MERC 2025 - Session 4

EXPLORATORY DATA ANALYSIS

Session 4 Outline

- Introduction to Exploratory Data Analysis.
- EDA – Case Study

Introduction to Exploratory Data Analysis (EDA)

QC and Imputation

- Visual Inspection
- Borehole Conditions
- Fill missing data

Basic Stats

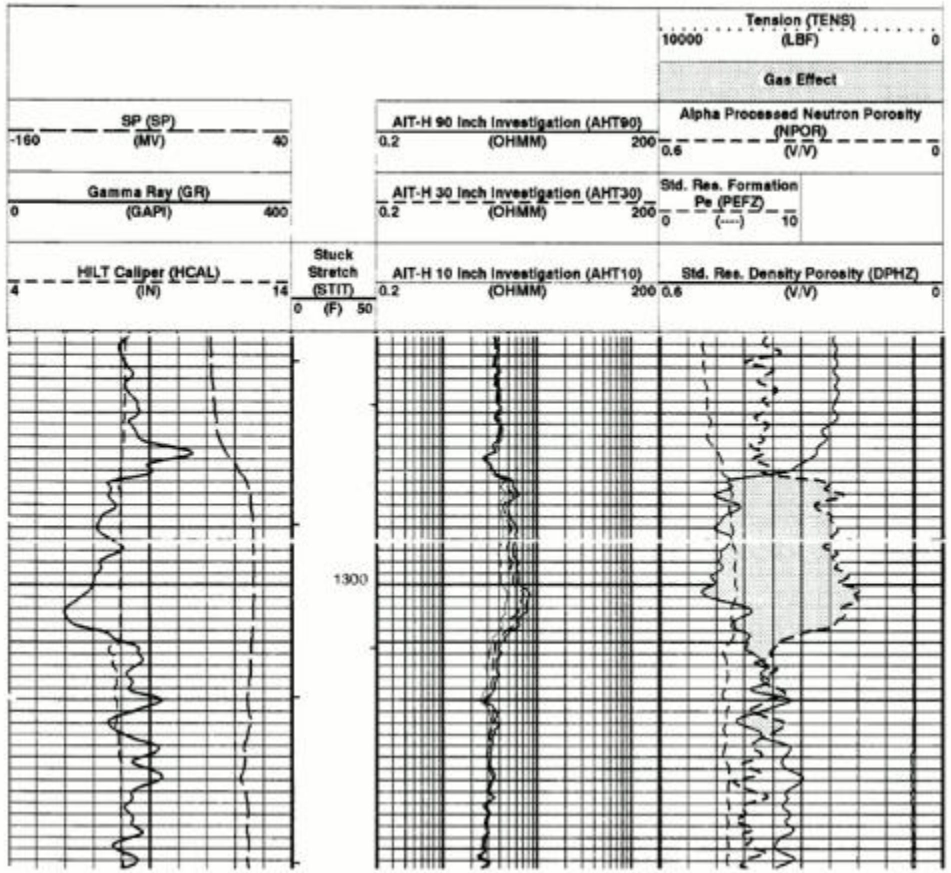
- Distributions
- Categorial Statistics
- Cross-plots
- Box and Violin Plots

Machine Learning Techniqus

- PCA
- t-SNE
- Cluster Analysis
- Self-Organizing Maps (SOMs)

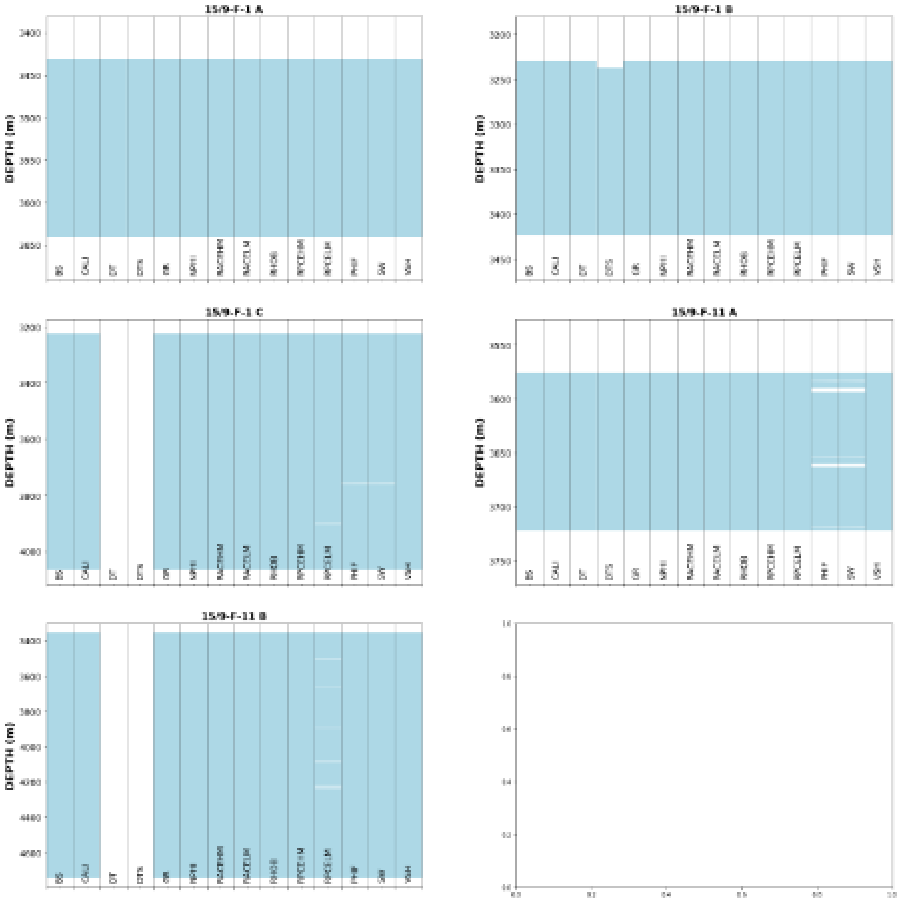
EDA – QA/QC Inspection

- Manual review of the data.
  - Examining for large spikes or unexplained changes in the data.
  - Are the values expected for the environment?
- Baseline Check
  - Are there shifts or changes to the log that are unrelated to geology (borehole or probe related?)
- Composite View
  - Viewing all the logs together to get a holistic view of each hole.



EDA – Filling in the gaps

- In preparation for advanced methods, ensuring there are no gaps in the data is critical.
  - Interpolation
    - Small gaps where obvious trends are continuous.
  - Imputation
    - Using correlated logs to help fill larger gaps.
  - Manual Correction
    - Manually filling in areas based on observed trends.

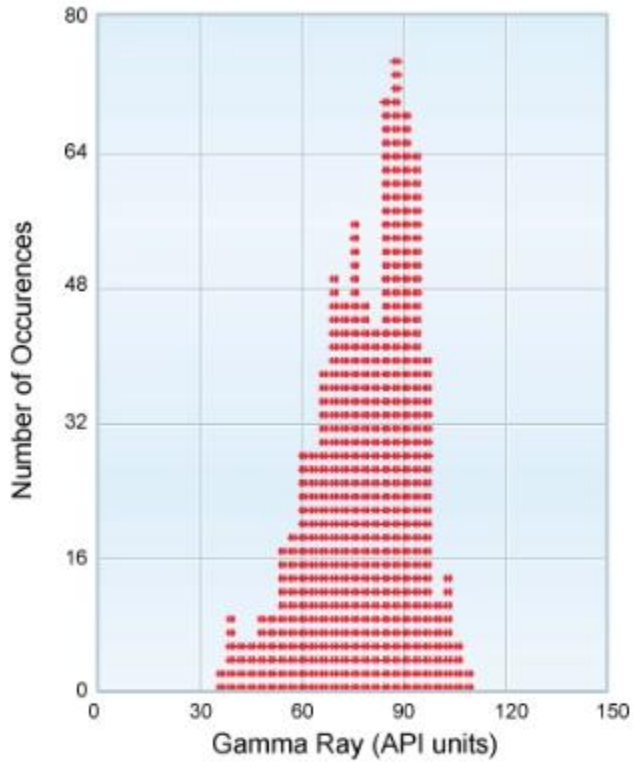


EDA – Examining Distributions

Histograms and Statistics

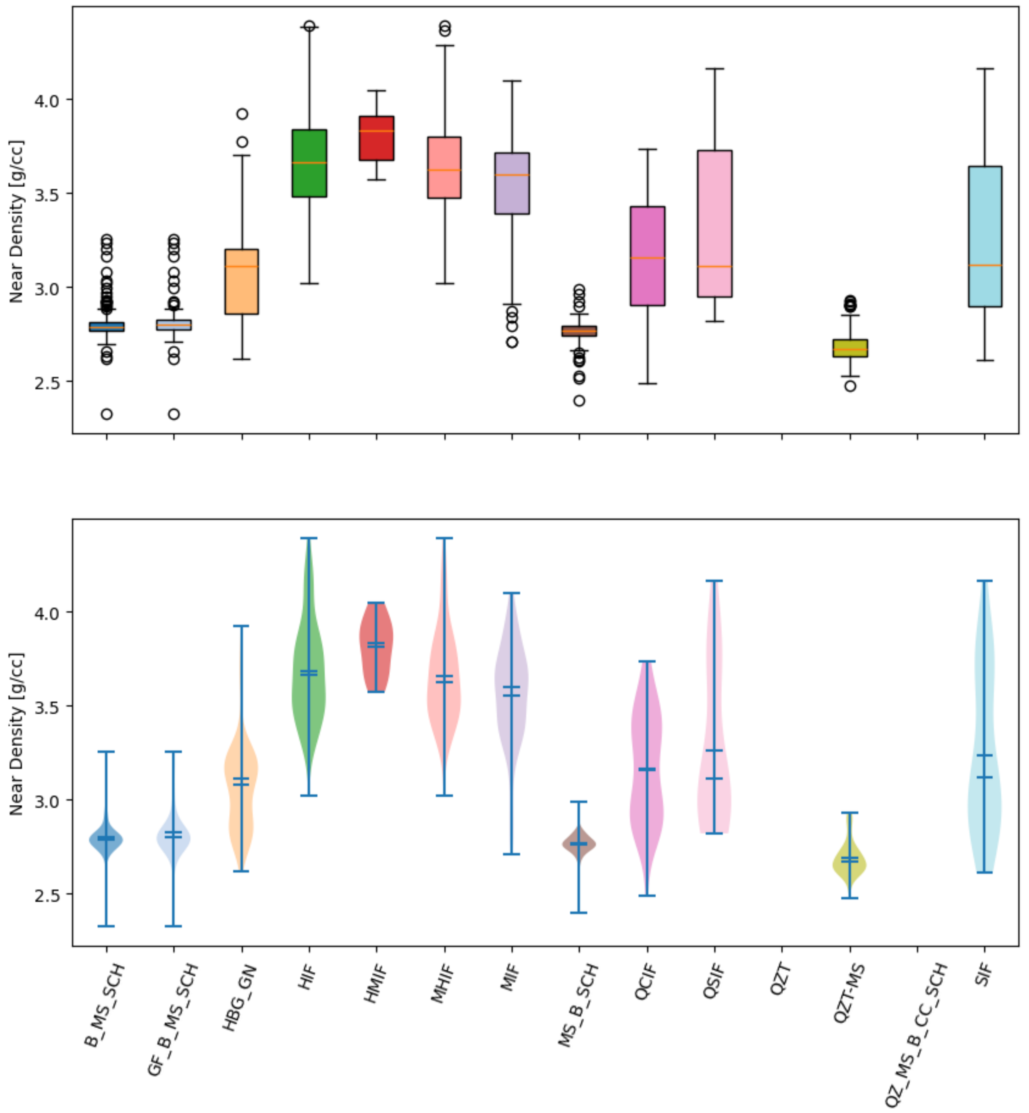
Analyzing the shape of data distribution helps in normalization and lithology identification.

- Modality
  - Bi-Modal or multi-modal distributions often indicate the presence of multiple geologic units or formations.
- Skewness
  - Some parameters (mainly magnetic and electrical) will follow a log-normal distribution and require additional transformation.
- Outliers
  - Values outside the 10th and 90th percentiles should be examined with care as they may indicate errors in the instruments or small (but significant) observations.



EDA – Statistics by Geological Unit

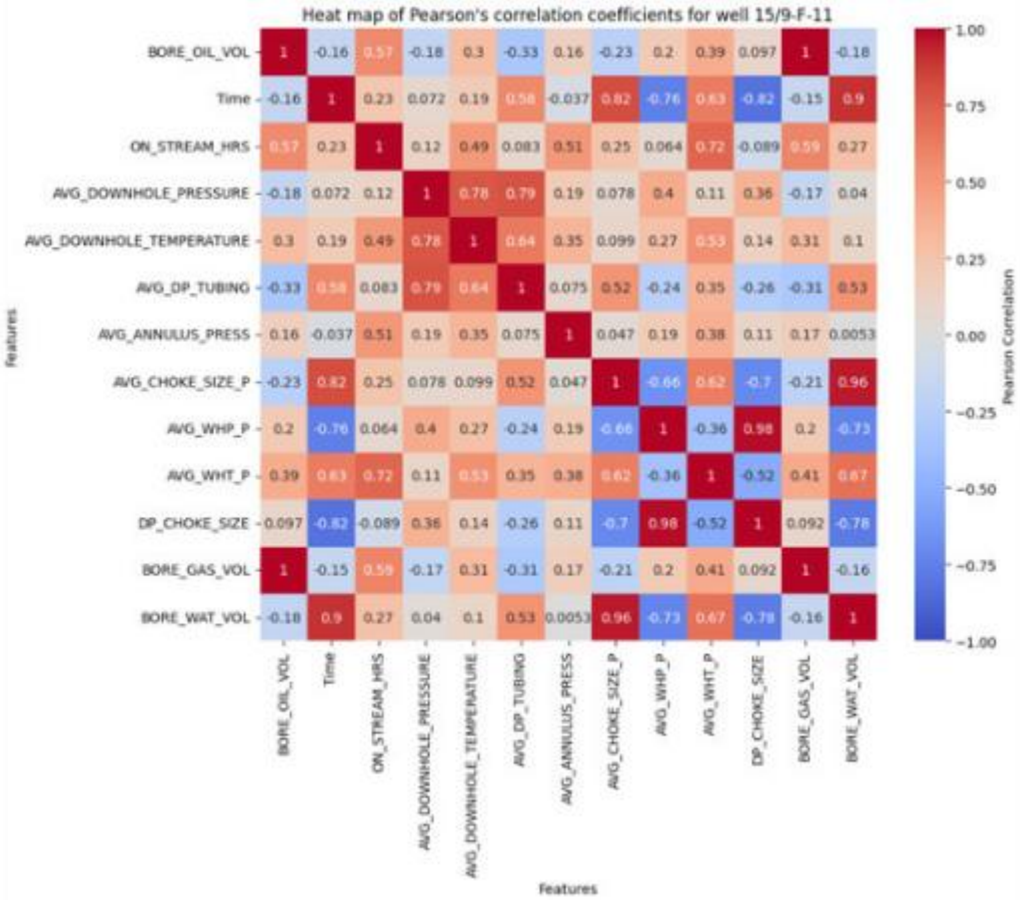
- Segment borehole data based on core-logged ‘from-to’ zones.
- Using visualization such as box or violin plots shows the distribution for each unit.
- Establishes a **characteristic response** for the formation and can be then used in modelling.



EDA – Feature Redundancy Analysis

Are there any redundant parameters?

