ACSE9: IRP Project Plan
Machine learning for automatic facies classification from 3D geophysical models
Author: Nitchakul Pipitvej
Supervisor: Dr. Michele Paulatto

Rationale and Project Objectives

In geology, identifying the type of rocks is an important process. Often, a site region will harbor multiple rock types but not all of them are the object of interest. Still, the conventional way of determining the lithology by laboratory studies of rock cylinder retrieved from drilling boreholes is very costly and time-consuming. Meanwhile, drilling boreholes can also collect well logs data along the hole, this is called downhole logging and is much faster and cheaper than laboratory measurements. Companies have been trying to devise methods based on machine learning (ML) to determine lithology from well logs, which are relatively high-resolution (1m - 10 m resolution) but are in 1D, thus only provide a single vertical profile.

The objective of the project is to try to determine lithology from physical properties measured in 2D or 3D with geophysical methods. These data are typically much lower resolution (500 m to 4 km resolution). This leads us to take ML into consideration. With the recent development of ML, there are many newly developed methods and techniques that were yet to be explored with 2D and 3D data. It is the aim of this project to see how well each method perform across combinations of available rock properties. It should also be possible to assess whether it will return a satisfying prediction result when smooth, noisy, or incomplete data is available.

Short Literature Reviews

Dating back to 2007, a research was conducted to find the correlation between some physical properties of the rock and its lithology (Bedrosian *et al.*, 2007). While each property does not show any correlation, the combination of them can be used to derive a meaningful relationship with the lithotypes. By computing the joint probability density function (pdf) of the interpolated magnetotelluric (MT) data and seismic data, it is possible to classify the resulting joint pdf into broad lithotypes. In this research, the important properties derived from MT and seismic data are the electrical resistivity and the compressional wave velocity (V_p).

Many of the later researches use this as a base knowledge to derive the lithotypes from resistivity and V_p. One journal in 2012 studied the relation of these geophysical models by using self-organizing maps (SOMs) instead of pdf (Bauer, Muñoz & Moeck, 2012). It was mentioned that by using SOMs, multiple geophysics properties can be analyzed at once. While the journal uses both resistivity and V_p, one additional property added was the V_p gradient. By introducing this additional property into the analysis, it gains an increase in the potential to classify the explored data into more classes compared to the pdf method introduced from its referenced study (Muñoz *et al.*, 2010). Similarly, (Braeuer & Bauer, 2015) also explore the use of SOMs with different geophysical models, V_p, shear wave velocity (V_s), and V_p/V_s ratio and achieve a broad classification of the sediments. It was also stated that SOMs can be used in place of k-means clustering (Bação, Lobo & Painho, 2005). A similar conclusion was also discovered upon testing the two methods with seismic data, stating that SOMs will outperforms k-mean in cases that the input is continuous (Coléou, Poupon & Azbel, 2003). Another classification method that was explored with resistivity and V_p is the Fuzzy c-means clustering (García-Yeguas *et al.*, 2017). Using the mentioned method, it achieved to attain a more precise location of the internal structure of the studied island.

Most of the mentioned researches uses only small number of geophysical properties. There is a possibility that by using the same method on more parameters can increase the classification precision. In the research that compares SOMs with Feed Forward Neural Network (FFNN), using seven rock properties (Konaté *et al.*, 2015). While the purpose of this study is to compare the performance of SOMs and FFNN, it shows that using SOMs on multiple properties can derived meaningful classification of the lithology.

While there are many researches focused on the MT and seismic data, not much of the other parameters of geophysics are explored. It is possible that there are other notable relations in the unexplored parameters that can derived a more accurate model for lithology classification. Likewise, many classification methods were left unexplored in this field as well.

Proposed approach

As mentioned in the literature review, there are many combinations of geophysical properties and ML methods that were yet to be explored. One of the properties was found be useful is the seismic wave velocity; namely V_p, V_s, and V_p/V_s ratio. Additional properties that could be explored are the density, and electrical resistivity.

To test each ML methods, it was necessary that a cross-sectional model of the Earth must be generated. The initial synthetic model shown in Figure 1 contains multiple lithological classes, which are Siltstone, Clastic, Lavas, Intrusions, Meta sediments, Upper crust, Lower crust, Diorite, and Gabbro. This is a representation of a depth cross-sectional earth model that contains multiple rock types. This model was generated by Dr Michele Paulatto, who is the supervisor of this project. This model represents the real-world geophysics model which contain sedimentary layers, a volcanic edifice, faults and magmatic intrusions. This model was generated using multiple geological parameters; which are V_p, V_s, porosity, temperature, melt content, and pore aspect ratio.

