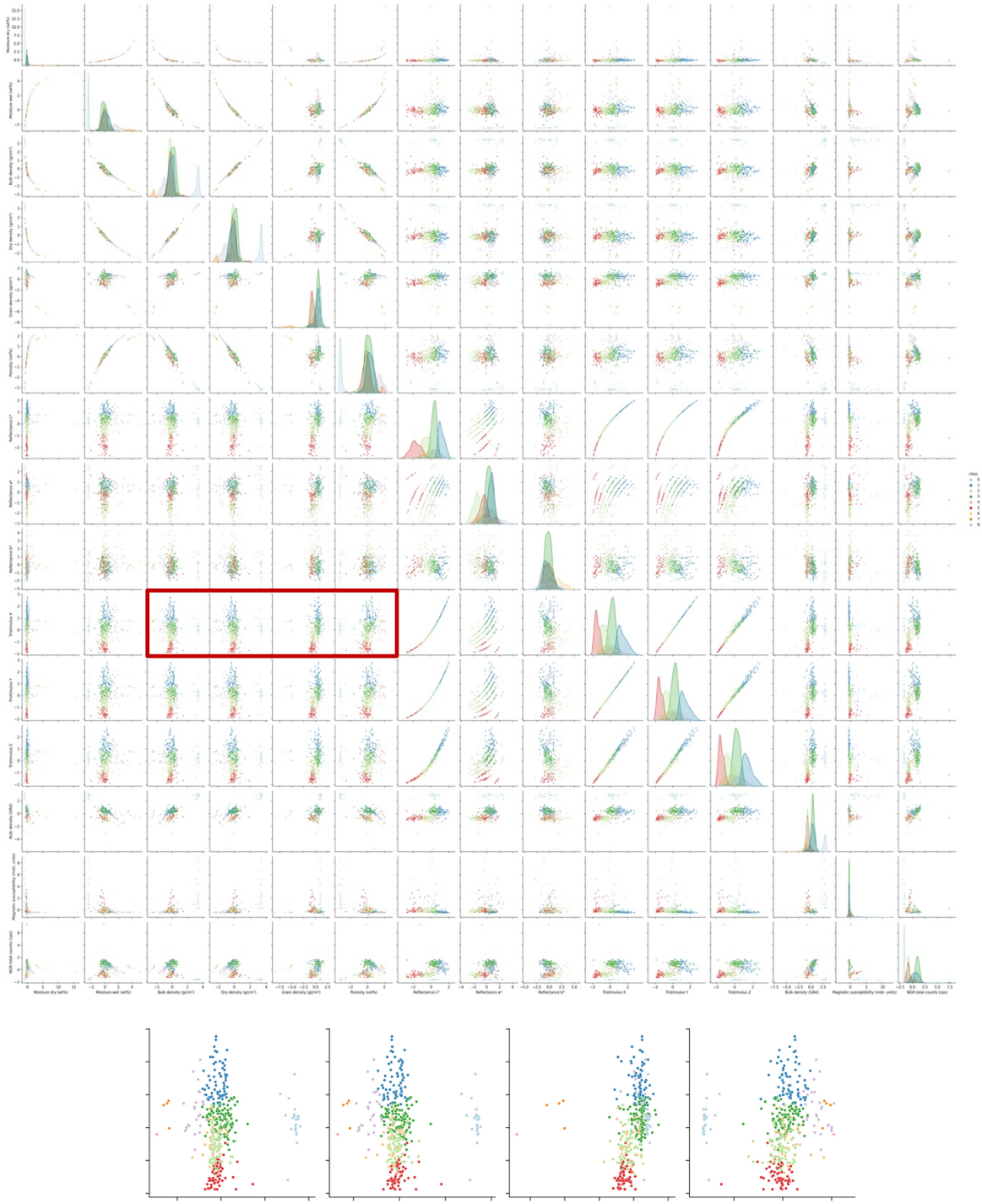


Figure 3: Pair plot of the features with clusters (k=9)

(the red box is enlarged below)

