Imperial College London

9 September 2019

Machine Learning for Automatic Facies Classification from Tomography Models

Individual Research Project

MSc in Applied Computational Science and Engineering

Nitchakul Pipitvej

Supervised by Dr. Michele Paulatto

Imperial College London

Lithological classification

- Identifying rock type from geophysical data
- Saves time and money compared to drill core analysis

Machine learning can help, applied to:

- Well logs
- 2D tomography sections

GOALS:

- implement code to assess different methods
- test with realistic synthetic data

Imperial College London

Classification methods: Unsupervised Learning

- Pros
  - does not need initial label
  - is not biased by the targeted output
- Cons
  - difficult to evaluate result
  - noise may affect the resulting cluster

Imperial College London

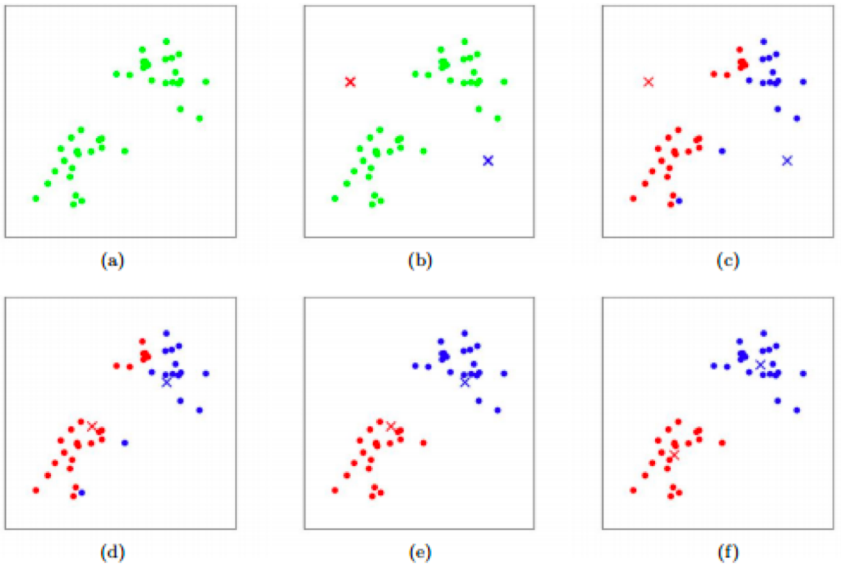
Methods tested

Method	Pros	Cons
K-Means	- Fast and Simple	- Limited to convex cluster shape
Fuzzy C-Means (FCM)	- Soft clustering	- Limited to convex cluster shape
Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBScan)	- Not limited to convex shape - Noise cluster	- May lost information in noise cluster
Self-Organizing Maps (SOMs)	- Dimensional reduction - Robust on large data input	- Many hyperparameters

Imperial College London

K-Means

1. Randomly choose K cluster centers.
2. Assign each data points to the nearest cluster center.
3. Update the cluster centers with the mean of all the data points in each cluster.
4. Repeat until cluster centers converge.



source: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Imperial College London

Fuzzy C Means (FCM)

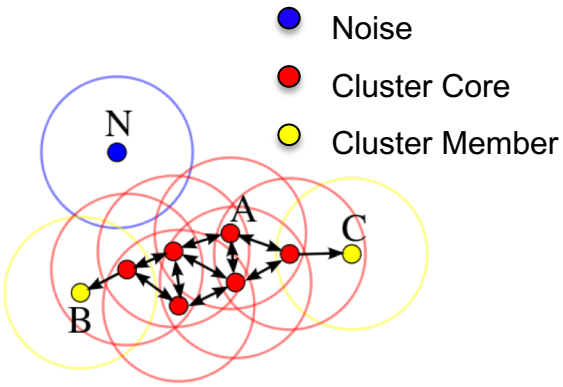
1. Randomly choose C cluster centers.
2. Randomly assign the membership degree of each data point.
3. Computes the centroid weighted by membership degrees
3. Update the membership degree by the distance of data points to each cluster center.
4. Repeat until the cluster centroid converges.

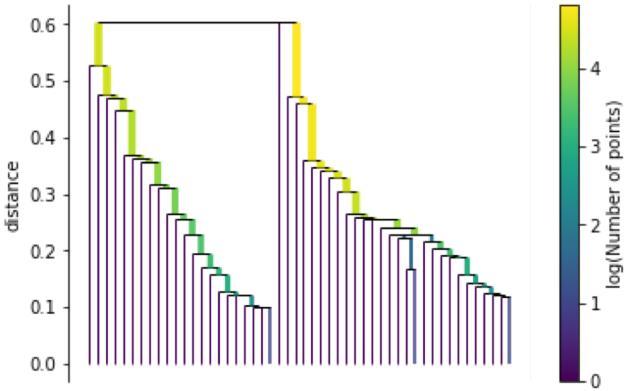
Imperial College London

HDBScan

Hyper-parameters: minimum number of points (min_sample), minimum cluster size (min_cluster_size)

1. Identify core points (having over min_sample neighbour within **ε** distance).
2. Computes single linkage dendrogram of the core points
3. Combines the branches based on the minimum cluster size