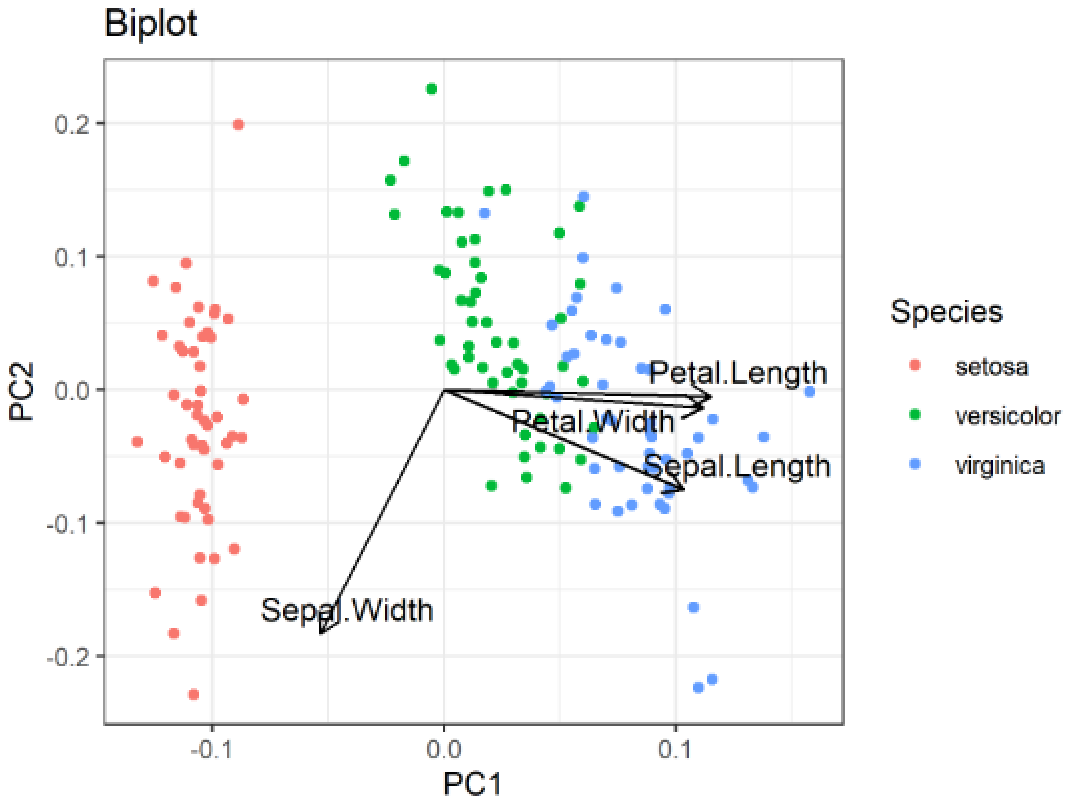
- This is an important consideration before using advanced machine learning techniques as it can introduce bias.
- Correlation Matrix
    - Heat-map showing the Pearson correlation coefficient between parameters.
        - Drop highly co-linear features in improve performance.



	BORE_OIL_VOL	Time	ON_STREAM_HRS	AVG_DOWNHOLE_PRESSURE	AVG_DOWNHOLE_TEMPERATURE	AVG_DP_TUBING	AVG_ANNULUS_PRESS	AVG_CHOKE_SIZE_P	AVG_WHP_P	AVG_WHT_P	DP_CHOKE_SIZE	BORE_GAS_VOL	BORE_WAT_VOL
BORE_OIL_VOL	1	-0.16	0.57	-0.18	0.3	-0.33	0.16	-0.23	0.2	0.39	0.097	1	-0.18
Time	-0.16	1	0.23	0.072	0.19	0.58	-0.037	0.82	-0.76	0.63	-0.82	-0.15	0.9
ON_STREAM_HRS	0.57	0.23	1	0.12	0.49	0.083	0.51	0.25	0.064	0.72	-0.089	0.59	0.27
AVG_DOWNHOLE_PRESSURE	-0.18	0.072	0.12	1	0.78	0.79	0.19	0.078	0.4	0.11	0.36	-0.17	0.04
AVG_DOWNHOLE_TEMPERATURE	0.3	0.19	0.49	0.78	1	0.64	0.35	0.099	0.27	0.53	0.14	0.31	0.1
AVG_DP_TUBING	-0.33	0.58	0.083	0.79	0.64	1	0.075	0.52	-0.24	0.35	-0.26	-0.31	0.53
AVG_ANNULUS_PRESS	0.16	-0.037	0.51	0.19	0.35	0.075	1	0.047	0.19	0.38	0.11	0.17	0.0053
AVG_CHOKE_SIZE_P	-0.23	0.82	0.25	0.078	0.099	0.52	0.047	1	-0.66	0.62	-0.7	-0.21	0.96
AVG_WHP_P	0.2	-0.76	0.064	0.4	0.27	-0.24	0.19	-0.66	1	-0.36	0.98	0.2	-0.73
AVG_WHT_P	0.39	0.63	0.72	0.11	0.53	0.35	0.38	0.62	-0.36	1	-0.52	0.41	0.67
DP_CHOKE_SIZE	0.097	-0.82	-0.089	0.36	0.14	-0.26	0.11	-0.7	0.98	-0.52	1	0.092	-0.78
BORE_GAS_VOL	1	-0.15	0.59	-0.17	0.31	-0.31	0.17	-0.21	0.2	0.41	0.092	1	-0.16
BORE_WAT_VOL	-0.18	0.9	0.27	0.04	0.1	0.53	0.0053	0.96	-0.73	0.67	-0.78	-0.16	1

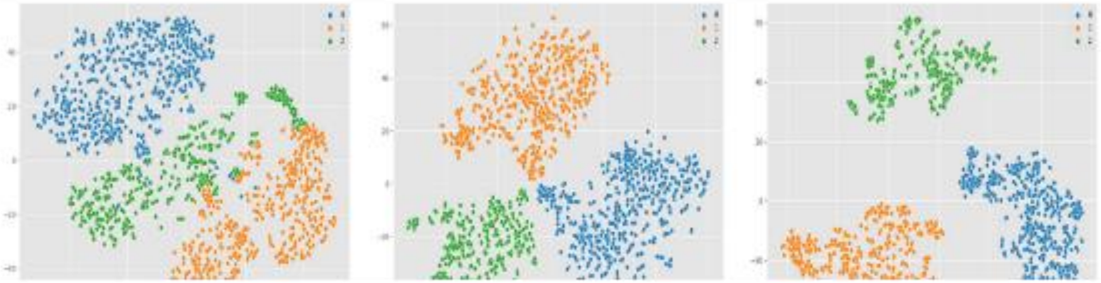
EDA – Principal Component Analysis (PCA)

- Linear Dimension Reduction
  - Borehole physical properties are generally multi-parameter (greater than 4) and understanding their relationship can be difficult.
  - PCA reduces the data set into the 2 or 3 components that can be compared with each other.
  - Assumes that all the parameters are **linear**.



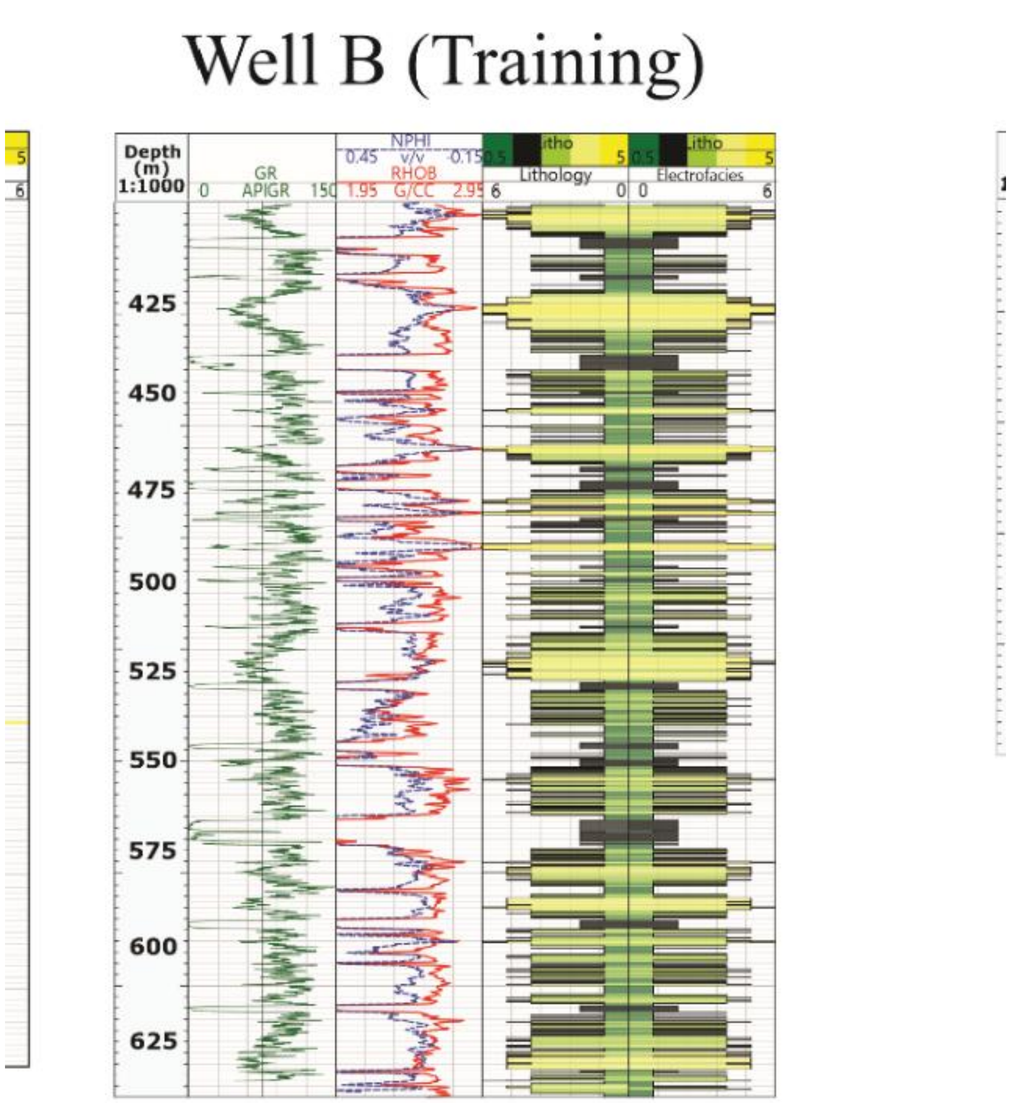
EDA – t-SNE Dimension Reduction

- Like PCA but uses a technique called **Manifold Learning** to reduce multi-dimensional data into 2 or 3 components.
- It's a stochastic method, such that each time it's performed it will be different.
- Is a non-linear technique and can often identify trends that are not apparent compared to PCA.



EDA – Cluster Analysis

- An alternative interpretation that defines new **domains** based on the natural divisions in the borehole logs.
- Different methods are available such as:
  - K-Means
  - DBSCAN
  - HDSCAN

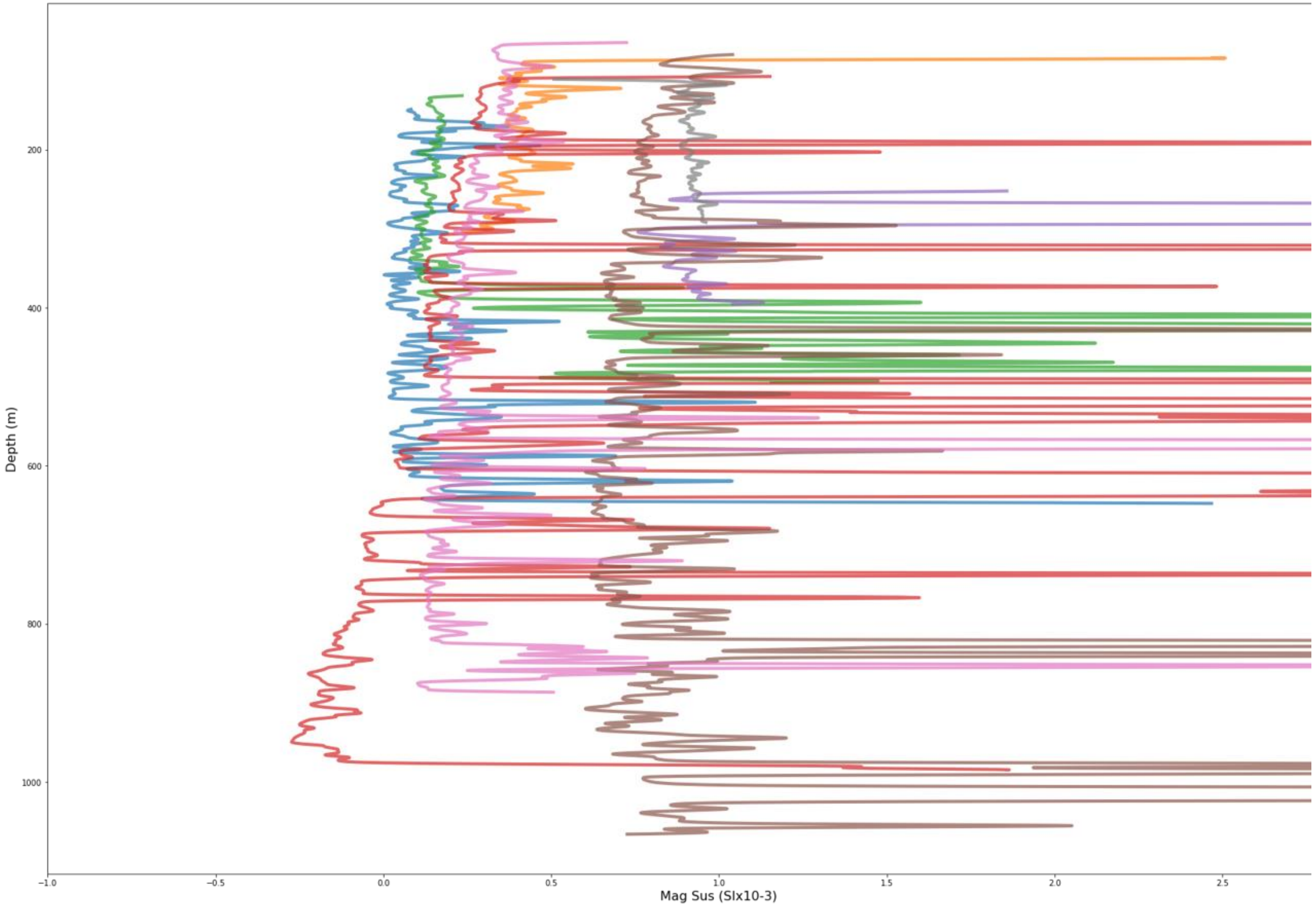


Physical Property Statistical Analysis Case Study

- 11 holes of physical property data acquired by third party with the following parameters:
  - Magnetic Susceptibility
  - Induction Conductivity
  - Induced Polarization
  - Full waveform sonic
  - Spectral Gamma
- Goal of the project was to examine the data quality, examine the relationships between the parameters and geology, and perform a cluster analysis.

