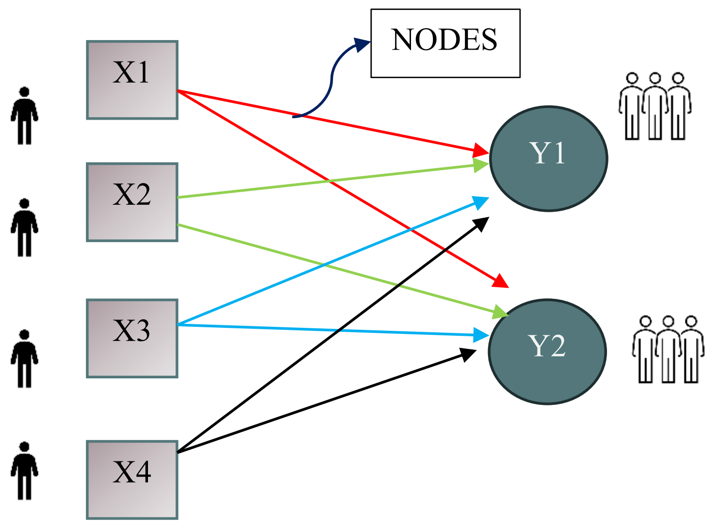
AIRBORNE MAGNETIC AND GRAVITY INTERPRETION USING ML

Course: Gravity and Magnetic methods

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

HUMAN ANALOGY FOR UNDERSTANDING SOM

BASIC REPERENSATION OF A SIMPLE SOM MAP:



Imagine attending a meeting with a group of random people (x1, x2, x3, x4). As you enter, you scan the entire group and try to recognize each person who seems most familiar to you (In SOM we use **Euclidean distance**). When you identify the most familiar person, known as the Best Matching Unit **(BMU),** you add them to your group. Gradually, you scan and remember everyone, creating a familiar group (y1,y2) around you. By the end of the meeting, you are surrounded by the most familiar people. This analogy mirrors how neurons in our brain help us recognize things around us. Similarly, in a Self-Organizing Map (SOM), **neurons (nodes)** find the closest node in the input space, mimicking this human brain process. The **sigma** parameter is like your ability to recognize people within a certain distance. It controls how far your recognition spreads, determining how many people around you can identify as familiar. On the other hand, the **learning rate** is like how quickly you adjust your understanding of who is familiar. It dictates how much you update your memory of familiar faces during the meeting. Together, these parameters help you form a group of familiar people around you, just like how neurons in a Self-Organizing Map (SOM) adjust and recognize patterns in data.

CLUSTERING USING SOM

Self-Organizing Map are a type of artificial neural network which mimics the way in which human brain organizes and process information. It was specifically designed for unsupervised learning. SOM basically reduces the high dimensional data into low dimension (2-D grid) similar like how our brain process big chunk of information into organized way. SOM excels at detecting hidden patterns, clusters, and anomalies — a core aspect of intelligent systems. Unlike traditional clustering methods, SOM preserves the topological structure of input data, making it effective for geophysical applications.

MATHEMATICAL FORMULATION OF SOM

Each neuron in the SOM grid is assigned a **weight vector** of the same dimension as the input data, chosen at random. For each input data point, the Euclidean distance between the weight vectors and the input data points is calculated. The input data point with the smallest distance to the weight vector is identified as the **Best Matching Unit (BMU).**

$$BMU = |x - w_i|$$

x= input data

w_i= *weight vector(i-th position)*

The weight vector for each input data (neuron) needs to be updated. The formula for updating the weight vector is provided below. *Source: https://medium.com/@amirali_62130* :

$$w_j(t+1) = w_j(t) + L(t) * (x - w_j(t))$$

$w_j(t+1) = \ New\ weight$

$w_i(t) = Old\ weight$

L(t)= Learning rate

x= input data

After processing all inputs, they are grouped into individual weight units (i), which is referred to as an **Epoch**. This process is repeated for multiple Epochs until the feature map stabilizes. The goal is to identify the nearest node to the input space based on the distance to the weight vector.

SOM PARAMETERS USED

1. **Sigma:** Specifies the width of the neighbourhood function, with higher values indicating a broader influence across neurons.
2. **Learning Rate (L):** Determines the magnitude of weight updates, which diminishes over time to ensure convergence.
3. **Max Iterations:** Denotes the total number of training steps executed during SOM training.
4. **Error Metric:** Typically employs quantization error to assess the distance between input vectors and their BMUs, thereby indicating the quality of convergence.