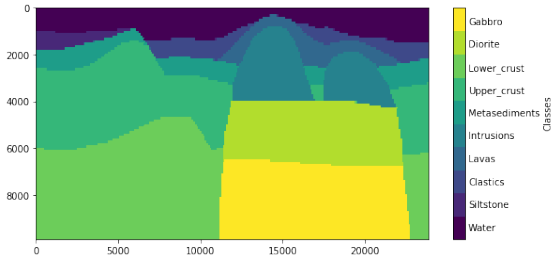

Figure 1Classes in the synthetic model

Using the mentioned geological parameter, the main input of this project is then generated as seen in Figure 2. The parameters available to test the ML method are the V_p, V_s, density, V_p/V_s ratio, compressional wave attenuation (Q_p), and shear wave attenuation (Q_s).

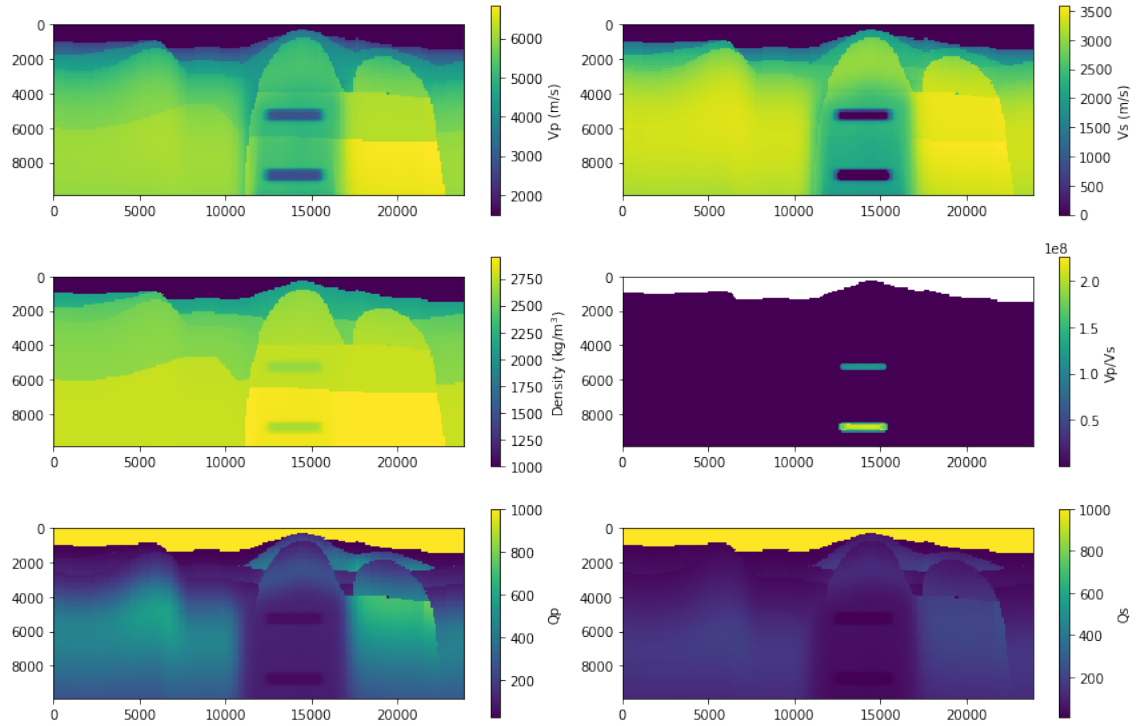

Figure 2 Properties in Synthetic model, Compressional wave velocity (1), Shear wave velocity (2), Density (3), Vp Vs ratio (4), Attenuation (5,6)

Boreholes data could also be captured from this model, Figure 3 is an example of plotted 1D data.

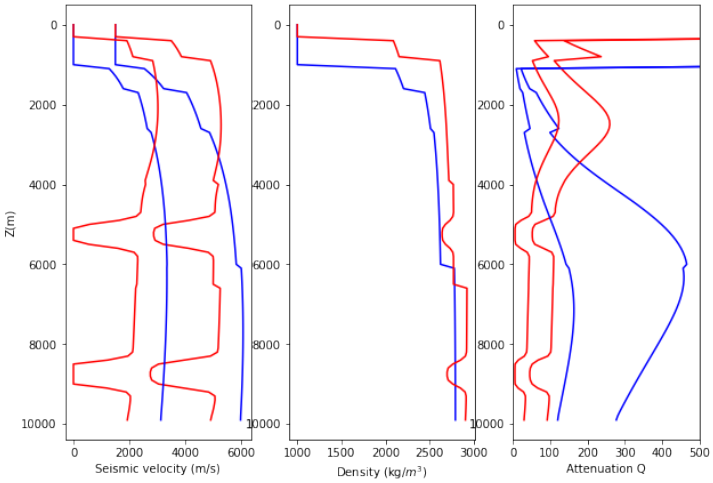

Figure 3. 1D vertical profile data from x = 15000

This synthetic model will be treated as part of the project input to train and test the accuracy of ML approaches.