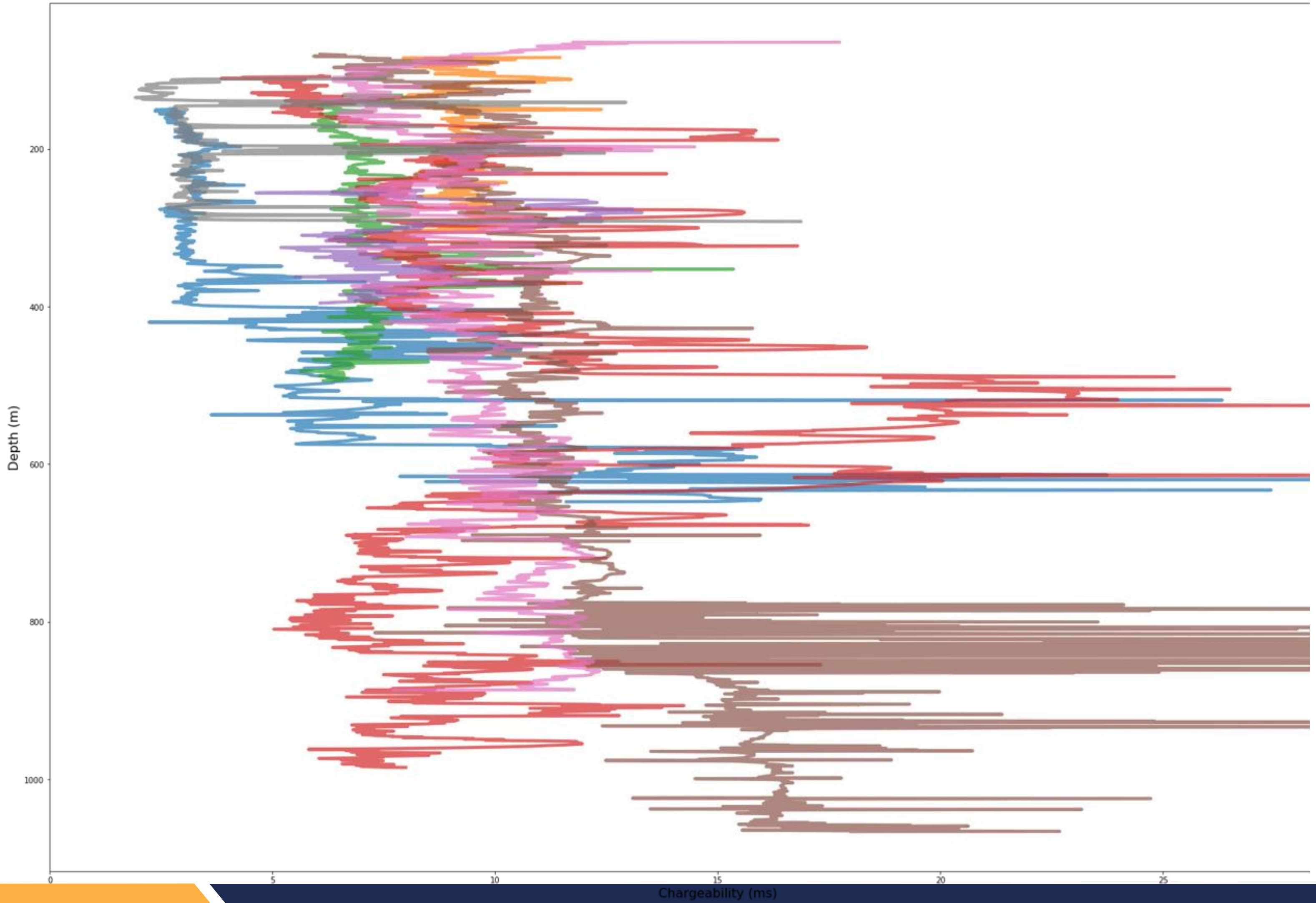
Data Quality Check – Magnetic Susceptibility

DGI GEOSCIENCE



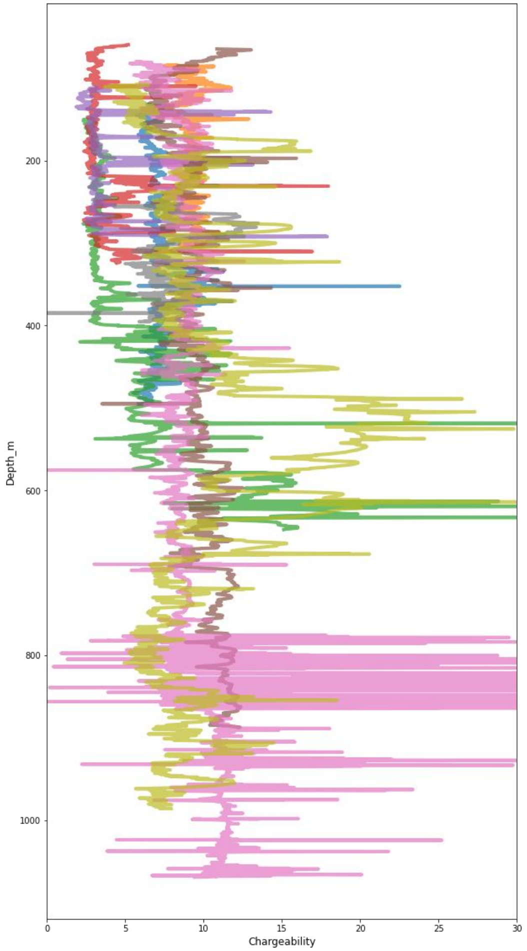
Data Quality Check - IP

DGI
GEOSCIENCE

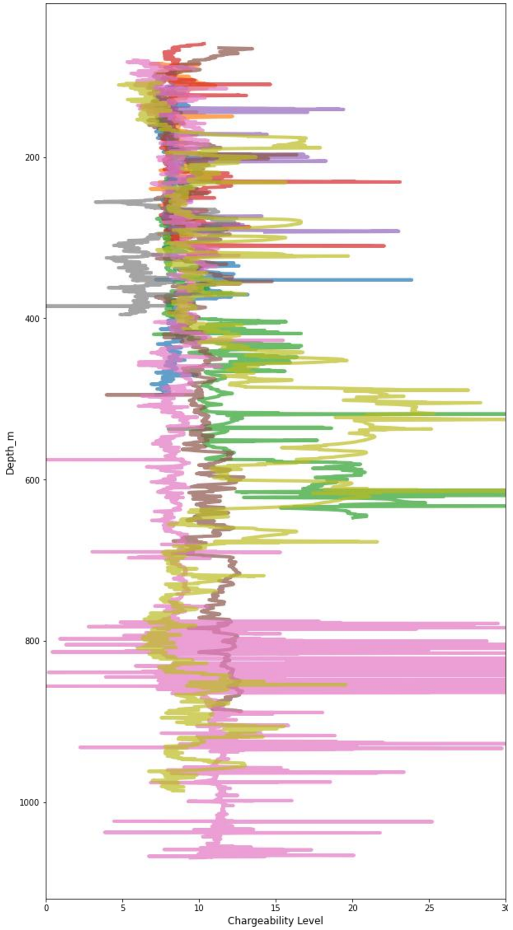


Data Preparation: Level Chargeability Data

Original Data

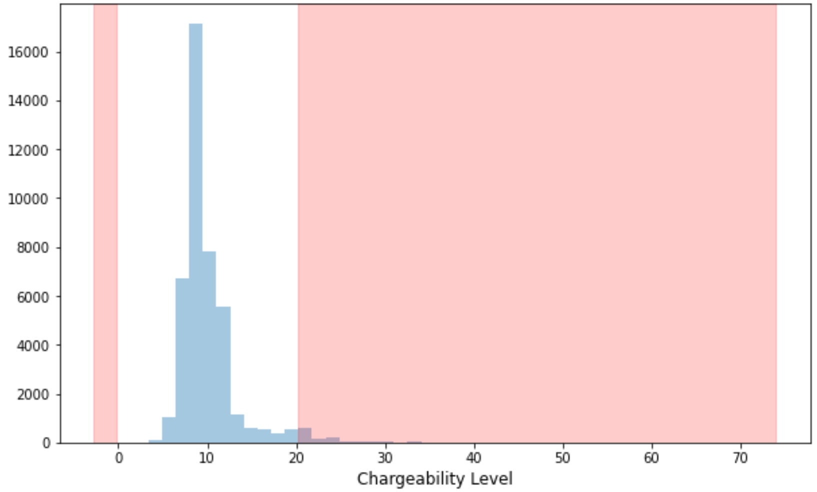


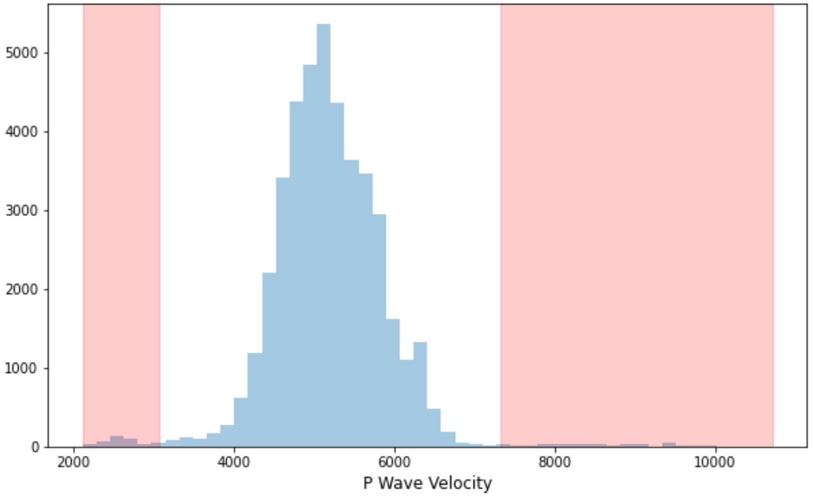
Leveled Data

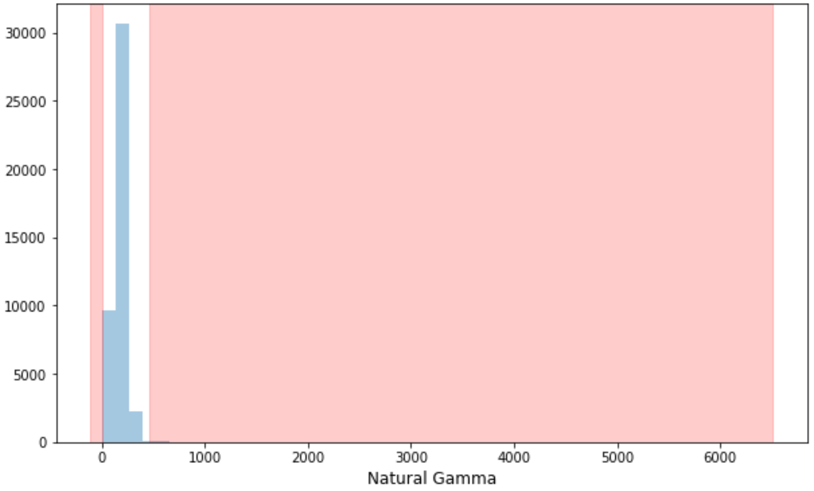


Data Preparation: Remove Outliers

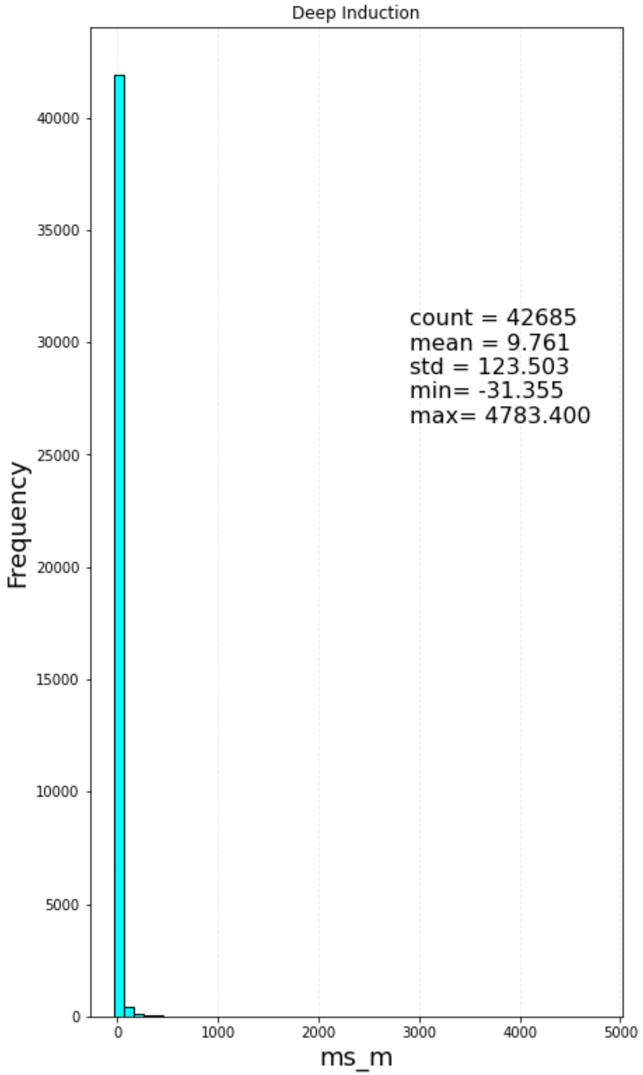
- Chargeability Capped at 20.
- P Wave Velocity Capped at (3000 & 7300)
- Natural Gamma capped at 450.

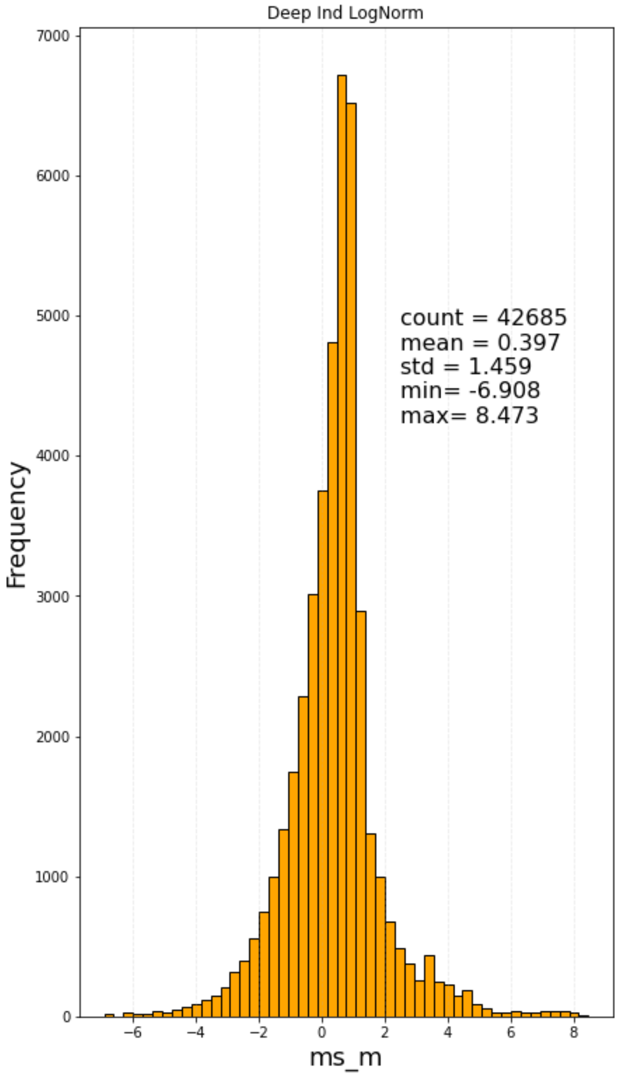




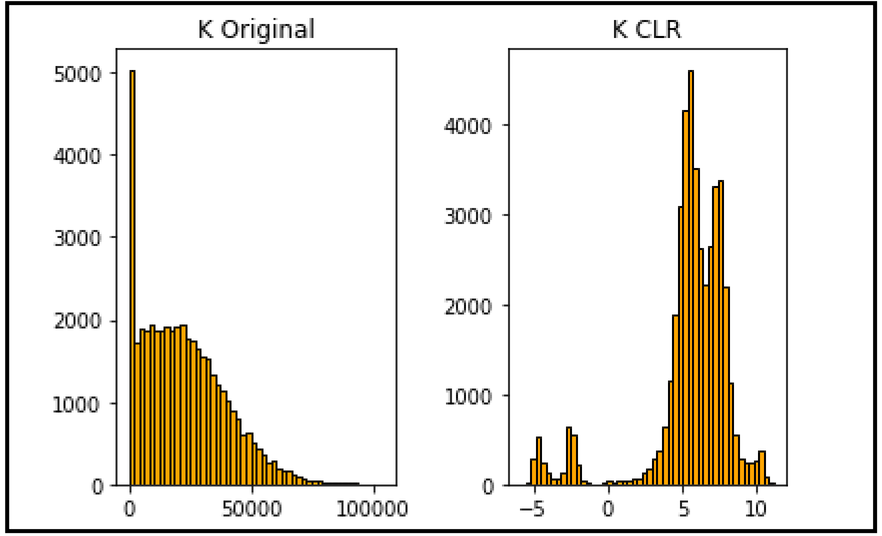


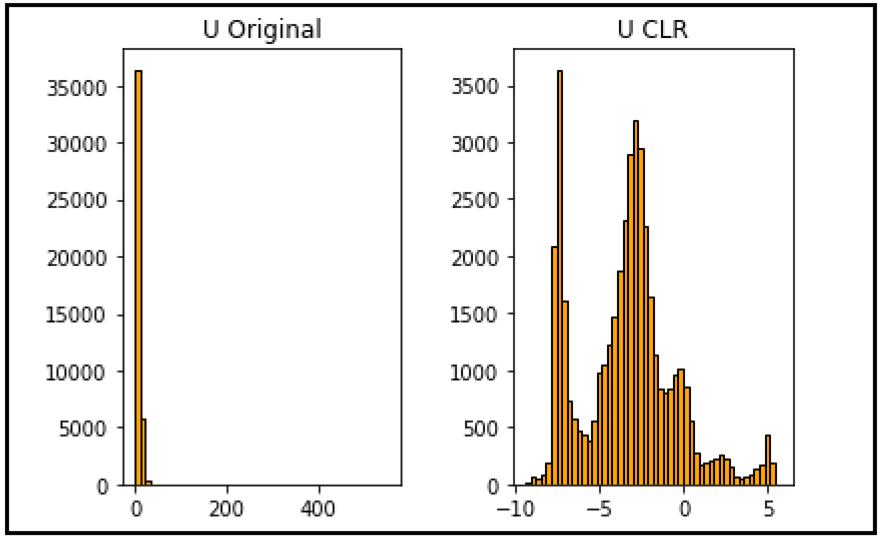
Data Preparation: Log Transform Induction