IMPACT OF PARAMETER TUNING ON SOM OUTPUT

Adjusting these parameters significantly affects the SOM output. For instance, a high sigma may result in overly smooth maps with less distinct clusters, while a low sigma may create fragmented clusters. Similarly, the learning rate and iteration count control the convergence speed and accuracy. Proper tuning is crucial for achieving optimal results.

High sigma results in smooth, generalized maps with less distinct clusters. Low sigma creates sharper, but possibly fragmented clusters. A high learning rate accelerates convergence but may overshoot optimal weights. A low rate leads to slower learning and possibly better accuracy. Too few iterations may result in underfitting. Too many can lead to overfitting or wasted computation. Monitoring error helps in understanding the SOM's learning progression and fine-tuning other parameters.

```
     quantization_error
0              0.054120
1              0.053532
2              0.050218
3              0.049740
4              0.054990
..                  ...
475            0.033493
476            0.036291
477            0.037570
478            0.033988
479            0.034822

[480 rows x 6 columns]
✅ Using Best Parameters: Grid=(5, 5), Sigma=1.0, LR=0.3, Iterations=10000
```

$$\text{QE}=\frac{1}{N}\sum_{i=1}^{N}\left\|x-w_j\right\|$$

Figure 11: *Quantization error(QE) over training iterations for the Self-Organizing Map (SOM). The error gradually decreases as training progresses, indicating improved mapping quality. The lowest error values occur near the final iterations, confirming convergence. The model used the best parameters: Grid=(5,5), Sigma=1.0, Learning Rate=0.3, and Iterations=10,000.*

WHY SOM OUTPERFORMS FCM AND DBSCAN

FCM uses probabilistic memberships, which can be affected by noise. SOM preserves spatial relationships between clusters, providing separation of geological anomalies. DBSCAN requires careful parameter tuning to handle regions with varying densities. SOM adapts to complex cluster structures without needing pre-defined density thresholds.

WHY USE SOM + K-MEANS?

The Self-Organizing Map (SOM) algorithm reduces the dimensionality of your dataset by mapping similar data points to neurons on a 2D grid. Each neuron has a weight vector that represents the learned feature pattern and every data point gets assigned to its BMU on this SOM grid. After training, each SOM neuron (grid cell) contains a weight vector that represents the feature space. Instead of clustering individual data points, we cluster these neuron weights because they represent the overall feature structure. We treat each neuron weight as a data point in a lower-dimensional space and apply K-Means to these neuron weights to group them into meaningful clusters. Each neuron is now labelled with a K-Means cluster ID. Every original data point is already assigned to a neuron (BMU), and instead of using BMU coordinates, we assign the data point the cluster ID of its BMU (from K-Means). This ensures that data points with similar features belong to one cluster instead of being scattered. SOM simplifies data and structures it meaningfully, while K-Means refines clustering using SOM's weight vectors. This hybrid method enhances interpretability and robustness, making it more effective for extracting clusters from the weight vectors themselves.

ARE WE FORCING CLUSTERS?

The Self-Organizing Map (SOM) naturally organizes similar data into neighborhoods, preserving the inherent structure without enforcing hard clusters. To enhance interpretability, K-Means clustering is applied **not to the raw input data**, but to the **weight vectors (w_j) of the trained SOM neurons**. This process is like coloring a map rather than drawing it from scratch. The SOM already provides a topologically ordered and dimensionally reduced representation of the input data, and K-Means simply groups these organized structures for better clarity. Therefore, we are not imposing clustering on the raw data; instead, we are quantifying and labeling the organization that the SOM has already established. This approach ensures that similar data points are grouped together in a meaningful way, enhancing the overall interpretability of the data structure.

$$\boxed{\textbf{Centroid=arg}\min_j \|\boldsymbol{x} - \boldsymbol{w}_j\||}$$

SOM Training and Optimal Parameter Selection Procedure

To ensure robust clustering performance using the Self-Organizing Map (SOM), a comprehensive hyperparameter optimization pipeline was implemented. This involved systematically evaluating multiple configurations based on clustering quality metrics and retraining the SOM using the best-found parameters for final clustering.