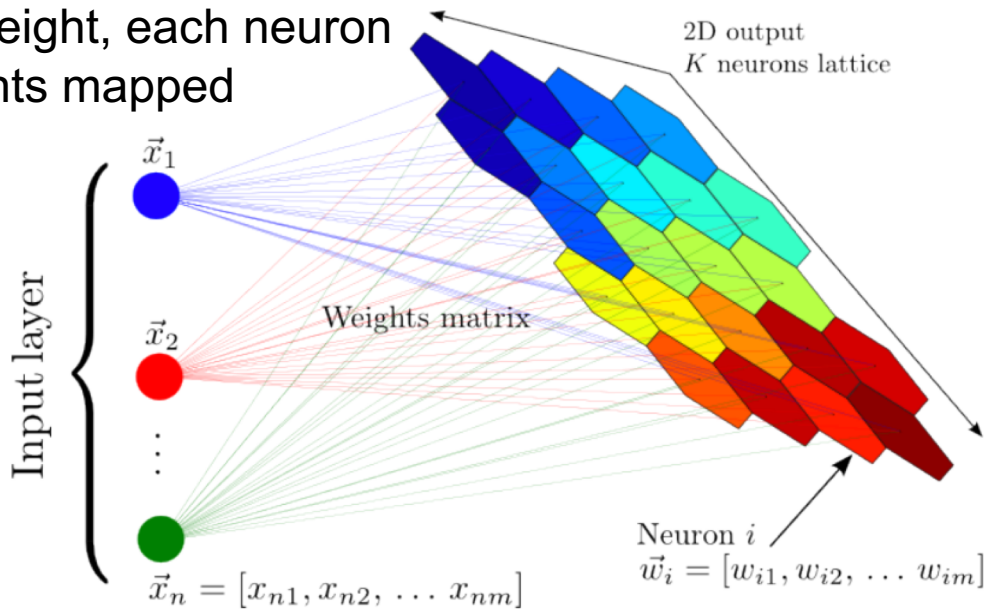




Imperial College London

SOMs

1. Initialize Neural network layer's weight (2D)
2. Compute each data point's Euclidian distance to each neuron, assign the point to the closest neuron.
3. Update the Neural network layer weight, each neuron will be more similar to the data points mapped to that neuron.

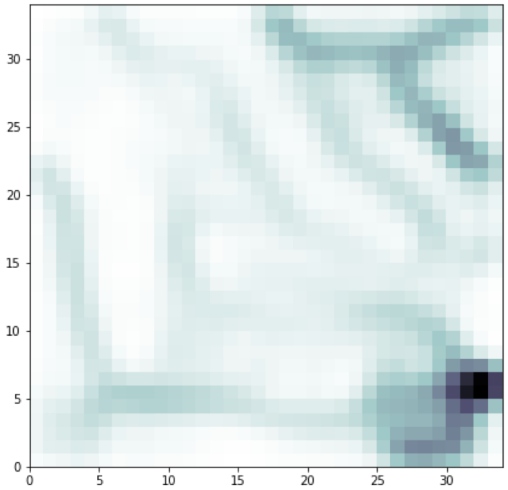


source: https://arxiv.org/pdf/1312.5753.pdf

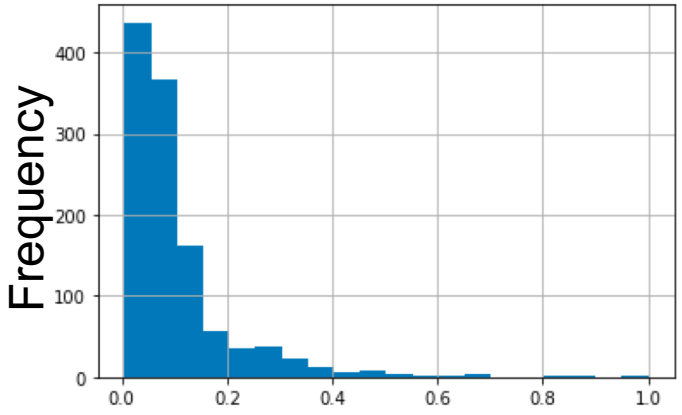
Imperial College London

SOMs (2)

4. Calculate u-matrix (distance between neurons)
5. Find clusters in neuron map:
   - Watershed segmentation (find ridges and valleys)
   - Other clustering methods: K-means, FCM, HDBSCAN



U-Matrix



Distance

Imperial College London

Verification methods

- Internal Indexes
  - Sum Squared Error (SSE)
  - Silhouette Score
  - Calinski Harabaz Score
- External Indexes
  - Entropy
  - Purity