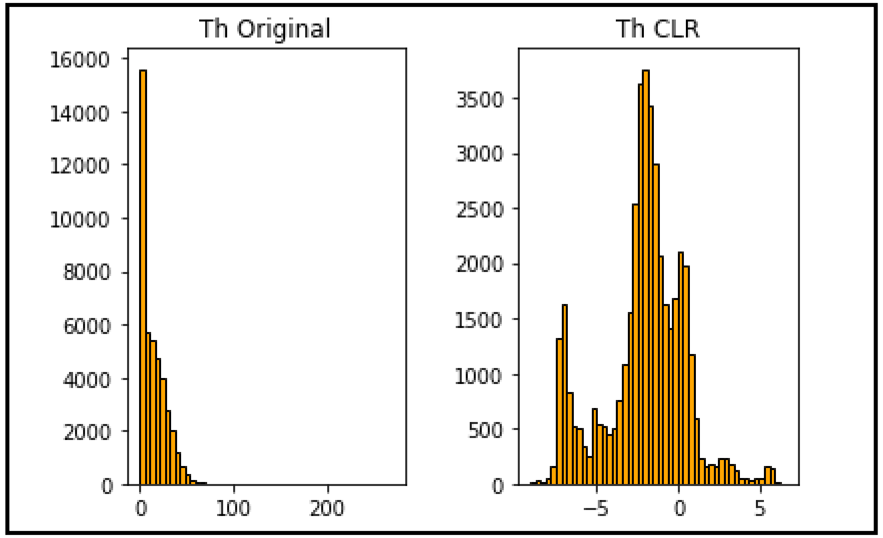




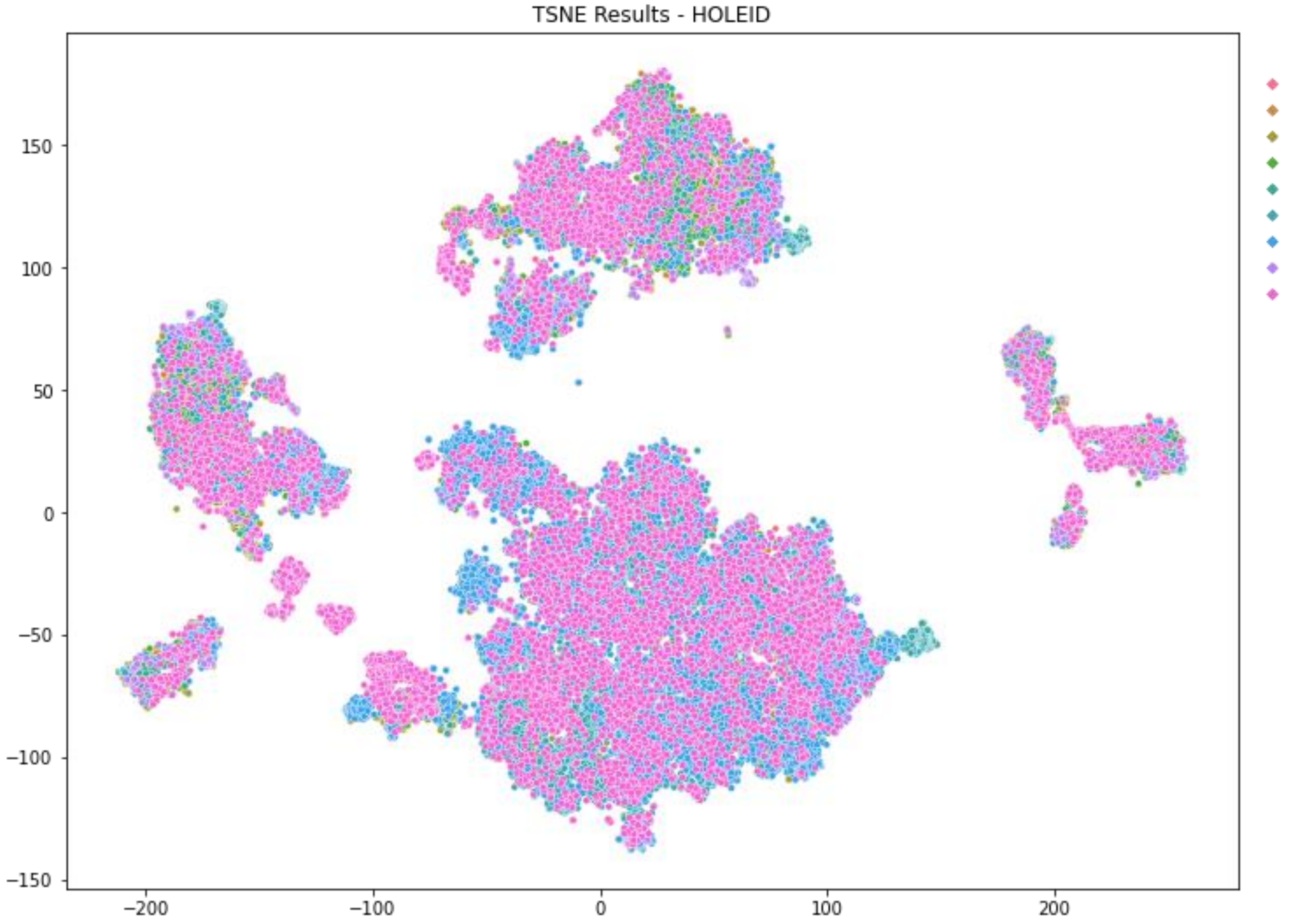
Data Preparation: Compositional Scaling of Spectral Gamma