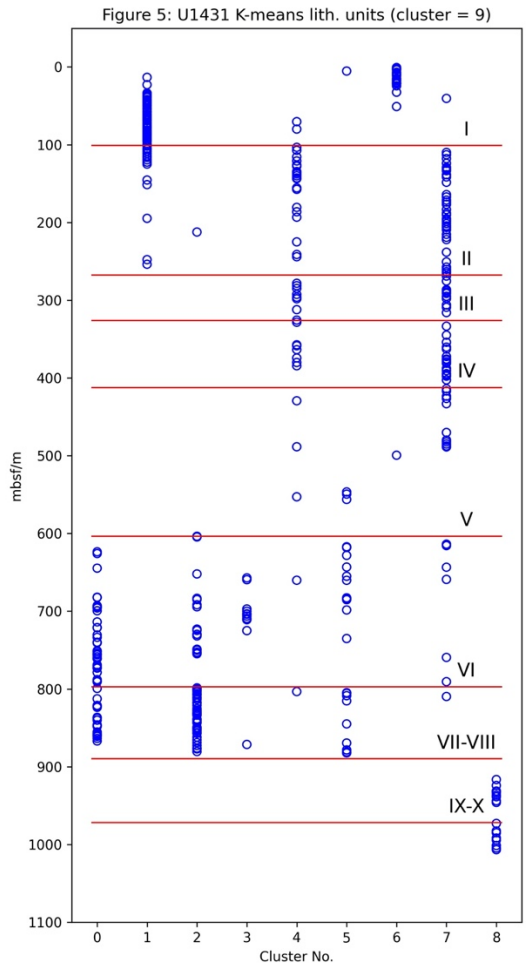
Figure 3 and Figure 4 shows the pair plot of the features with clustering results indicated. For Figure 3, the hyper parameter k of the K-means is 9, which is equal to the number of lithological units as indicated in Fig. 3 in Tse et al.[2]. For Figure 4, the hyper parameter k of the

K-means is 5, which is equal to the number of geological times as indicated in Fig. 3 in Tse et al.[2]. The contents in the red box are enlarged blow the pair plot.

From Figure 3 and Figure 4, we can see that the clustering results of K-means are great. From the contents in the red box, we can see that the data points represented by each color are grouped together to form clusters, and the classification effect is significant. Whether k==9 or k==5, the clustering results are excellent. Such result can also be observed on other parts of the pair plot. Therefore, we can confirm that the K-means algorithm works very well.

Figure 4: Pair plot of the features with clusters (k=5)

(the red box is enlarged below)



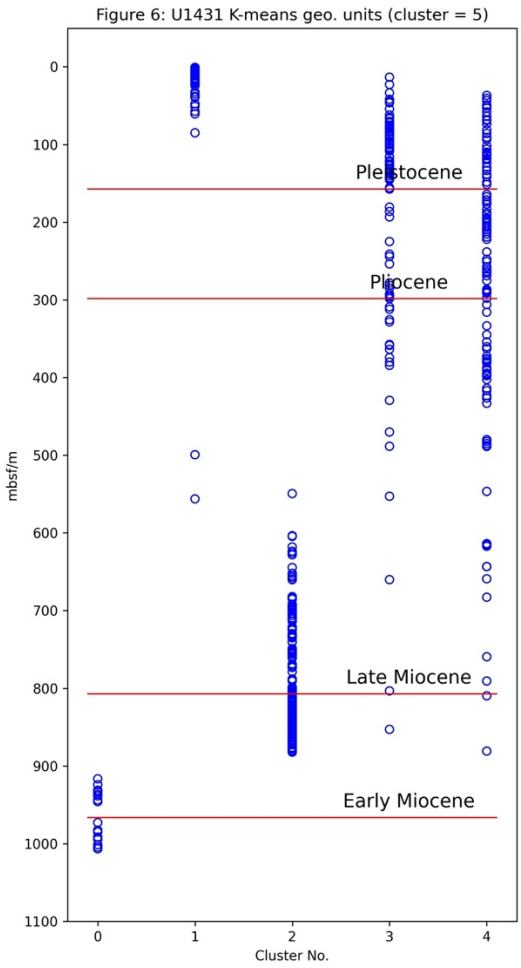


Figure 5 and Figure 6 shows the plots of cluster vs. time (just like Fig. 3 in Tse et al., 2019 [2]). Each data points in Figure 5 & 6 represents an observation and the red lines are the boundaries of lithologic or geologic time scale units that are already assigned to IODP ocean floor sediment cores by conventional methods.