Imperial College London

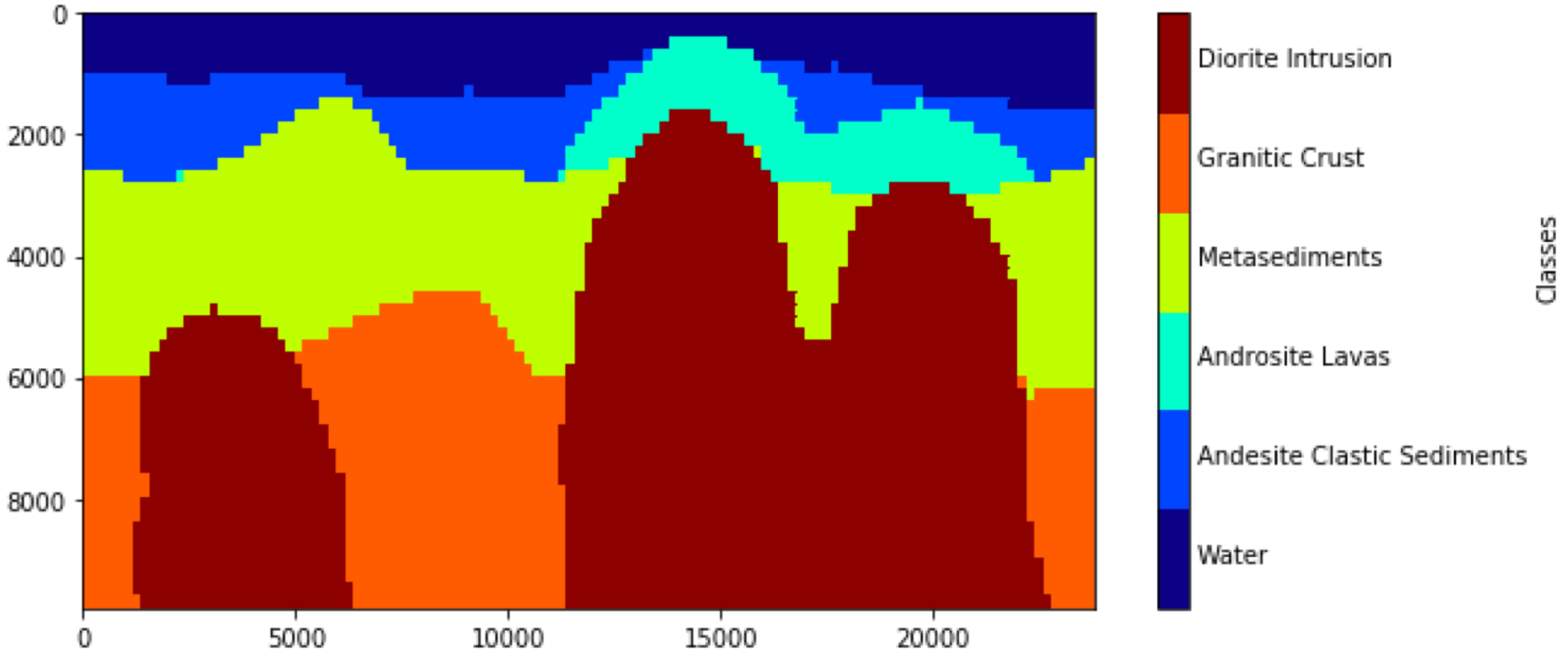
Program Flow

1. Data pre-processing
2. Classification Methods
   - K-Means
   - Fuzzy C means (FCM)
   - HDBSCAN
   - Self-Organizing Maps
     - SOMs – K-means
     - SOMs – FCM
     - SOMs – HDBSCAN
3. Verification and Validation