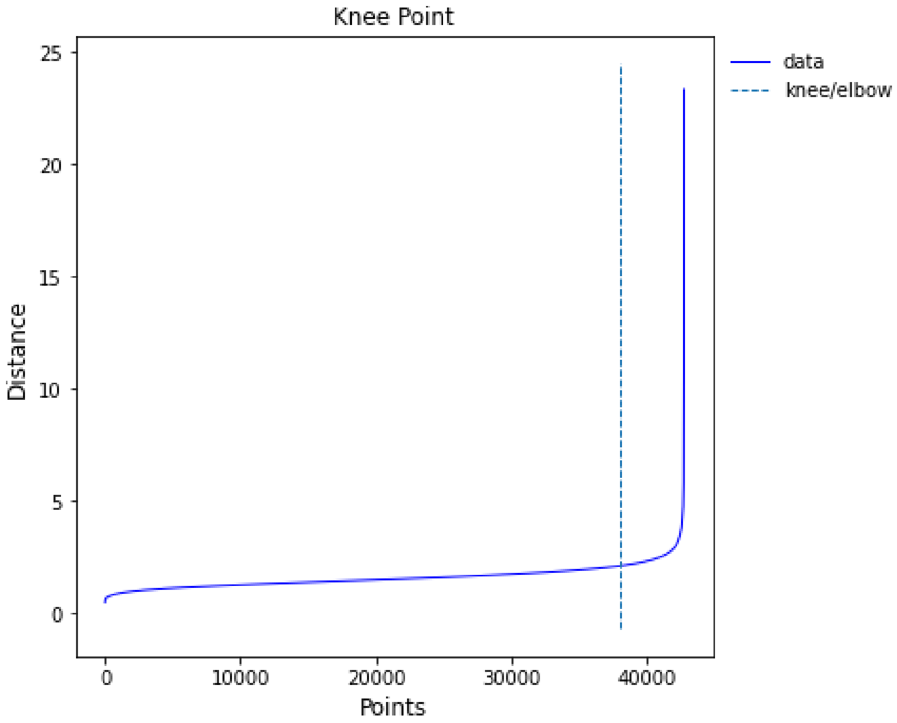


Dimensional Reduction using TSNE

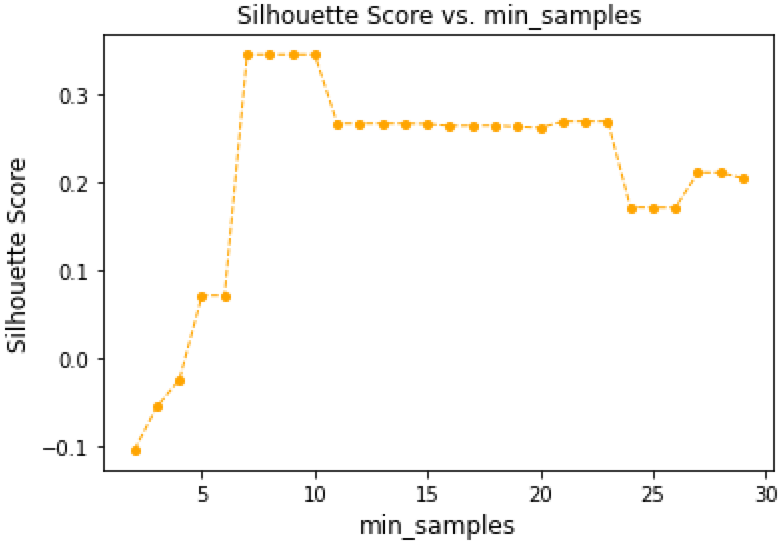


Clustering Tuning Parameters

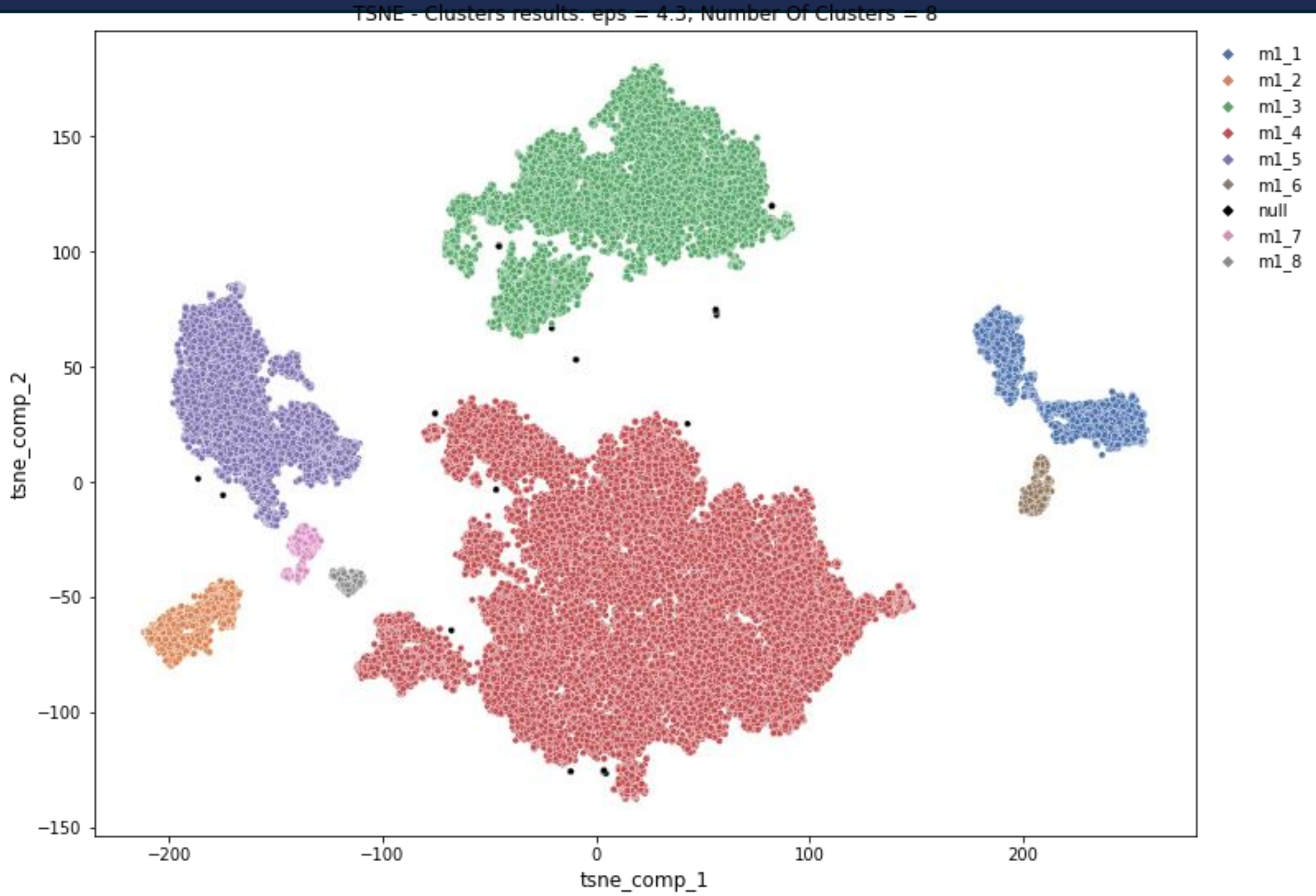
Epsilon Value



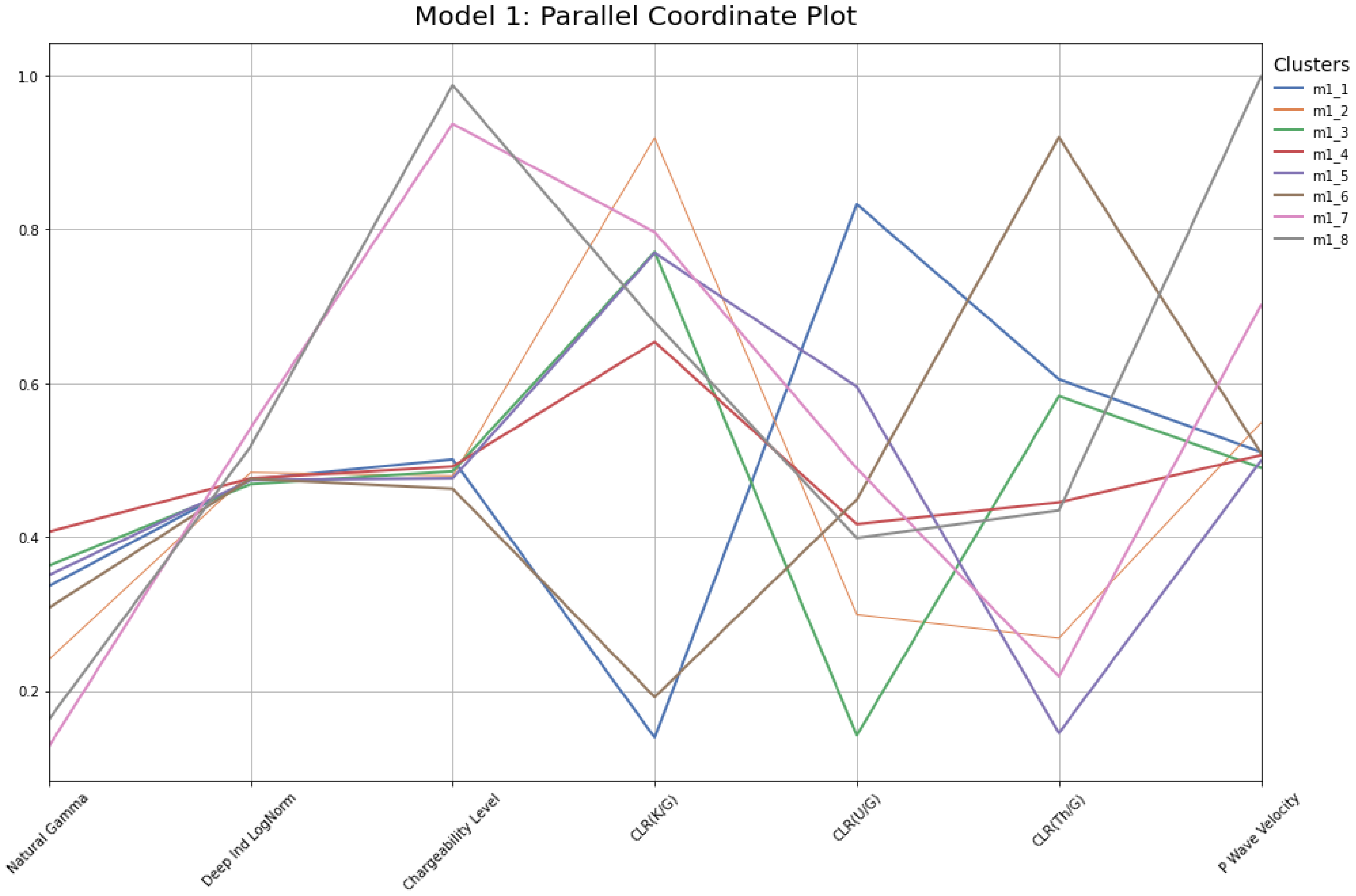
Minimum Samples



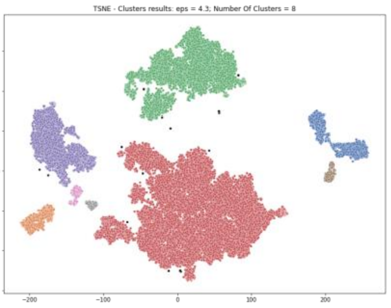
Cluster Analysis: 8 Clusters in the model

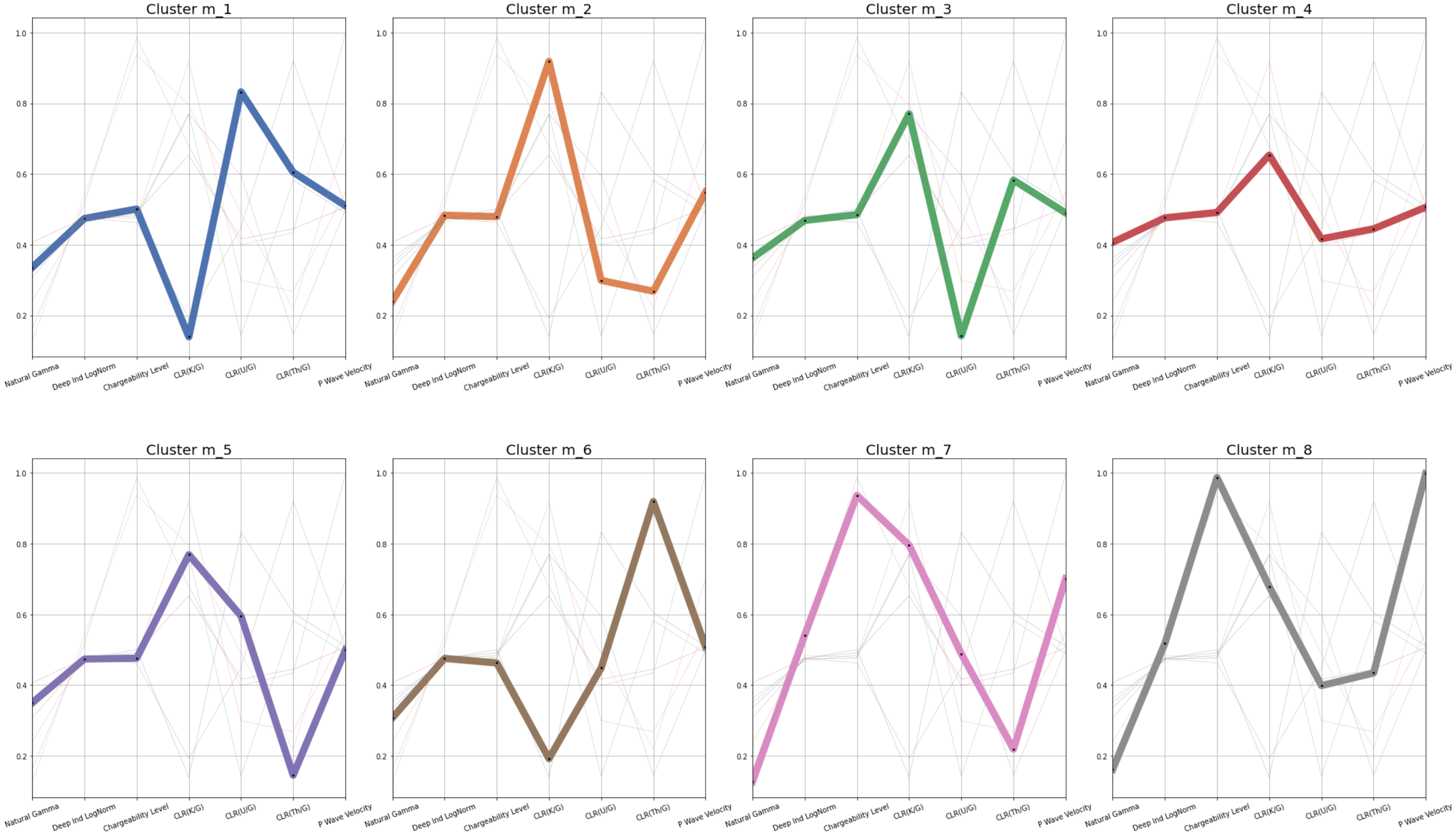


Parallel Coordinate Plot



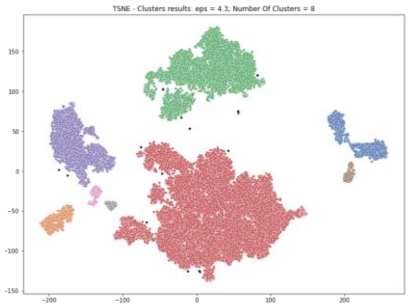
Parallel Coordinate Plots of 8 Clusters

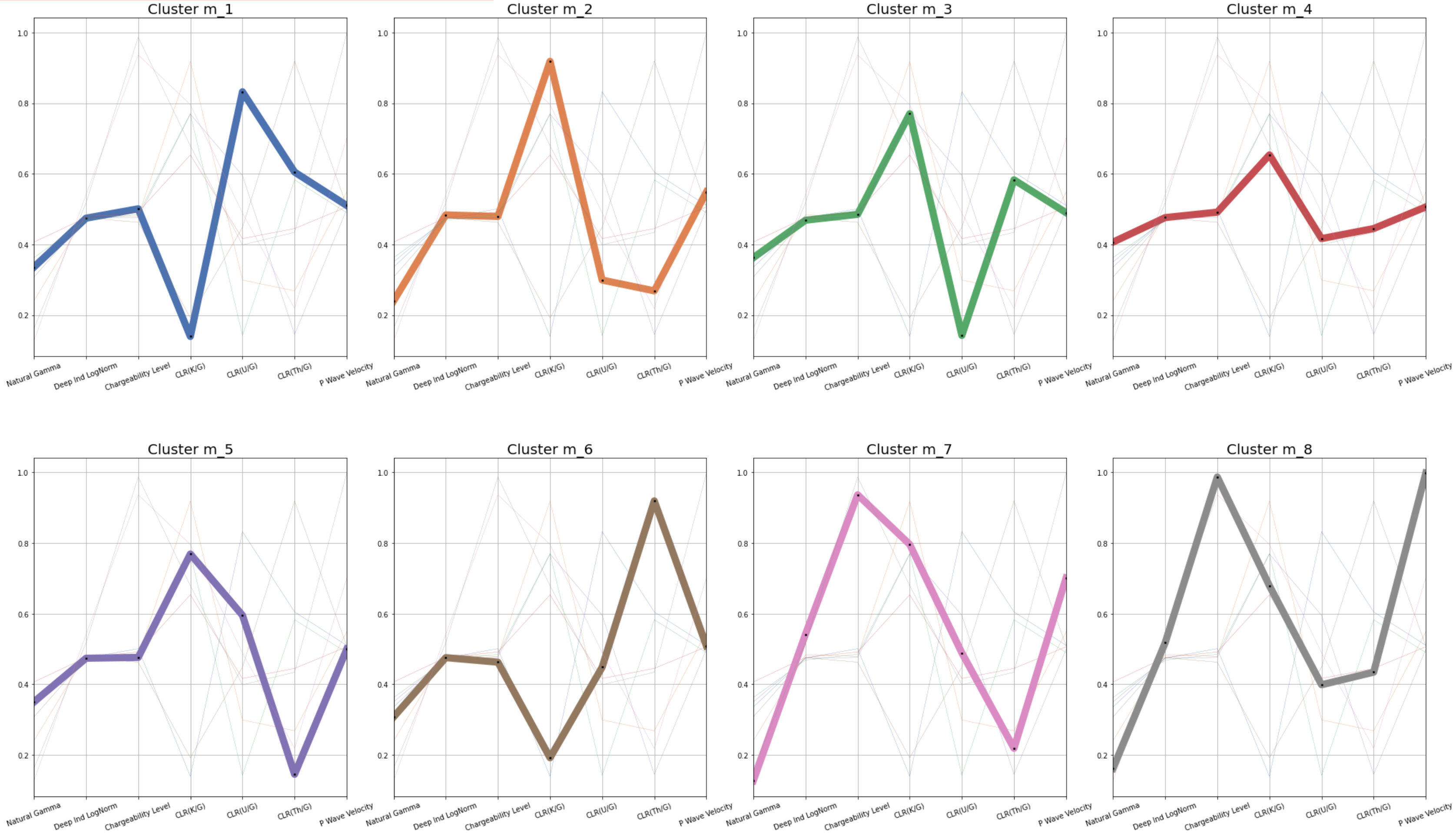




Parallel Coordinate Plots of 8 Clusters

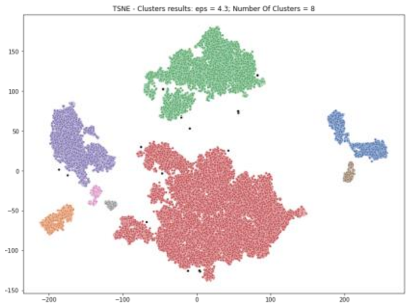
Cluster 1 shows a low K and high U.

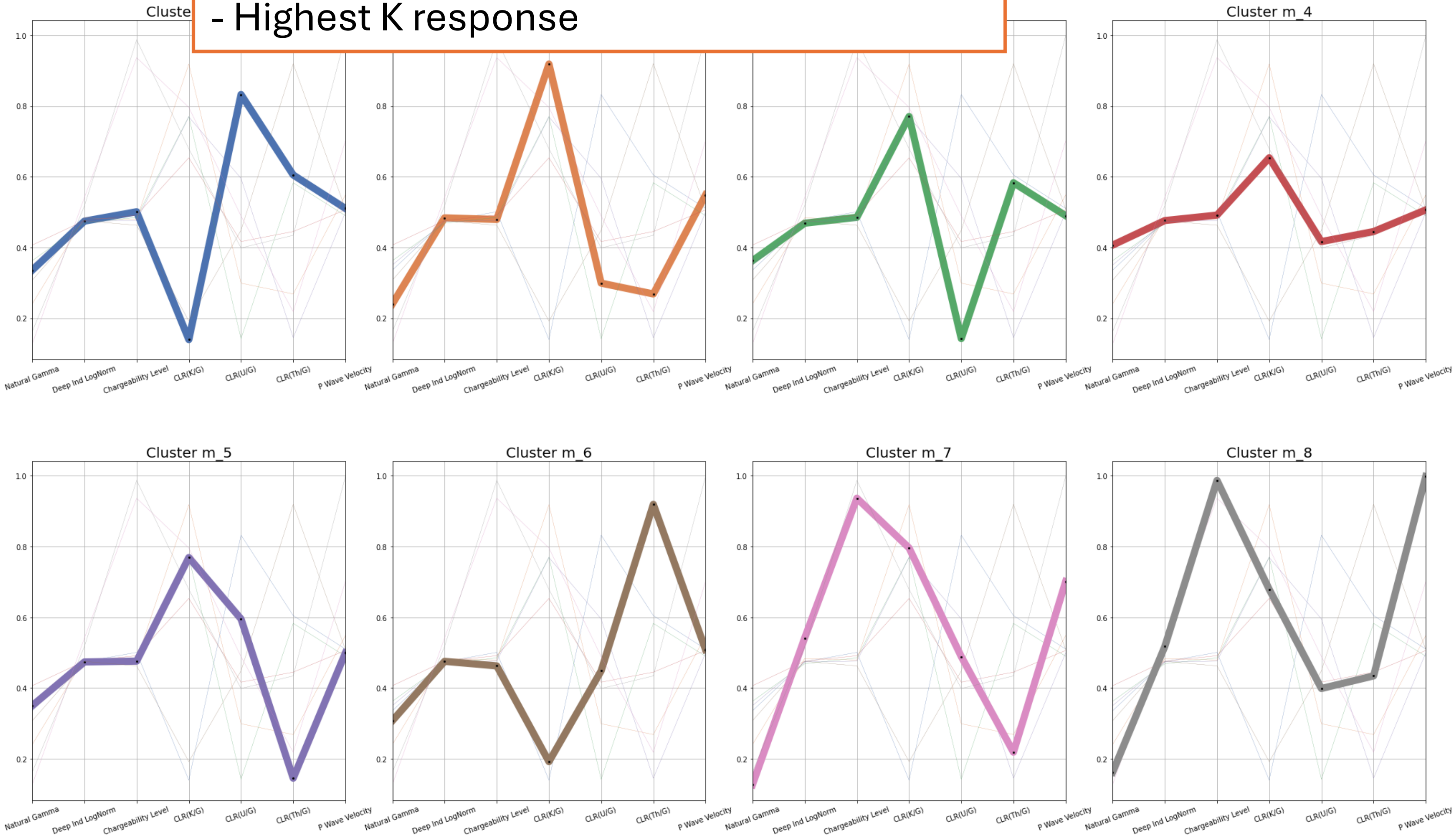




Parallel Coordinate Plots of 8 Clusters

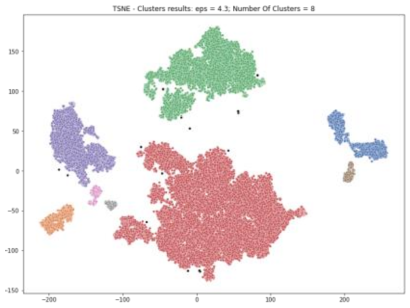
Cluster 2 is one of the smaller groups.
- Highest K response