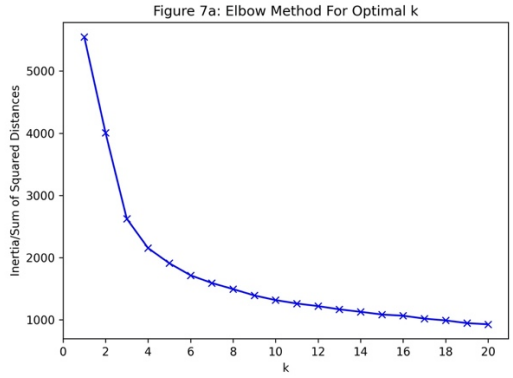
By comparing Figure 5 & 6 to the corresponding images given by [2], we can see that Figure 5 & 6 are similar to the results in [2].

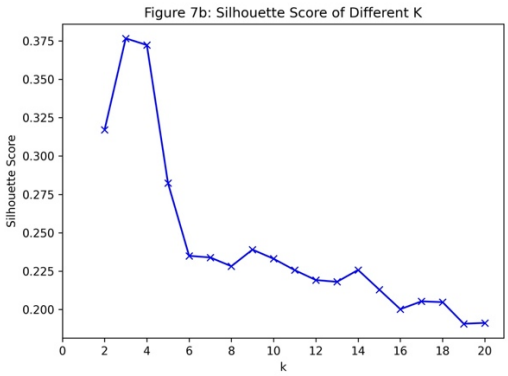
From Figure 5, we can see that the range of lithologic Unit VI to Unit VIII is mainly stretched by cluster 0 & 2 & 3 & 5. The range of lithologic Unit I is mainly stretched by cluster 1 & 6. The range of lithologic Unit II, III, IV, V is mainly stretched by cluster 4 & 7. The range of lithologic Unit IX, X, XI is stretched by cluster 8. Such result is similar to the results in [2]

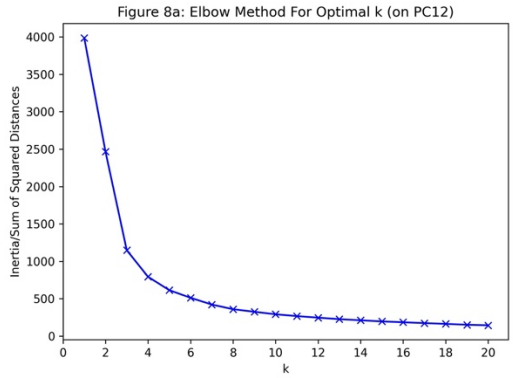
From Figure 6, we can see that cluster 3 terminates approximately around the Pliocene/late Miocene boundary. Such result is also similar to the results in [2].

From Figure 5 & 6, we can see that the clusters of K-means(k=9) correspond to lithological units well and clusters of K-means (k=5) correspond to geological time scale greatly. Due to the high correspondence to the lithological units and geological time scale, we can confirm that the K-means algorithm reaches excellent performance in lithology prediction.

Selecting optimal K for K-means







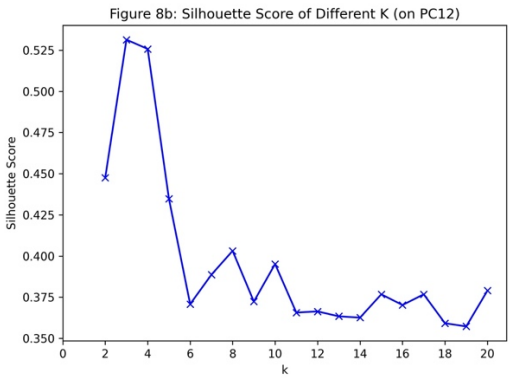
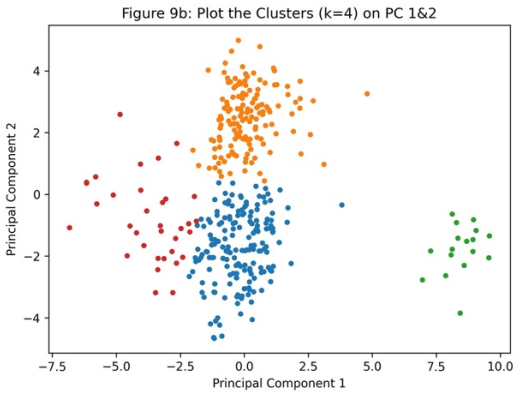
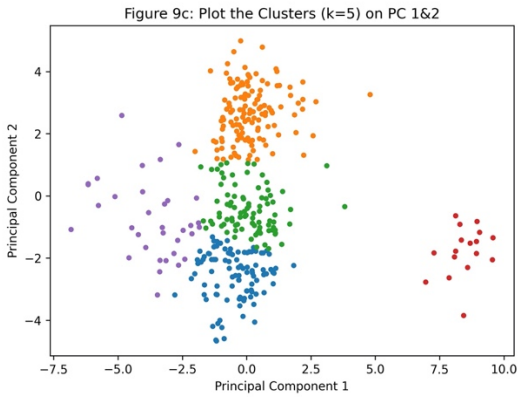