As for the ML methods to test, some of the tested methods seen in the literature review are as follow: SOMs, FFNN, Autoencoder, k-means Clustering, and Fuzzy c-mean clustering. While those methods were tested on other works, it is still possible to use those methods with this synthetic model and see which combination of inputs would give out the best result. Some other methods that can be further explored are support vector machines (SVM), k-nearest neighbors, Naïve Bayes, and Decision Tree. It is also possible to apply ensemble methods once the models are trained.

Due to the timespan of the project, it is not possible to choose all the mentioned methods to test with each combination of input data. This thus led to the following list of possible methods to explore.

- SOMs: an unsupervised learning artificial neural network introduced by (Kohonen, 1990). Its notable feature is its ability to take in an input of unlimited dimension and represent them as a map of lower dimension, which corresponds to the project's input of having multiple dimensions in the input. This method can be considered as a dimensional reduction approach. This output can then be processed and analyze to either manually or automatically assign the lithological class of each observed data cluster via some clustering method, then maps back to the original topography.

- K-means clustering: an unsupervised learning method. The general purpose of this method is to divide a multivariant input into a specific number of classes (Macqueen, 1967). The algorithm starts by initializing several clusters for the input to be assigned to, then assign the centroid position. Each data point will then be assigned to the class of the nearest centroid. The centroid position is then updated using the mean of all the data points assigned to that class. The final clustering is reached when the position of the centroid does not update anymore.
  One drawback of this method was that we need to know the initial number of clusters it correctly classifies the inputs. A popular approach to find this number is the elbow method (Ketchen & Shook, 1996), which iterates the number of cluster and plots it to find the point where the gradient of the mean distant to the centroid changes the most. That point demonstrates the number of optimal clusters in the observed dataset.

- Naïve Bayes: a probabilistic classifier, classified as a supervised learning method. It utilizes the Bayes theorem and predicts the class according to the given input parameters. A limitation was that each parameter should be independent of each other, which may not be well represented in the observed synthetic geological model.

- Feed Forward Neural Network: a supervised learning method. The network may contain multiple layer with several neurons. Each of these neurons will interpret the input by applying some activation function and updates the weight of the neuron at each iteration. By applying a Softmax function at the last layer, the probability of the input belonging to each class of lithology can be found (Montavon, Samek & Müller, 2018). It would be possible to modify a neural network architecture from other work with similar input and apply to this project's input. This can lessen the time needed to come up with a working architecture. However, this method posed a problem that it will require us to know the number of classes beforehand. An approach that can deal with this is to add a limitation to the software to only classify a data point into a class when the probability is over some threshold. OpenMax model layer (Bendale & Boult, 2015) could also be implemented to deal with the problem of unknown number of classes.

- K-nearest neighbors: a supervised learning method. By storing each input parameters as a vector, the distance between each trained data point to the new one can be found. As the name suggest, it will compare the *k* nearest neighbor class and assign the new data point's class based on those neighbors' class. This method may be used with the output SOMs or other dimensional reduction methods.

These methods are considered as they are some commonly used technique in ML and should be able to be implemented in a reasonable time. As for SOMs, the technique is often used in classifying lithology thus should be tested as a benchmark compared to the other techniques. Depending on the work progress and timeframe, additional methods that retain spatial relation may be tested.
It should also be noted that dimensional reduction and some data augmentation may be incorporate as part of the data preprocessing. Ensemble method could be added once all the other models are completed.

Testing Approaches
For the cases of supervised method, labeled data are needed to train the ML model. Using the script that generated the 2d synthetic model, it should be possible generate a variety of synthetic model to be used as the inputted data. Those models could be divided into several subsets of training set, test set, and validation set. These subsets should be divided using stratified random sampling, in which each subset will contain all the classes presented in the input, to train it in accordance to the representation of the whole data population. While it is not possible to know how many types of rocks are presented in an observed area in real life, it could be inferred that the trained ML model has been trained on all possible rock facies and will be able to predict unforeseen earth model.
Each subset of the inputted data should be used to train each supervised model. K-fold cross validation method is one of the popular methods that can generalize the error quite well while not being too complex.