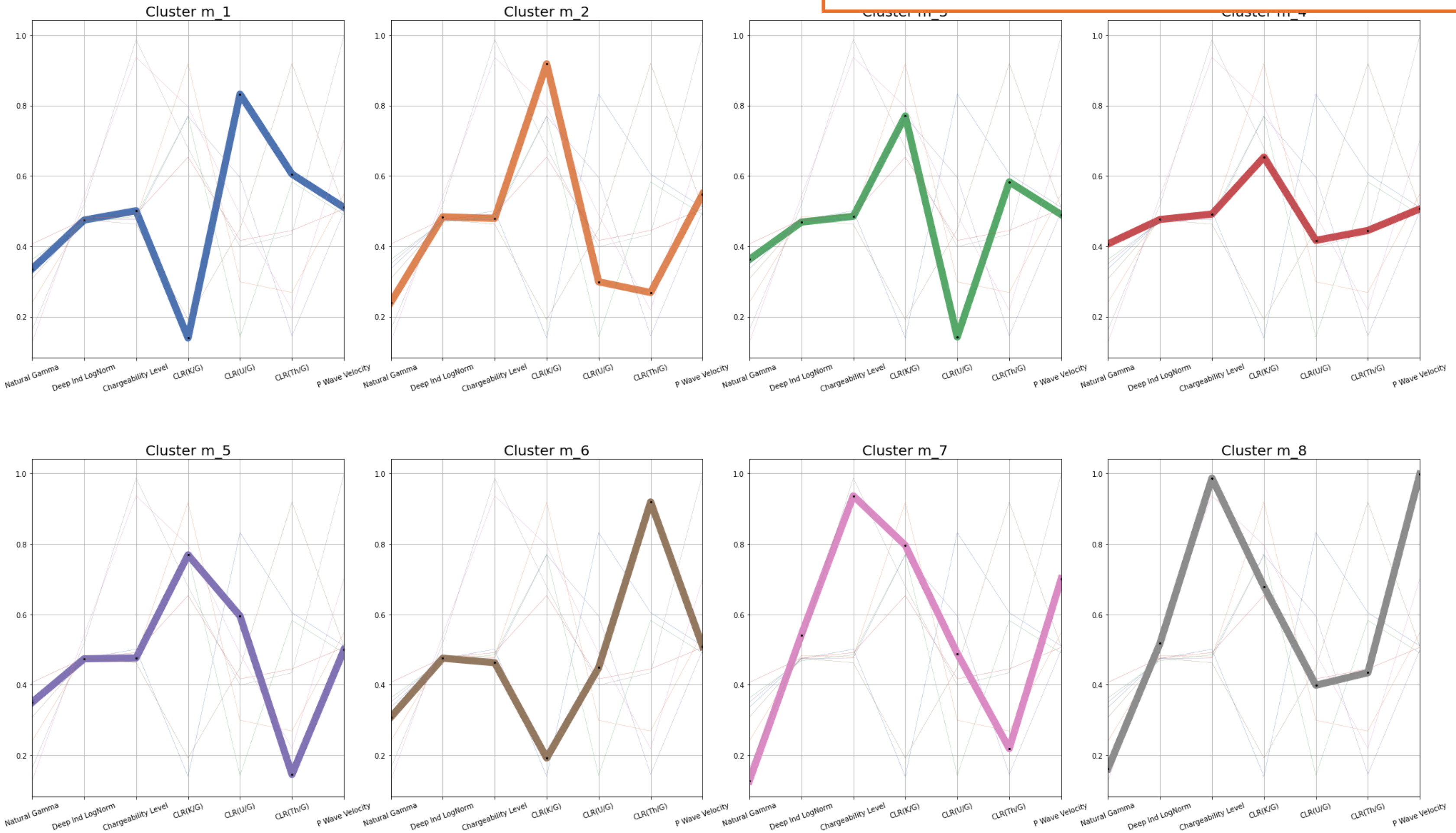




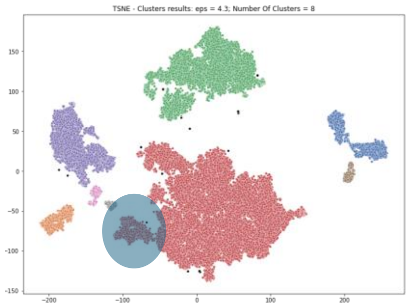
Parallel Coordinate Plots of 8 Clusters

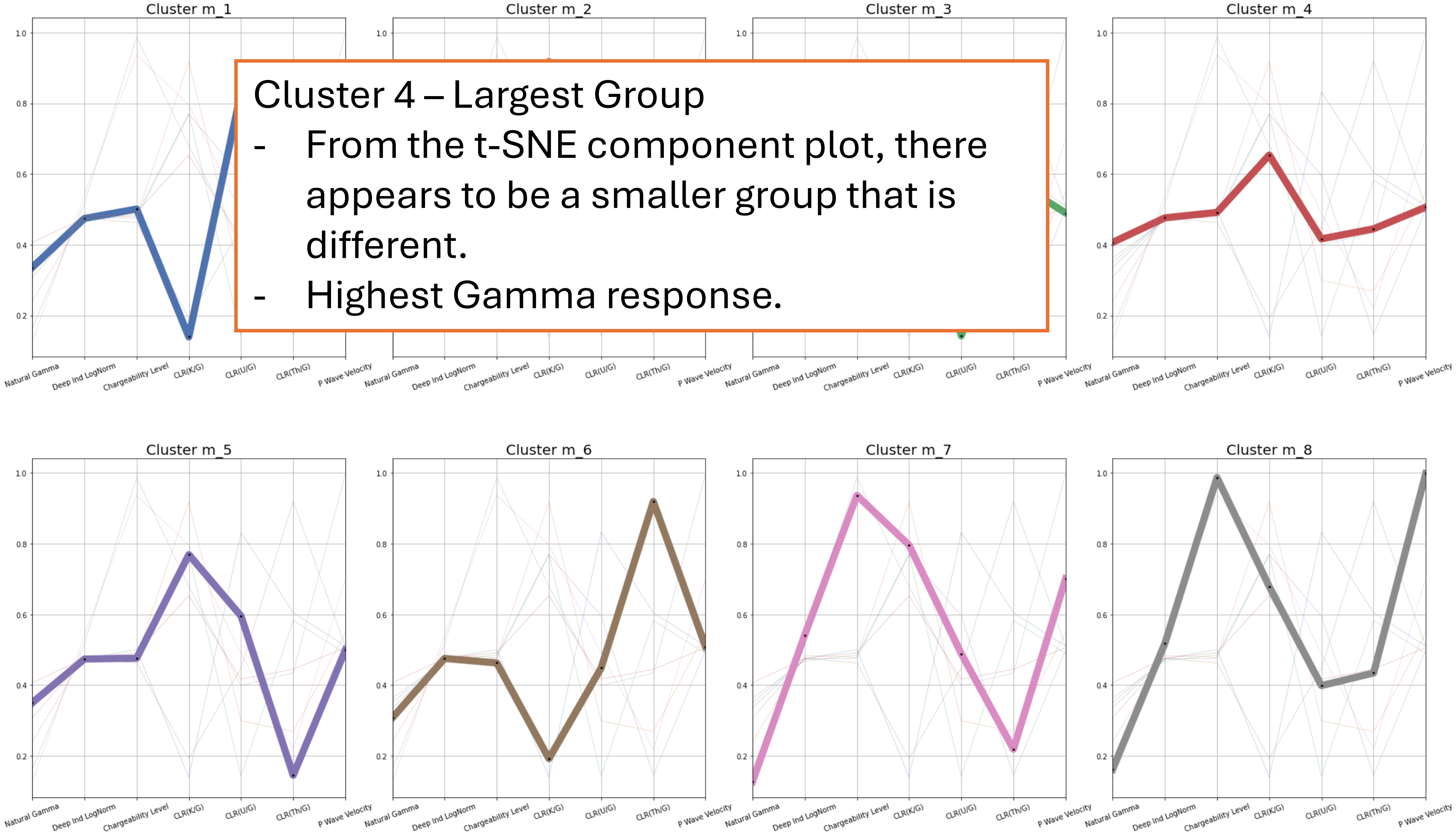
Cluster 3 – Low U response.





Parallel Coordinate Plots of 8 Clusters

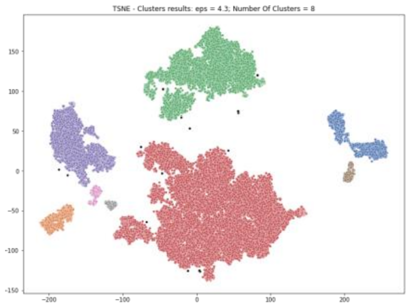


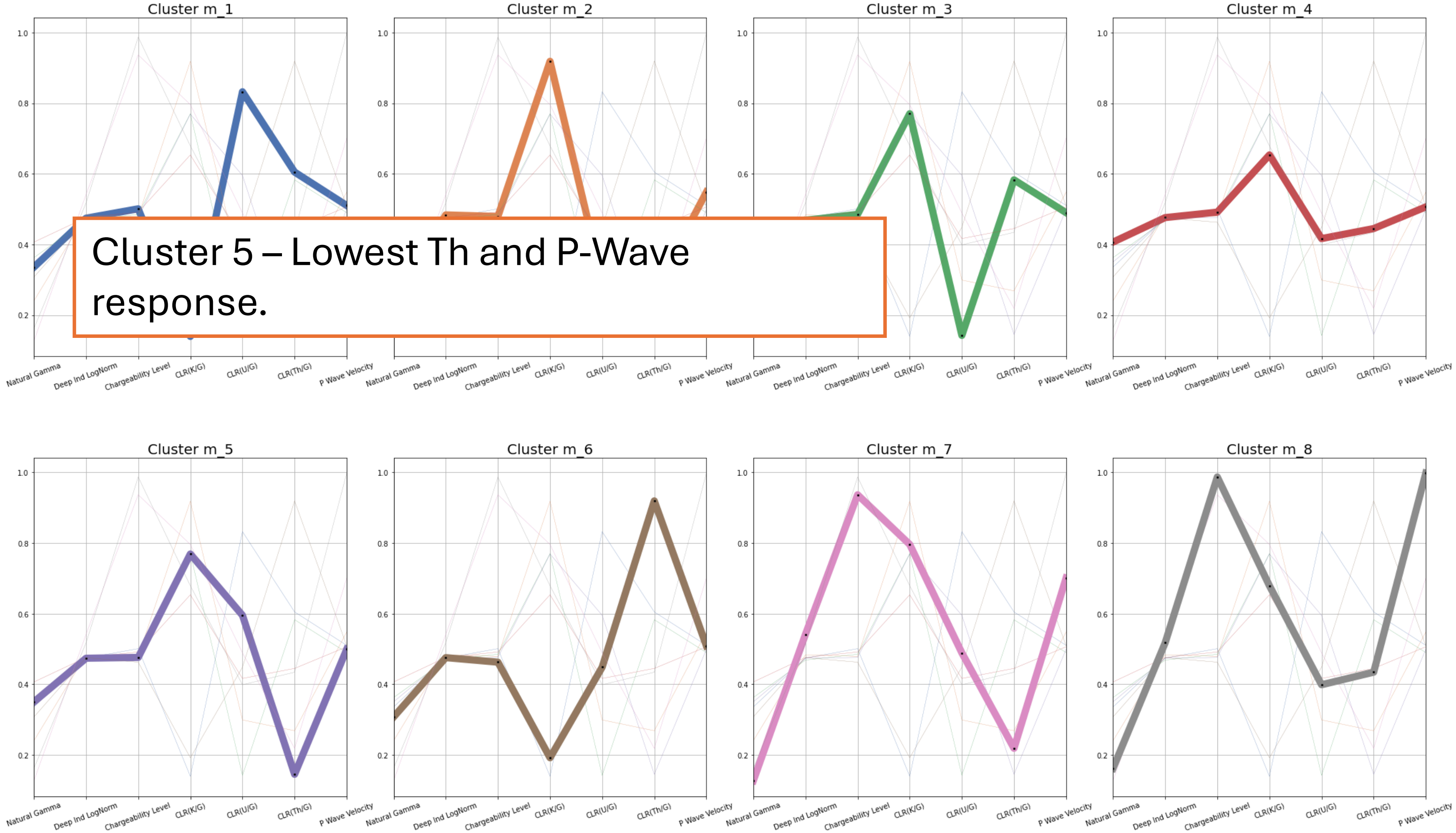


Cluster 4 – Largest Group

- From the t-SNE component plot, there appears to be a smaller group that is different.
- Highest Gamma response.

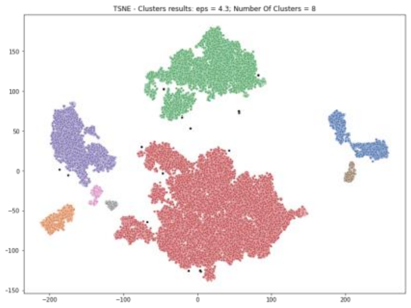
Parallel Coordinate Plots of 8 Clusters

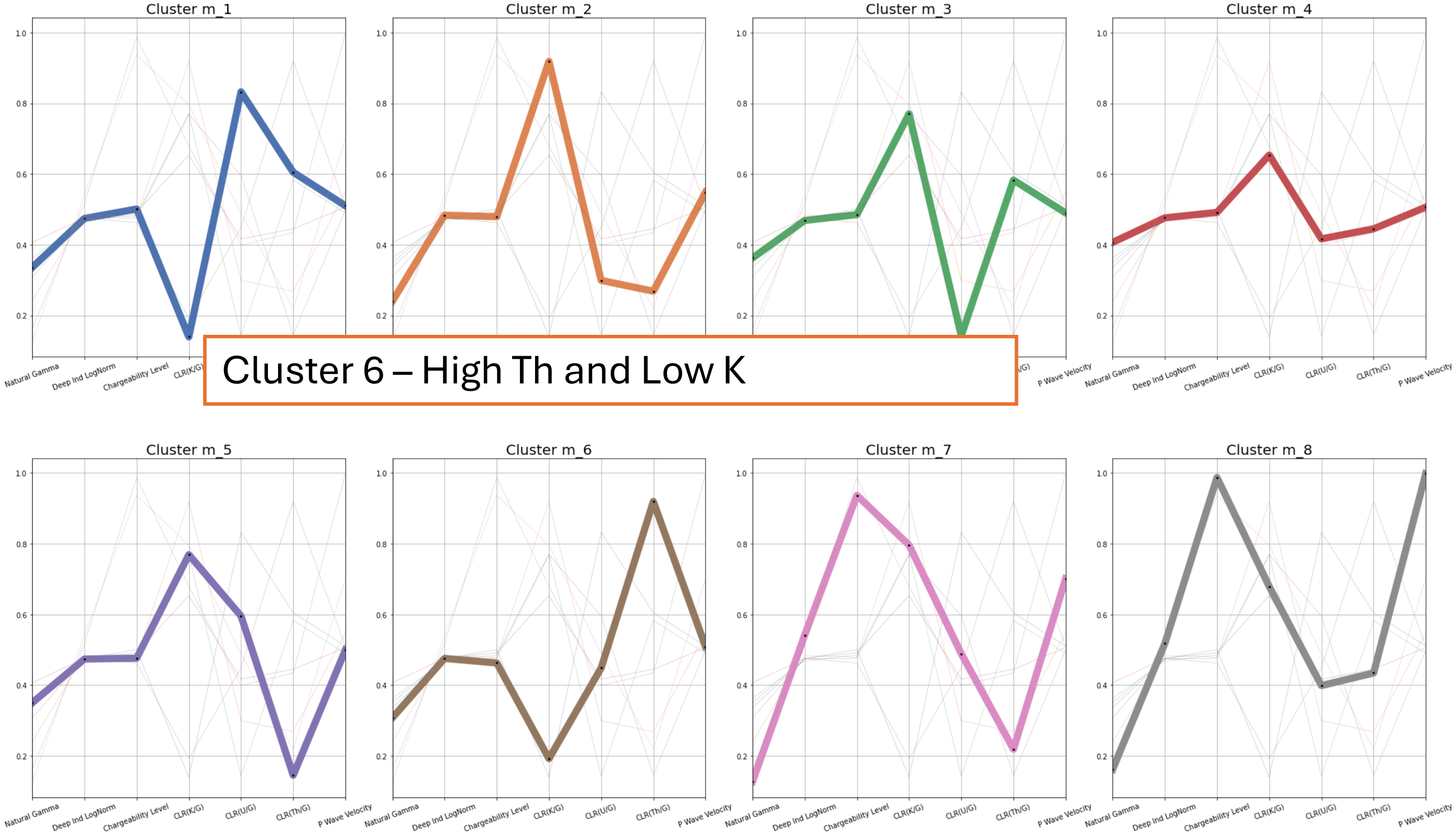




Cluster 5 – Lowest Th and P-Wave response.