Imperial College London

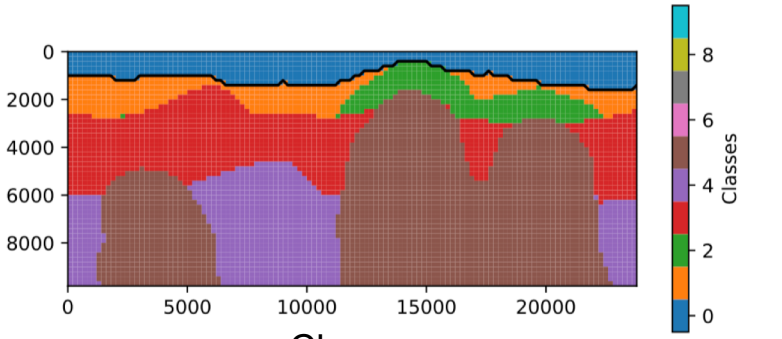
Test model: synthetic arc crust

5 different input lithological classes

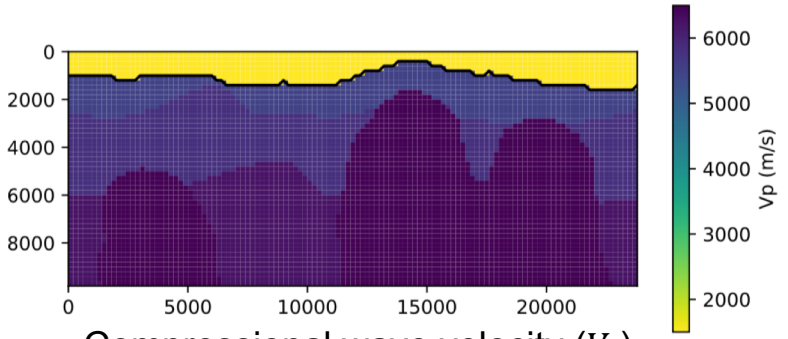


Imperial College London

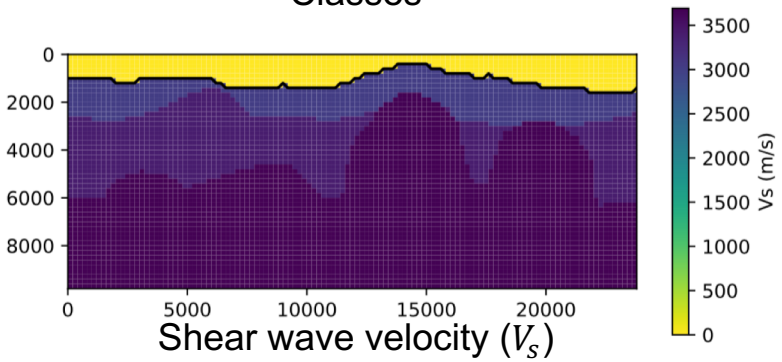
Underlying rock parameters



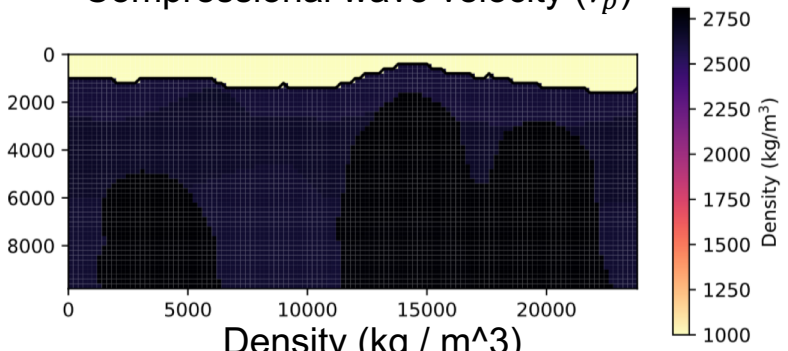
Classes



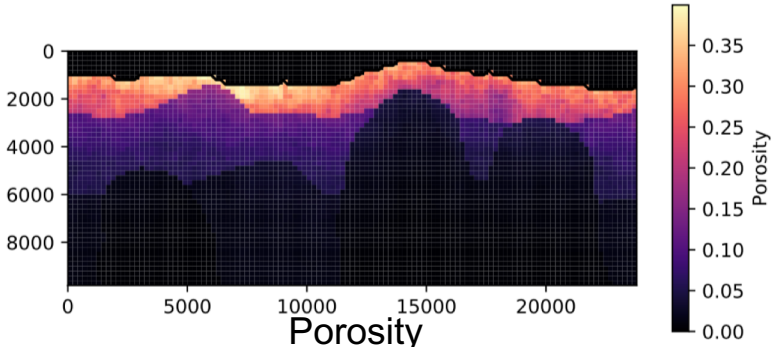
Compressional wave velocity (V_p)



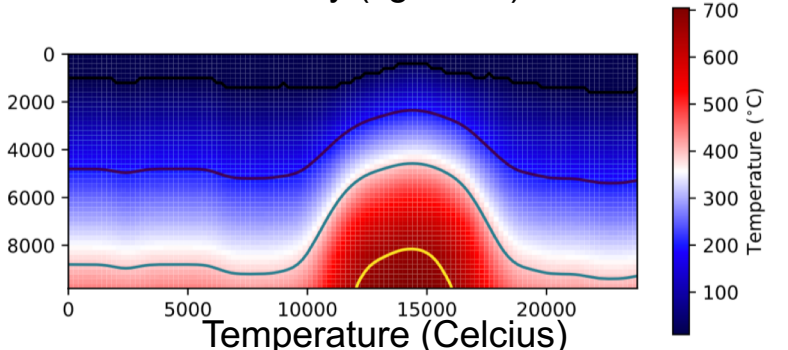
Shear wave velocity (V_s)



Density (kg / m^3)



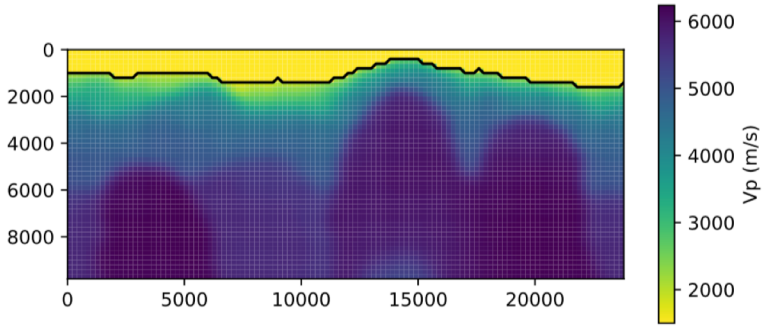
Porosity



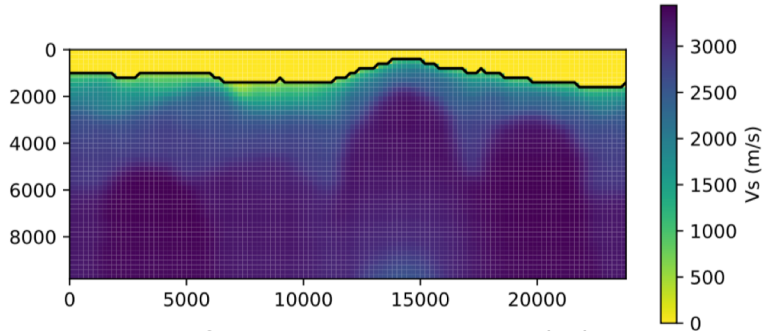
Temperature (Celcius)

Imperial College London

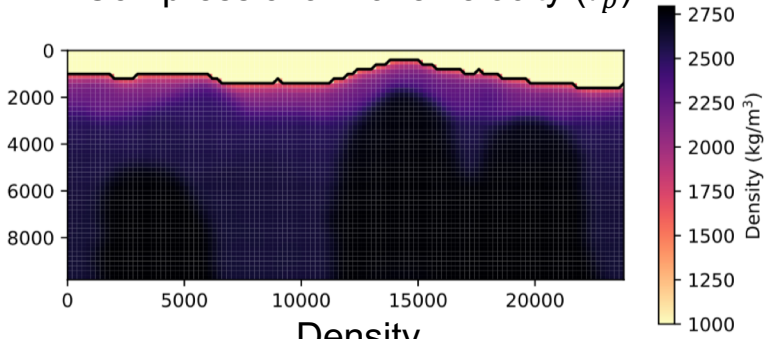
Input physical properties



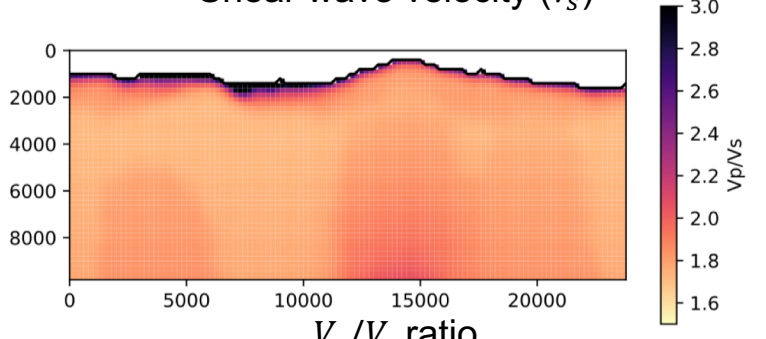
Compressional wave velocity (V_p)