As for the unsupervised methods, there are many clustering validation approach that could be taken. Further reading on this subject can be found here (Halkidi, Batistakis & Vazirgiannis, 2001).

Upon finding the most promising classification model, it should be tested with an unseen synthetic model that will be generated in a similar notion as the initial model. Additional test may be pursue using the trained model. Additional synthetic model with smoothening or noises added can be tested to see how well the trained model performed. This could represent the actual accuracy of the model since data recorded from well-logs may contain noises.

Software Specification and Implementation
With the objective of testing each ML methods in mind, python was chosen as the implementation language. The broad range of methods in the available libraries such as scikit-learn and TensorFlow could help jump start the implementation process. As of current, the code is developed on Jupyter Notebook environment and runs locally without any GPU. Upon further development, this could be ported to services like Google Colab which provide free access to GPUs and can help speed up the training and testing process.

The structure of the code will be separated into functions of each implemented methods and stored as a library. This library should be imported and used while Jupyter Notebook will act as an interface of the project. The input of each classification method should be either the whole or subset of the training dataset, and outputs the trained model of the method. This trained model should then be used to classify the testing dataset, which will then be feed to the validation function to evaluate the accuracy of the model.

List of Task and Milestones

The ideal timeline of the project is shown in Figure 4. Notable dates are as follows:

August 28, 2019	Project Plan completion
July 9, 2019	Prototype of first ML method
July 23, 2019	Prototype of second ML method
August 6, 2019	Prototype of third ML method
August 7, 2019	Software Prototype
August 17, 2019	Software clean up and performance check
August 30, 2019	Final report and project submission