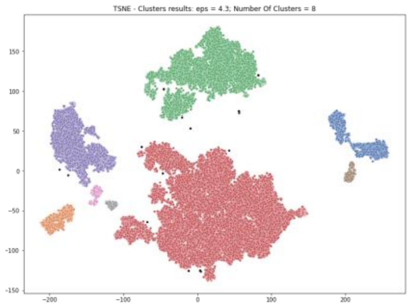
Parallel Coordinate Plots of 8 Clusters

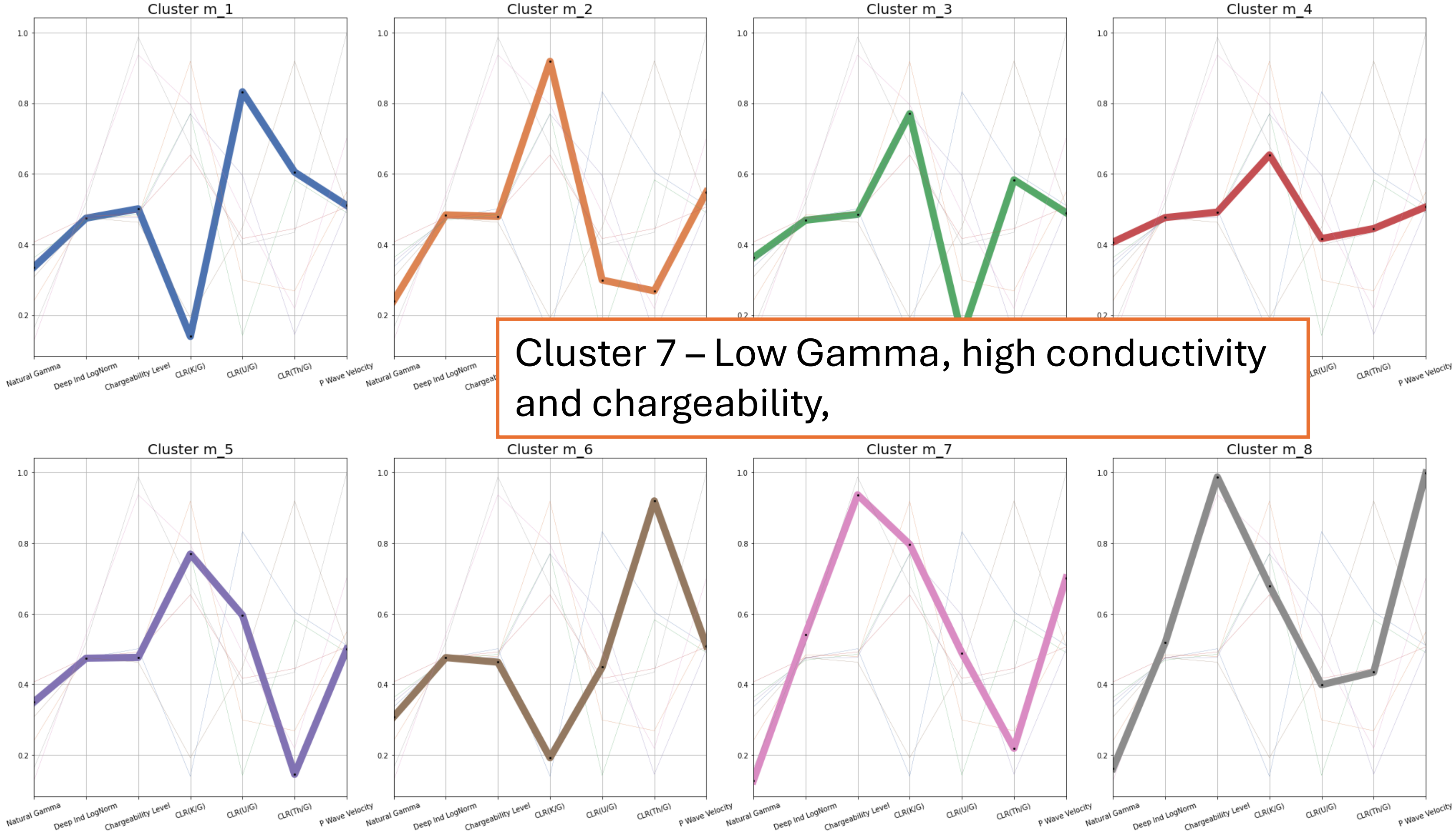




Cluster 6 – High Th and Low K

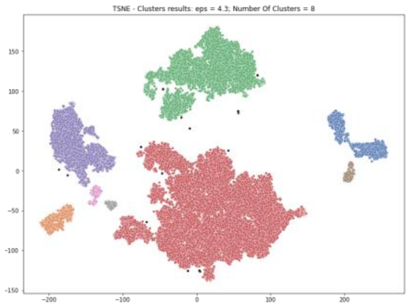
Parallel Coordinate Plots of 8 Clusters

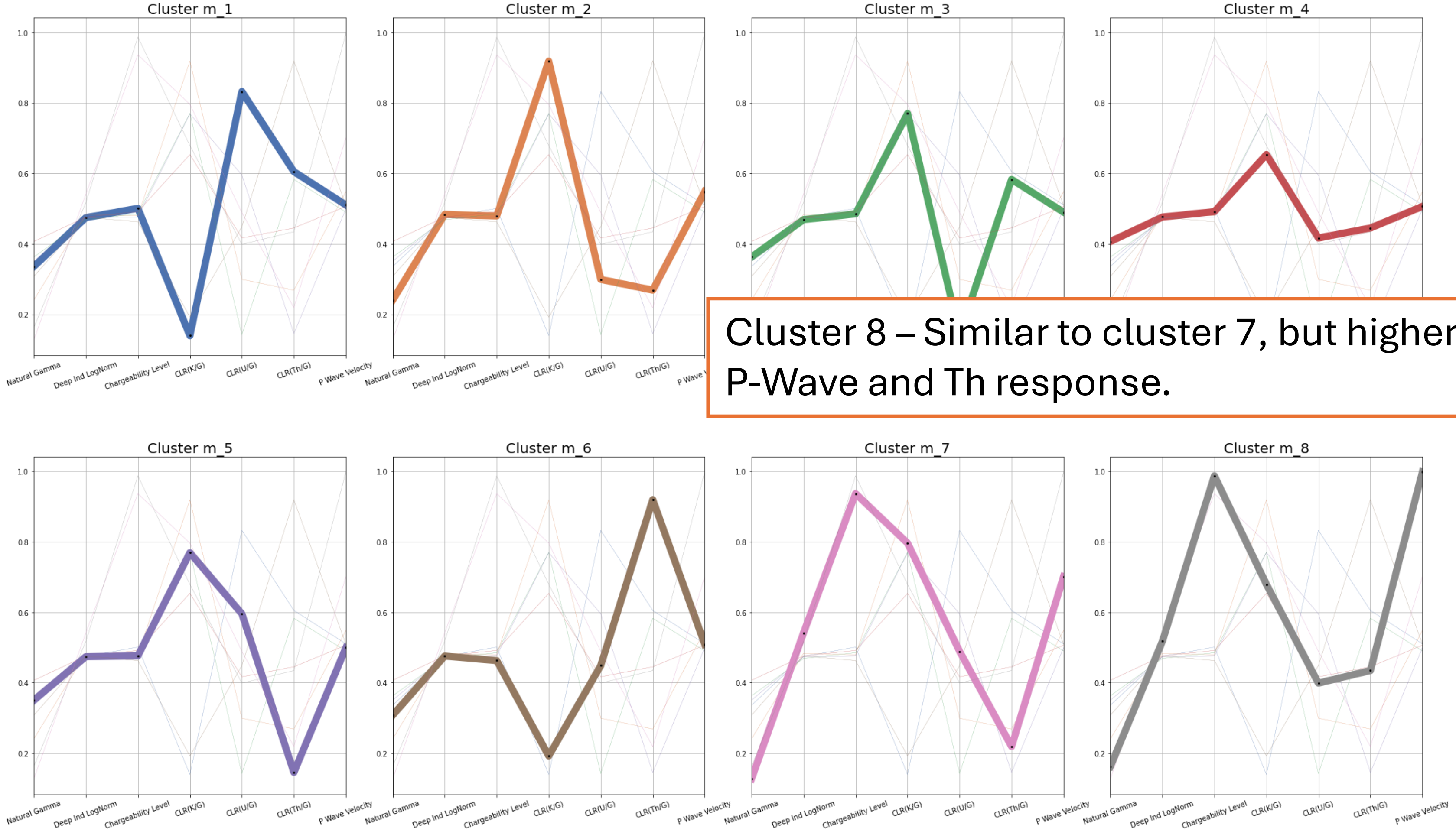




Cluster 7 – Low Gamma, high conductivity and chargeability,

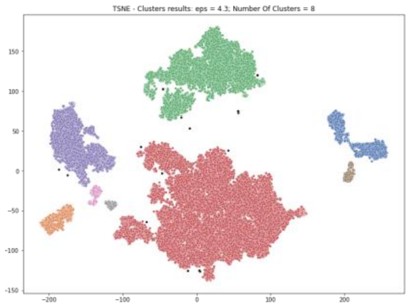
Parallel Coordinate Plots of 8 Clusters

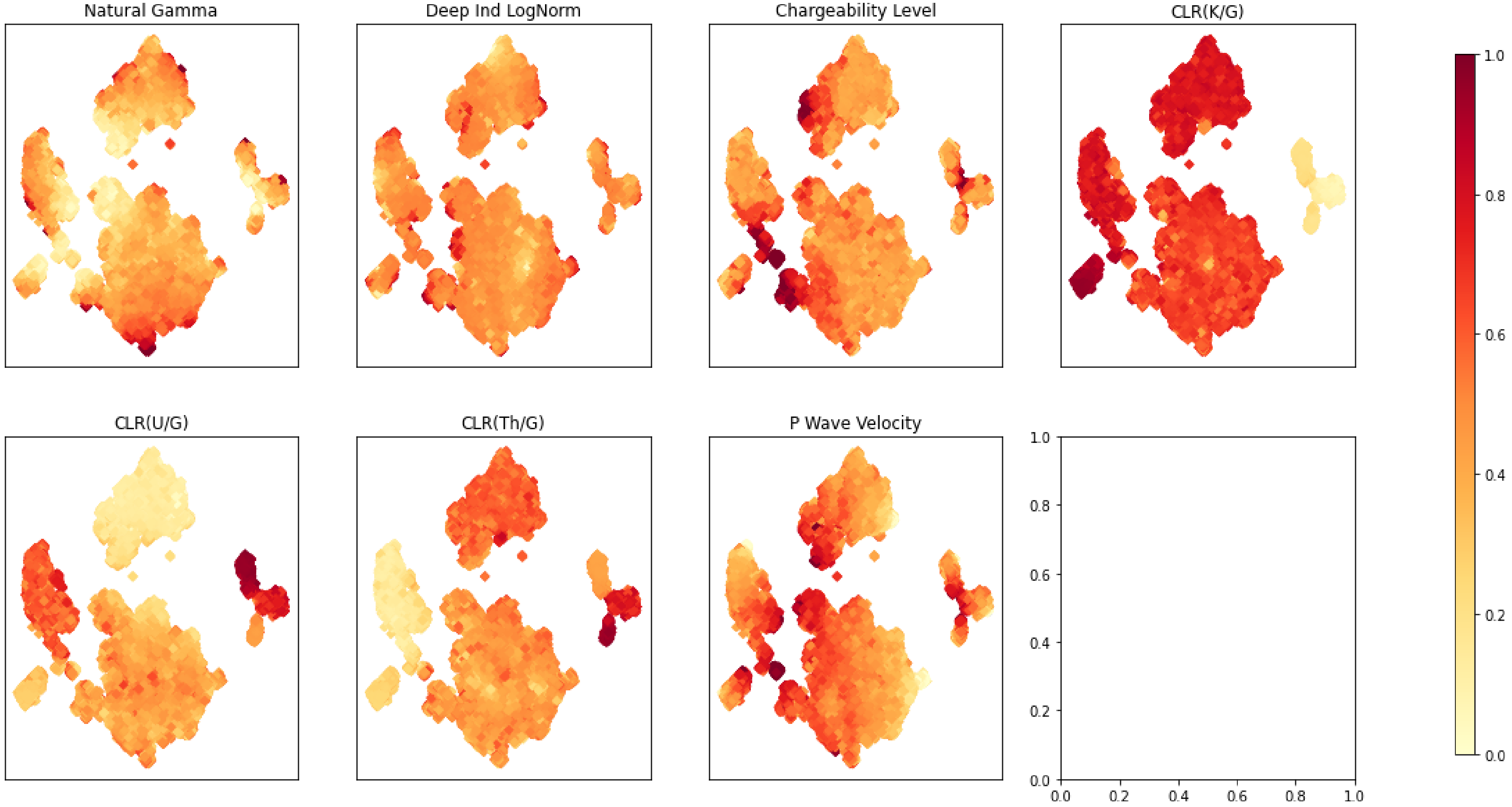




Cluster 8 – Similar to cluster 7, but higher P-Wave and Th response.