Figure 7a shows the inertia (sum of squared distance / loss function) of the K-means model with different value of k. It seems that from k=2 to k=3 and from k=3 to k=4, the value of inertia drops quickly, and after that the curve is not as steep as before. Therefore, k=3 or k=4 may be the best value of k according to Elbow Method (k=3 or k=4 is the elbow).

To confirm this, silhouette score is calculated to check the best k. The silhouette score is the evaluation of the density and dispersion of the cluster. The closer the silhouette score is to 1, the tighter is the interior of each cluster, and the further away is between different clusters, which means the clustering result is better. Figure 7b shows the silhouette score of K-means models with different value of k. Apparently, when k=3 or k=4, the silhouette score is the largest. Therefore, the best value of k is 3 or 4 (between 3 & 4, 3 is slightly better, but almost the same).

Figure 8a & 8b does the same thing on the K-means model on Principal components 1 & 2 instead of the complete raw data. The result is the same: the best value of k is 3 or 4 (between 3 & 4, 3 is slightly better, but almost the same).





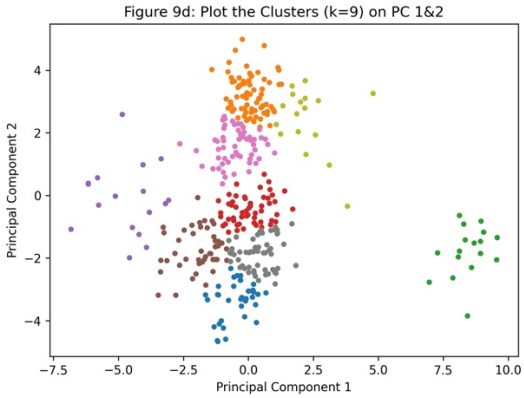


Figure 9a & 9b & 9c & 9d shows the PC1 and PC2 with clusters indicated (k = 3, 4, 5, 9). The K-means model is based on only PC1 and PC2 instead of the complete raw data. We can see that for k=3 and k=4, the clustering result is perfect. The data points of different colors form different clusters and the clusters are compact. This proves that k=3 or 4 are the best values of k. Besides, results of k=5 and k=9 are also shown. The results of k=5 is almost satisfactory, but the result of k=9 is not so well.

Discussion and Conclusions

In this work, unsupervised machine learning method, K-means, is applied to well-log data to predict lithology. Different from supervised learning methods, the unsupervised model in this work does not require any priori information and only needs input features. It turns out the clustering result of K-means model highly corresponds to the lithological units and geological time that are assigned to the dataset by conventional methods. Therefore, the K-means model in this work reaches excellent performance. Such powerful model is of significant practical meaning as it gives reliable lithology prediction, which helps humans to understand underground structure better and may benefit some important industry, such as oil extraction.

Besides, the elbow method is used to explore the best k for K-means. Such hyper parameter exploring method could generate model with stronger power of expressiveness.

In conclusion, the unsupervised learning model in this work shows excellent performance in predicting lithology based on well-log data, which is of great practical meaning.

References

[1] https://web.iodp.tamu.edu/LORE/

[2] 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." Frontiers of Earth Science 13 (1): 180–90.

[3] https://en.wikipedia.org/wiki/K-means_clustering