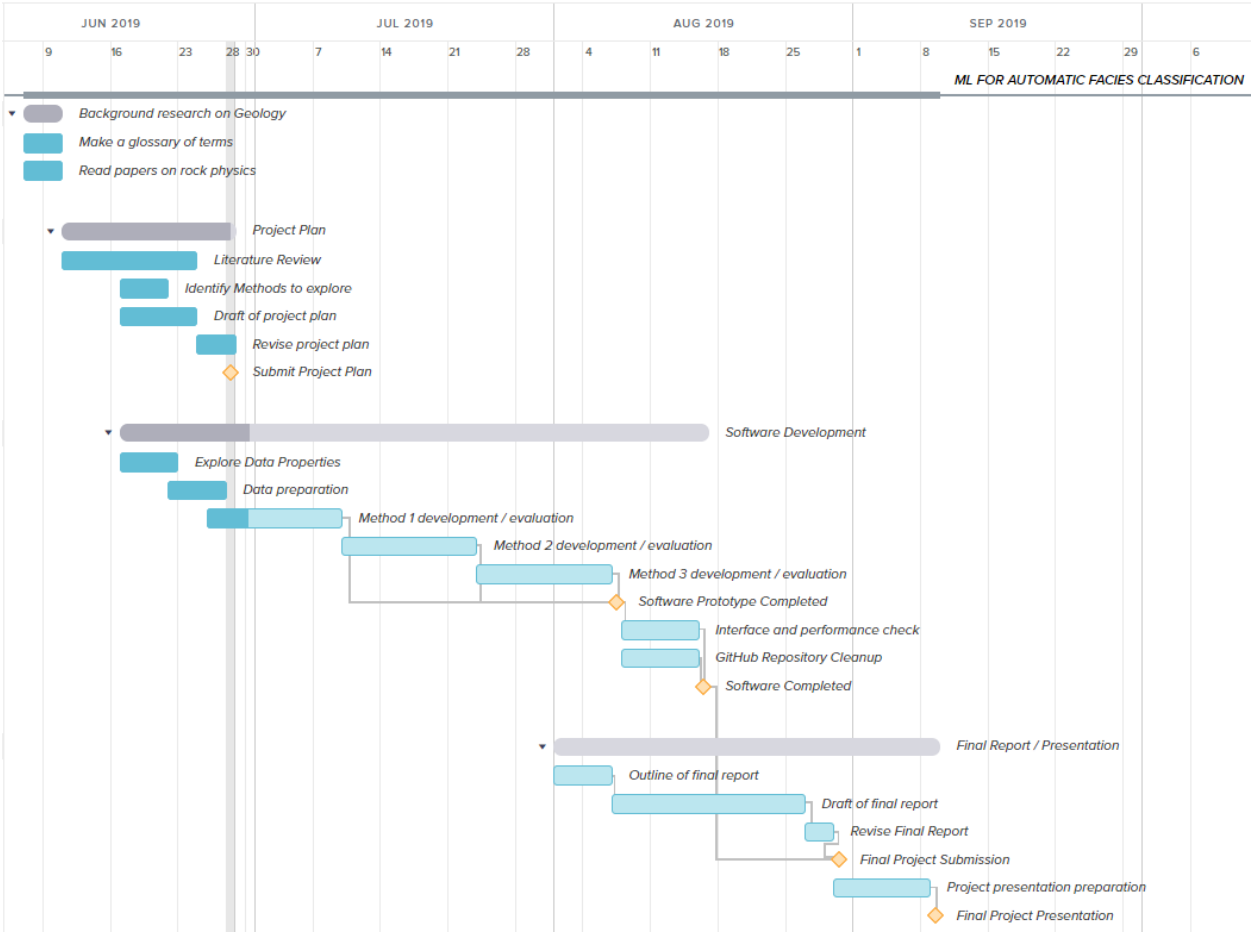

Figure 4 Gantt Chart of the project timeline

References

Bação, F., Lobo, V. & Painho, M. (2005) Self-organizing Maps as Substitutes for K-Means Clustering. In: Vaidy S. Sunderam, Geert Dick van Albada, Peter M. A. Sloot, & Jack Dongarra (eds.). *Computational Science – ICCS 2005*. Lecture Notes in Computer Science. 2005 Springer Berlin Heidelberg. pp. 476–483.

Bauer, K., Muñoz, G. & Moeck, I. (2012) Pattern recognition and lithological interpretation of collocated seismic and magnetotelluric models using self-organizing maps. *Geophysical Journal International*. [Online] 189 (2), 984–998. Available from: doi:10.1111/j.1365-246X.2012.05402.x.

Bedrosian, P.A., Maercklin, N., Weckmann, U., Bartov, Y., et al. (2007) Lithology-derived structure classification from the joint interpretation of magnetotelluric and seismic models. *Geophysical Journal International*. [Online] 170 (2), 737–748. Available from: doi:10.1111/j.1365-246X.2007.03440.x.

Bendale, A. & Boult, T. (2015) Towards Open Set Deep Networks. *arXiv:1511.06233 [cs]*. [Online] Available from: http://arxiv.org/abs/1511.06233 [Accessed: 26 June 2019].

Braeuer, B. & Bauer, K. (2015) A new interpretation of seismic tomography in the southern Dead Sea basin using neural network clustering techniques. *Geophysical Research Letters*. [Online] 42 (22), 9772–9780. Available from: doi:10.1002/2015GL066559.

Coléou, T., Poupon, M. & Azbel, K. (2003) Unsupervised seismic facies classification: A review and comparison of techniques and implementation. *The Leading Edge*. [Online] 22 (10), 942–953. Available from: doi:10.1190/1.1623635.

García-Yeguas, A., Ledo, J., Piña-Varas, P., Prudencio, J., et al. (2017) A 3D joint interpretation of magnetotelluric and seismic tomographic models: The case of the volcanic island of Tenerife. *Computers & Geosciences*. [Online] 109, 95–105. Available from: doi:10.1016/j.cageo.2017.08.003.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2001) On Clustering Validation Techniques. *Journal of Intelligent Information Systems*. [Online] 17 (2), 107–145. Available from: doi:10.1023/A:1012801612483.

Ketchen, D.J. & Shook, C.L. (1996) The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*. [Online] 17 (6), 441–458. Available from: doi:10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.

Kohonen, T. (1990) The self-organizing map. *Proceedings of the IEEE*. [Online] 78 (9), 1464–1480. Available from: doi:10.1109/5.58325.

Konaté, A.A., Pan, H., Fang, S., Asim, S., et al. (2015) Capability of self-organizing map neural network in geophysical log data classification: Case study from the CCSD-MH. *Journal of Applied Geophysics*. [Online] 118, 37–46. Available from: doi:10.1016/j.jappgeo.2015.04.004.

Macqueen, J. (1967) Some methods for classification and analysis of multivariate observations. In: *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*. 1967 pp. 281–297.

Montavon, G., Samek, W. & Müller, K.-R. (2018) Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*. [Online] 73, 1–15. Available from: doi:10.1016/j.dsp.2017.10.011.

Muñoz, G., Bauer, K., Moeck, I., Schulze, A., et al. (2010) Exploring the Groß Schönebeck (Germany) geothermal site using a statistical joint interpretation of magnetotelluric and seismic tomography models. *Geothermics*. [Online] 39 (1), 35–45. Available from: doi:10.1016/j.geothermics.2009.12.004.