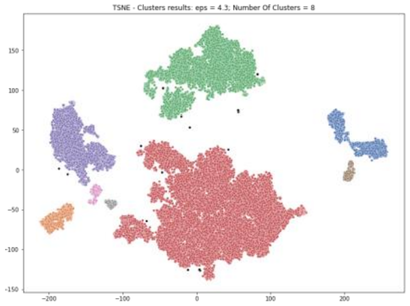
PRP Parameter Heat Map

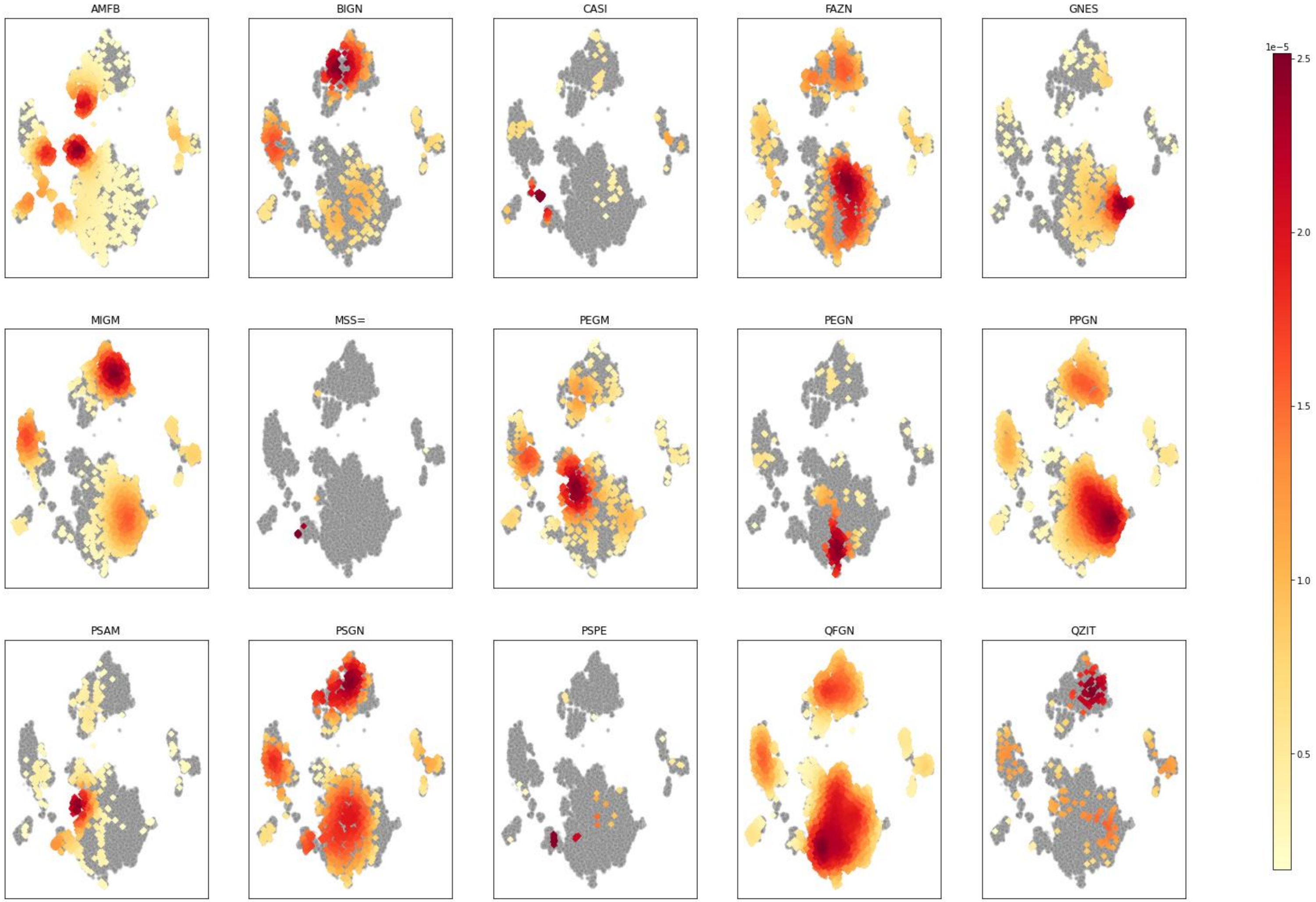




Lithology Heat Map

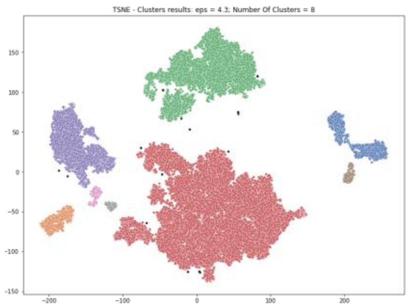
DGI
GEOSCIENCE

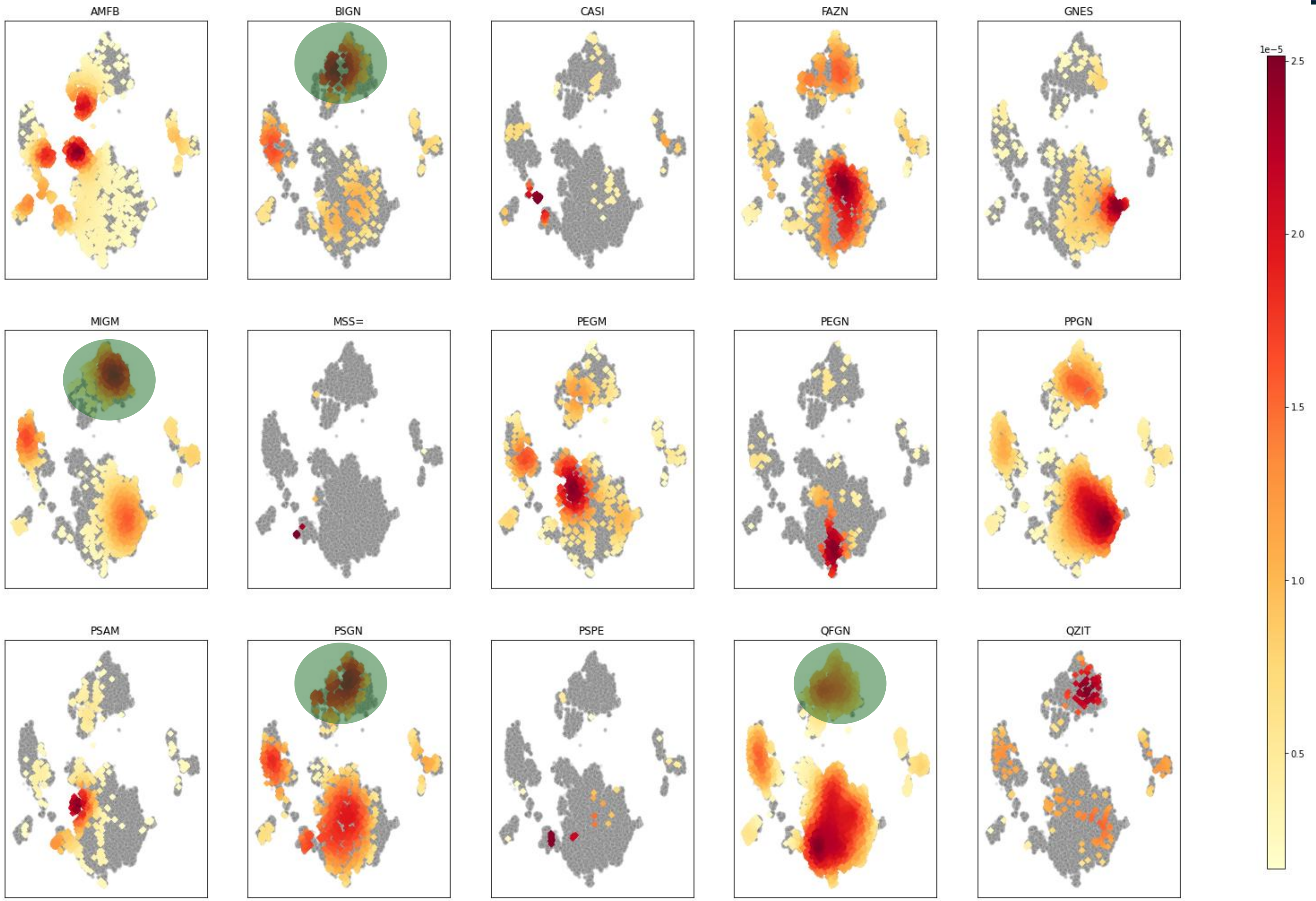




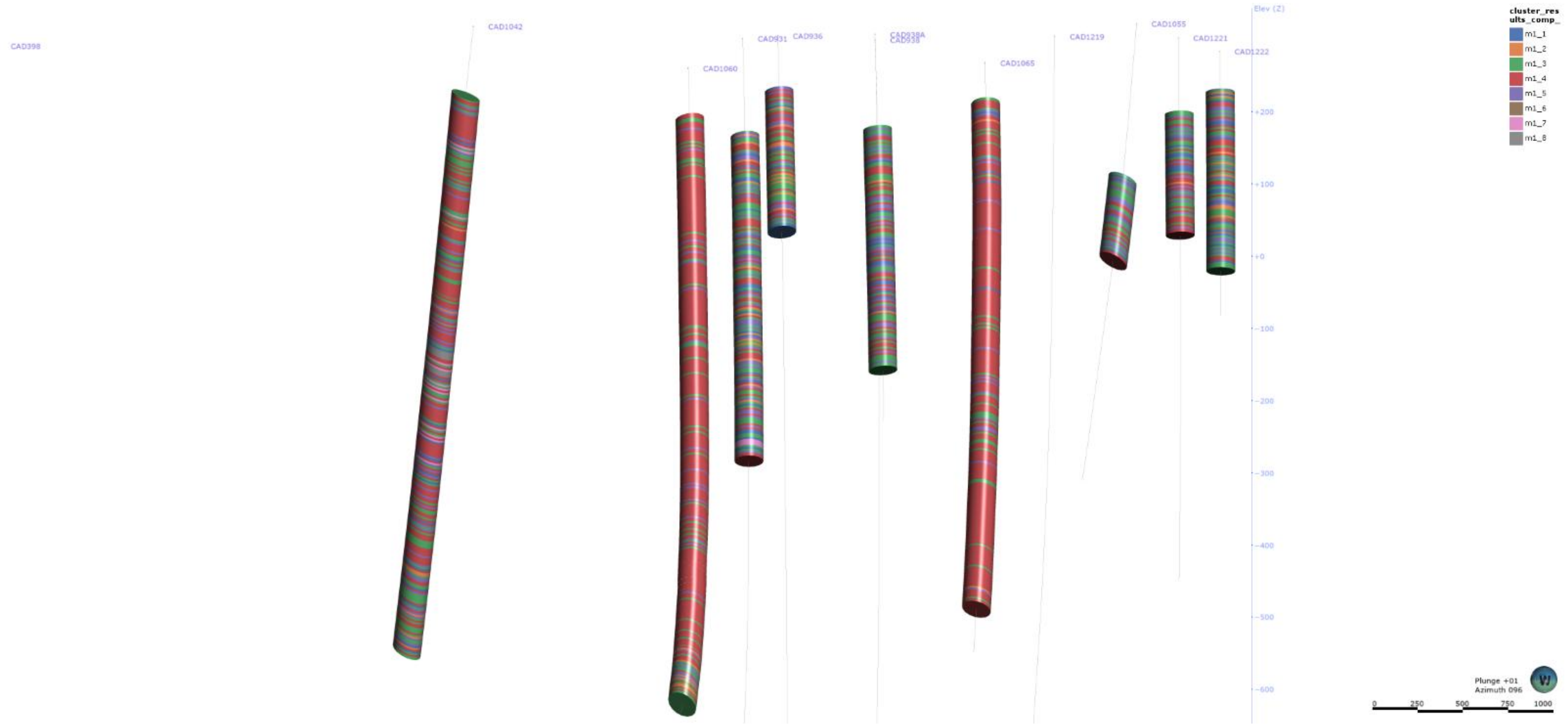
Lithology Heat Map

DGI GEOSCIENCE

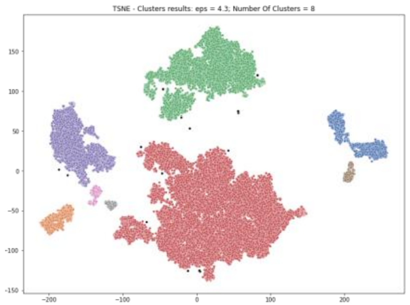


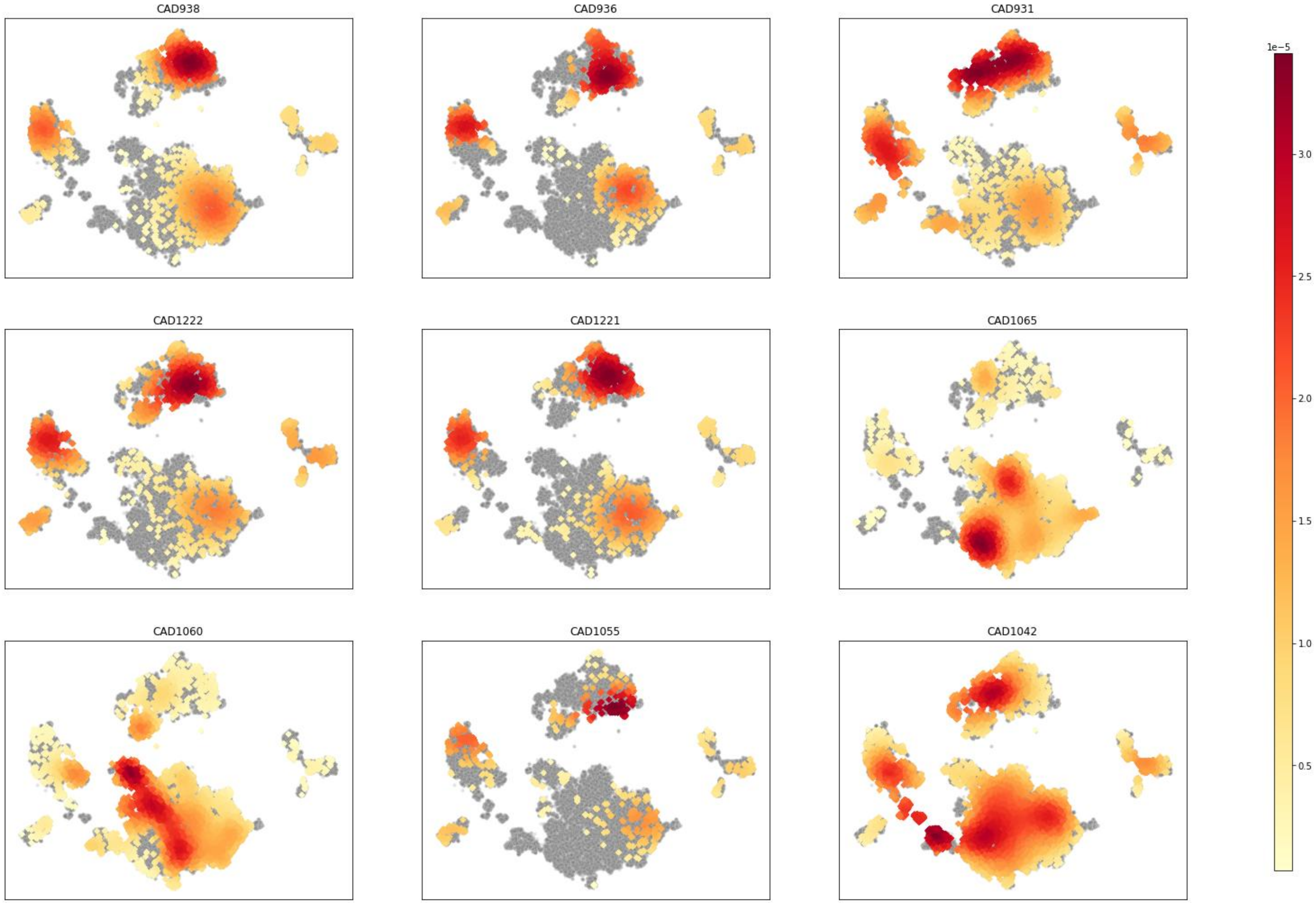


Spatial Distribution of Clusters



HOLEID Heat Map





Remarks

- PRP domains are an alternative interpretation based on the downhole logs.
- Will often match a mix of lithology, alteration, and/or mineralization.
- Same methodology can be applied to other datasets (geochemistry).
- It’s an iterative process.
  - Clusters can identify anomalies in the data.
    - How can we explain these anomalies?
      - Geology Related?
      - Borehole Conditions?
      - Probe Related?

DGI
GEOSCIENCE

Exercise
Python Analysis and Visualization

Click Here to Open