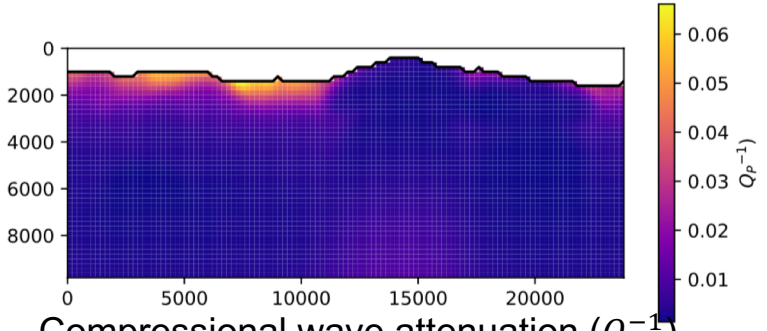
Shear wave velocity (V_s)



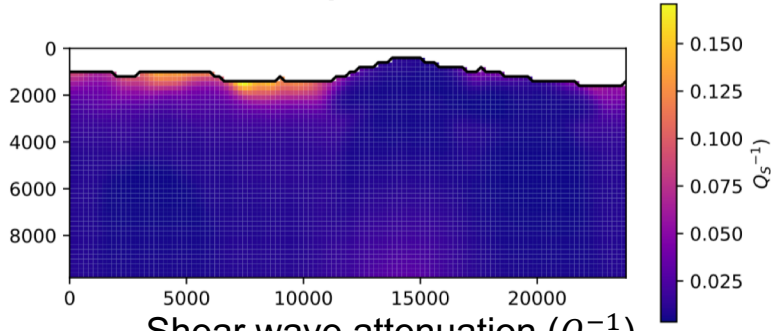
Density



V_p/V_s ratio



Compressional wave attenuation (Q_p^{-1})

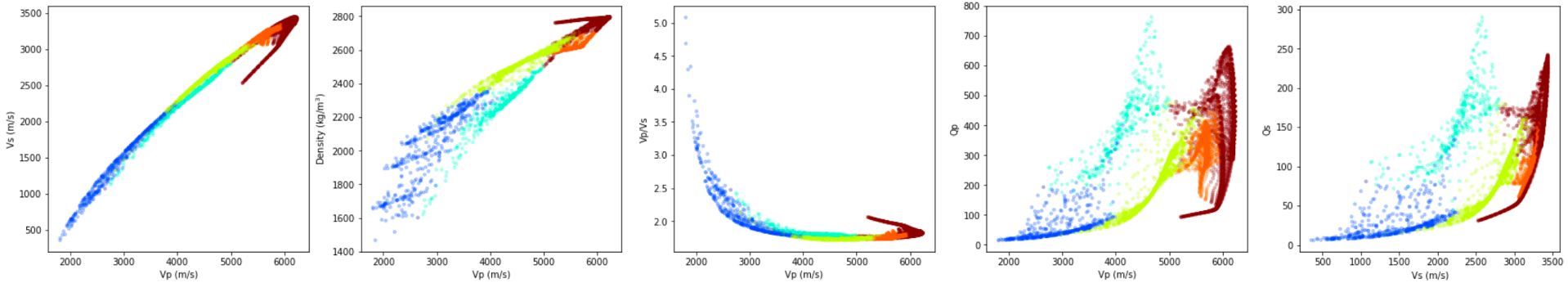


Shear wave attenuation (Q_s^{-1}).

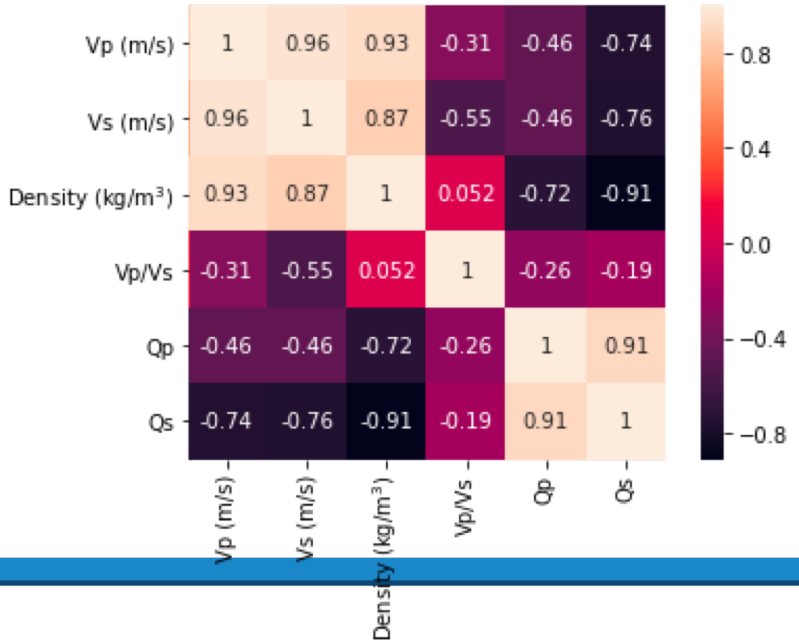
Imperial College London

Cross-plots

Each class has a different colour:



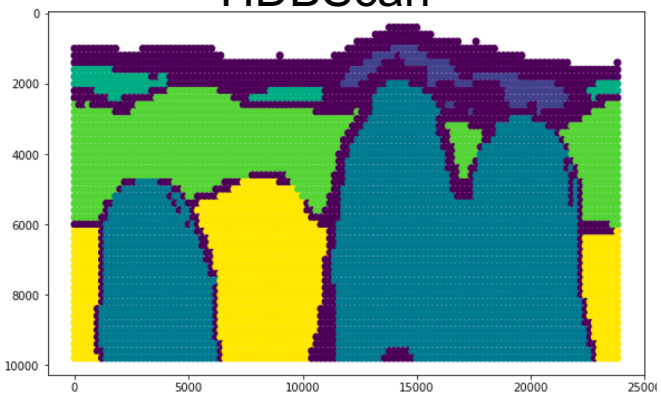
- V_p, V_s, density, V_p/V_s ratio, Q_p^{-1}, and Q_s^{-1}
- V_p, V_s, and V_p/V_s
- V_p, Q_p^{-1}, and density
- V_p and V_s



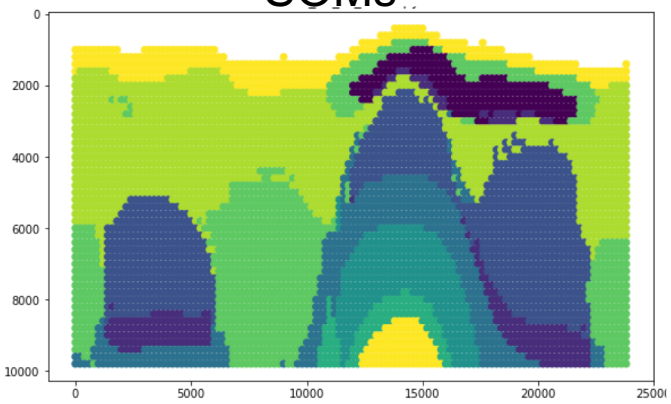
Imperial College London

Each method test result

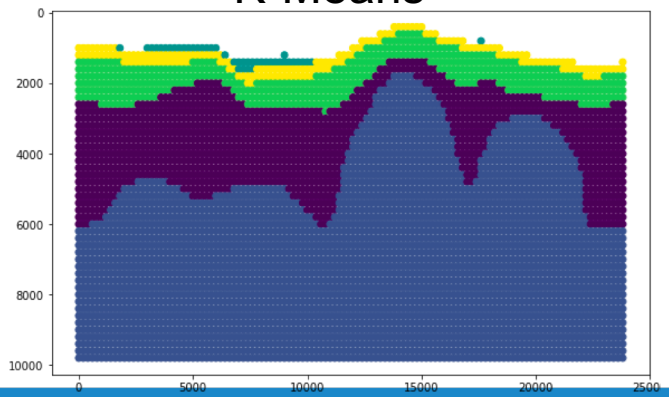
HDBScan



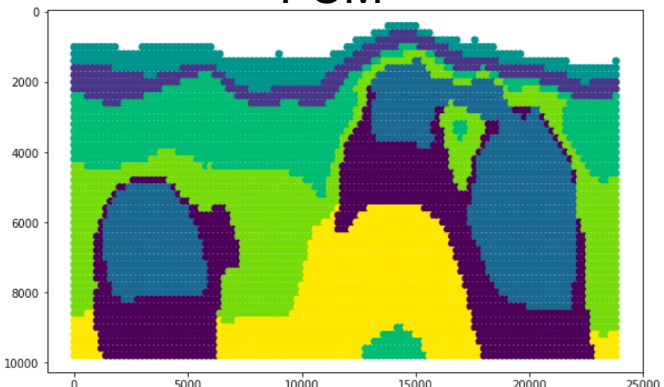
SOMs



K-Means



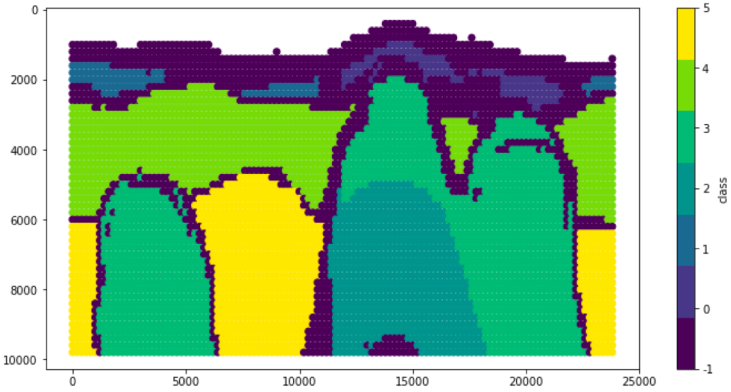
FCM



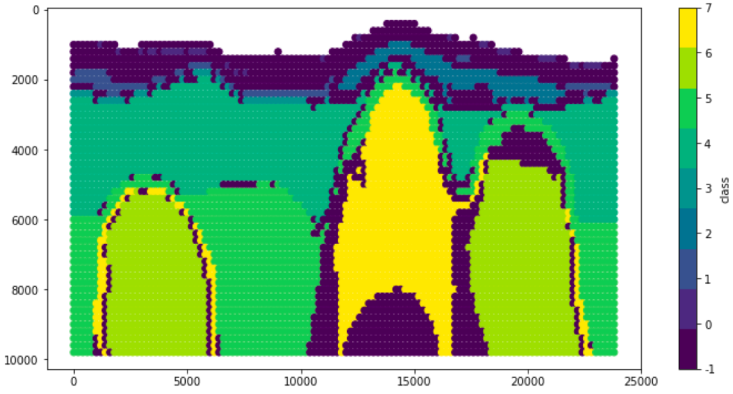
Imperial College London

HDBScan

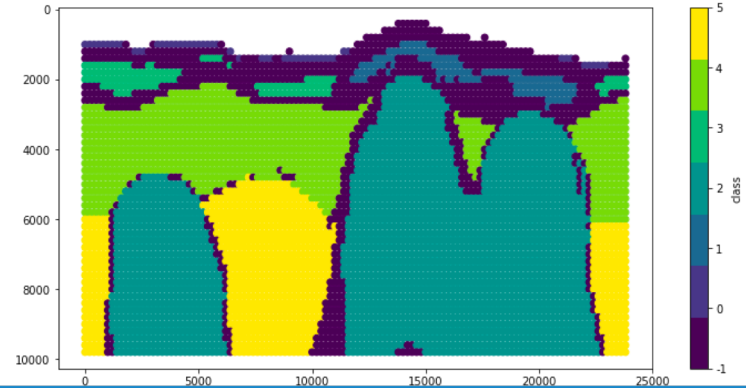
Full Parameters



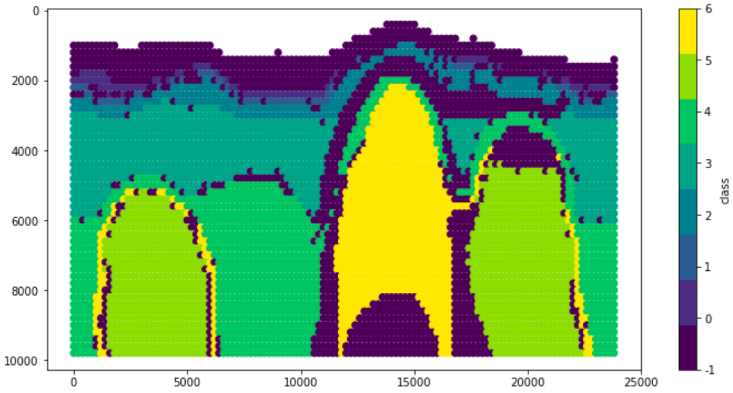
Vp, Vs, Vp/Vs



Vp, Qp, Density

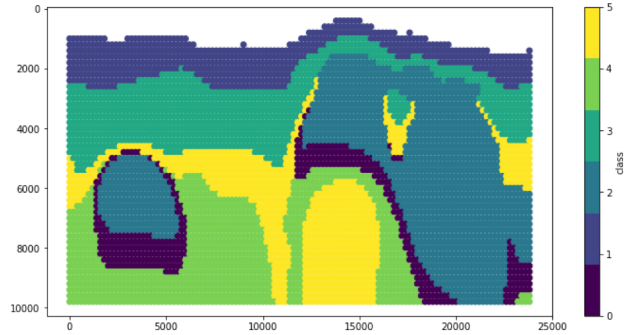


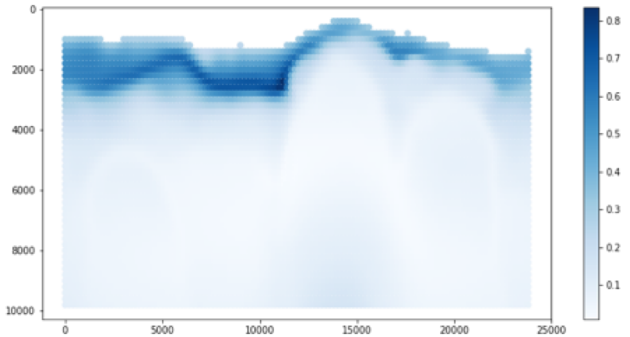
Vp, Vs

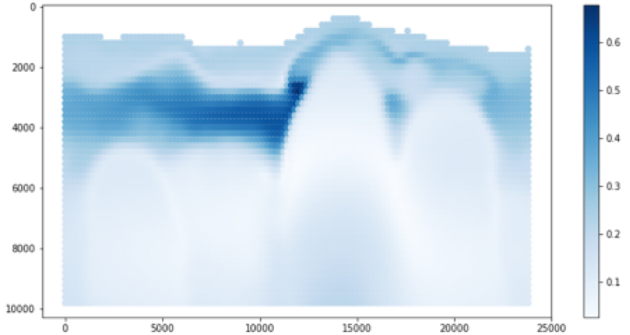


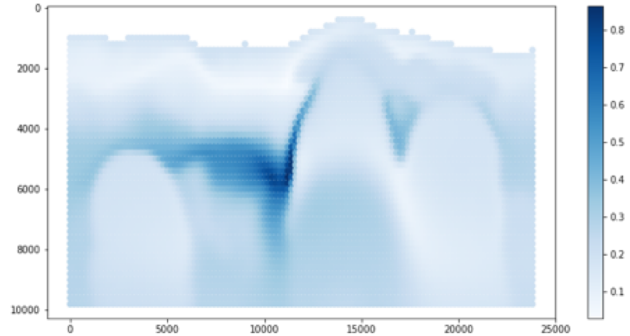
Imperial College London

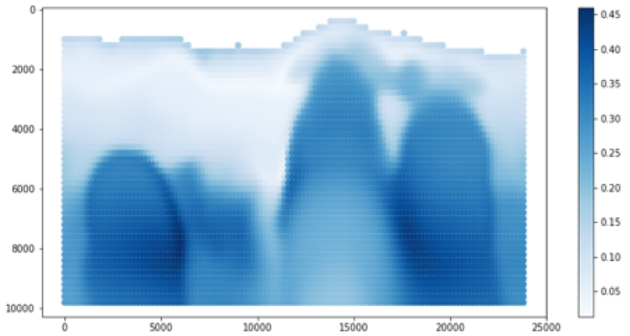
FCM

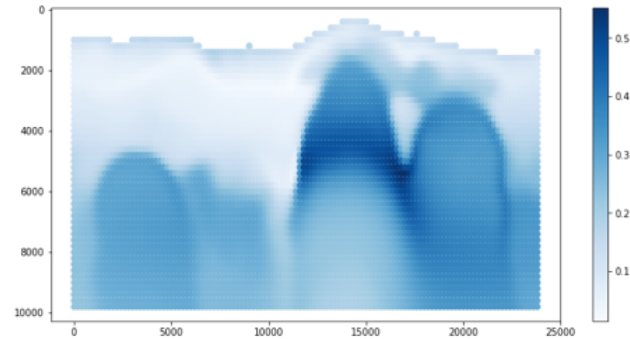






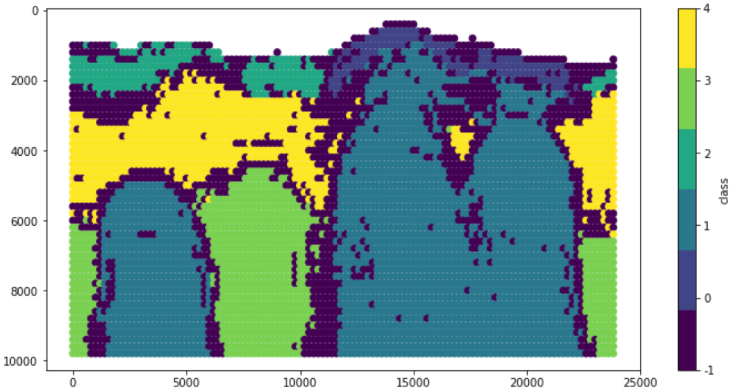




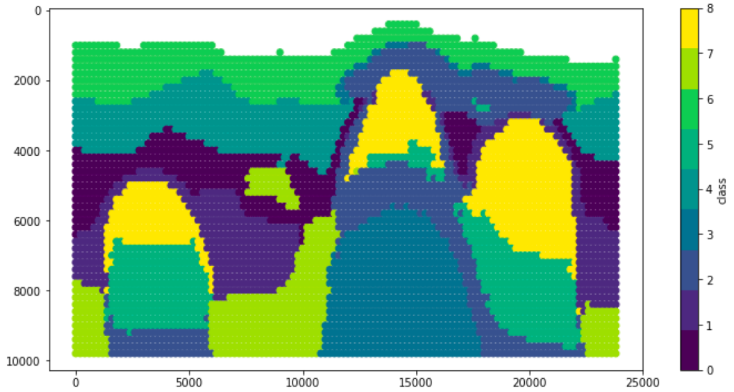


Imperial College London

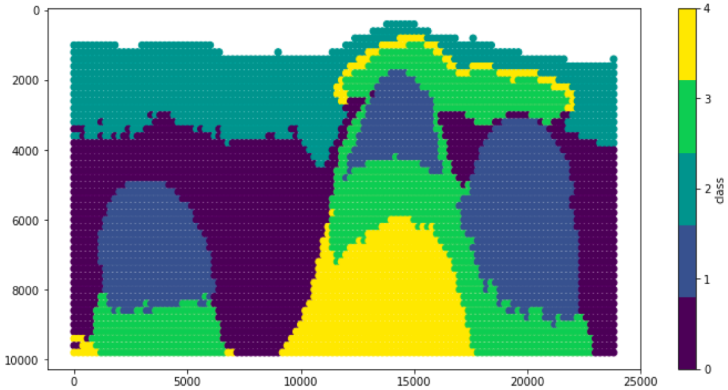
SOMs with other methods



SOMs - HDBScan



SOMs - FCM



SOMs – K-Means

Imperial College London

Verification & Computation time

	SSE	Silhouette Score	Calinski Harabaz Score	Entropy	Purity	Computation time (seconds)
K-Means	2339.9577	0.5741	10793.8328	0.6616	0.7307	0.1543
FCM	8559.9661	0.3253	3695.8394	0.6086	0.7636	80.1082
HDBSCAN	-	0.2145	857.4271	0.3754	0.8379	0.1320
SOMs	-	0.1809	2064.0543	0.6280	0.7181	40.1225
SOMs-HDBScan	-	0.1462	974.3359	0.4864	0.8133	106.3040
SOMs-FCM	140498.0742	0.3453	5388.8095	0.4655	0.7941	40.2202
SOMs-Kmeans	229979.9824	0.4086	5753.7709	0.6616	0.6834	40.4529

Imperial College London

Conclusions

- HDBSCAN works best when there is noise present in the input data and when there are multiple physical parameters to learn from.
- With very smooth fields, a fuzzy method would be preferable.
- SOMs: works in those conditions but have a lower purity score
- For FCM and K-means: better results with three or fewer parameters - works best with V_p, Q_p^{-1}, and density.
- Combining other clustering methods with SOMs works well, improves FCM results, computational time doubles while the validation scores improves very little.

Imperial College London

Thank you ☺