1. **Hyperparameter Grid Search**

A nested loop was used to iterate through combinations of key SOM hyperparameters:

- **Grid Size** (e.g., (5,5), (6,6), …)
- **Sigma (σ)**: Controls the neighbourhood radius.
- **Learning Rate (LR)**: Governs the rate of weight updates.
- **Iterations**: Determines the number of training epochs.

For each parameter combination, the SOM was trained, followed by K-Means clustering on the trained weight vectors. The clustering performance was evaluated using:

- **Silhouette Score**: To assess cluster cohesion and separation.
- **Quantization Error**: To measure topological accuracy and convergence quality.

2. **Best Configuration Selection**

The configuration yielding the **highest Silhouette Score** was selected as the optimal setup.

3. **Final SOM Training and Clustering**

After identifying the optimal configuration, the SOM was **retrained** using the best parameters obtained. Post-training, **K-Means clustering** was reapplied on the SOM neuron weight vectors. InThe final clustering result, visualized in the report, is therefore a direct outcome of this retrained SOM model using the most effective hyperparameters discovered during the tuning process.

CLUSTERING RESULTS

The clustered data is plotted, indicating geological structures. Self-Organizing Maps (SOM) categorize data into clusters representing various subsurface characteristics.

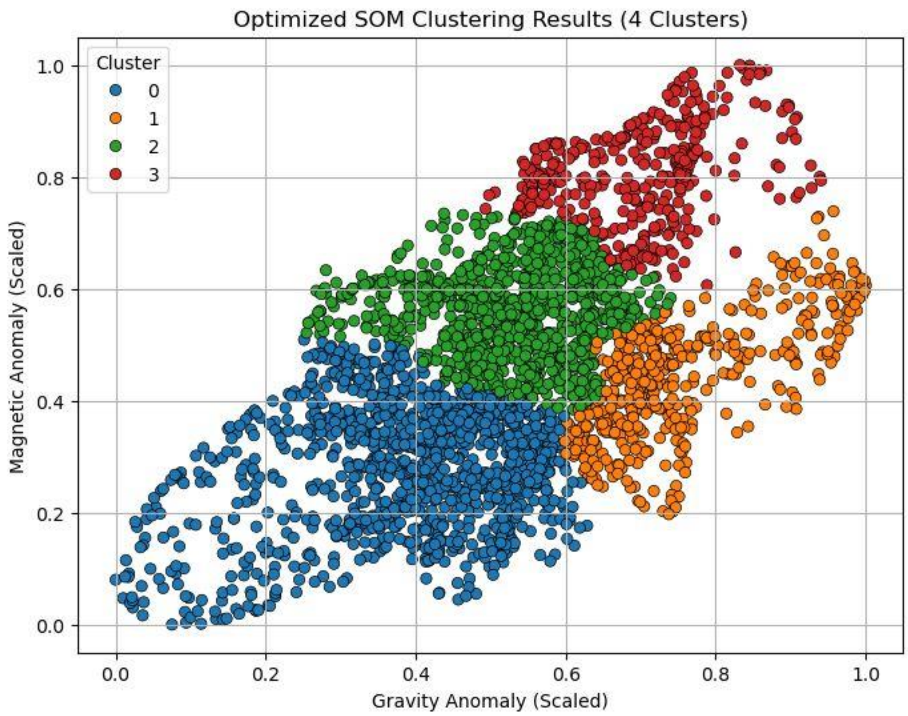


Figure: *Clustered representation of scaled gravity and magnetic anomalies using a Self-Organizing Map (SOM) and K-Means. Four distinct clusters were identified, highlighting zones of geophysical similarity. Data scaling using MinMaxScaler allowed improved feature representation and visualization of the underlying structure in the anomaly space.*

The clustering plot shows that SOM preserves topological relationships while grouping data based on gravity and magnetic anomalies. Each color-coded cluster represents unique geophysical properties, which may indicate different subsurface lithologies or structural domains. **MinMaxScaler** standardizes input features for balanced weightage during clustering, creating sharp cluster boundaries that support strong separability and enhance geological interpretability.

STATISTICAL ANALYSIS OF CLUSTERS

Means and standard deviations for Bouguer Gravity and Magnetic Anomalies are calculated. A statistical plot shows cluster-wise variations.

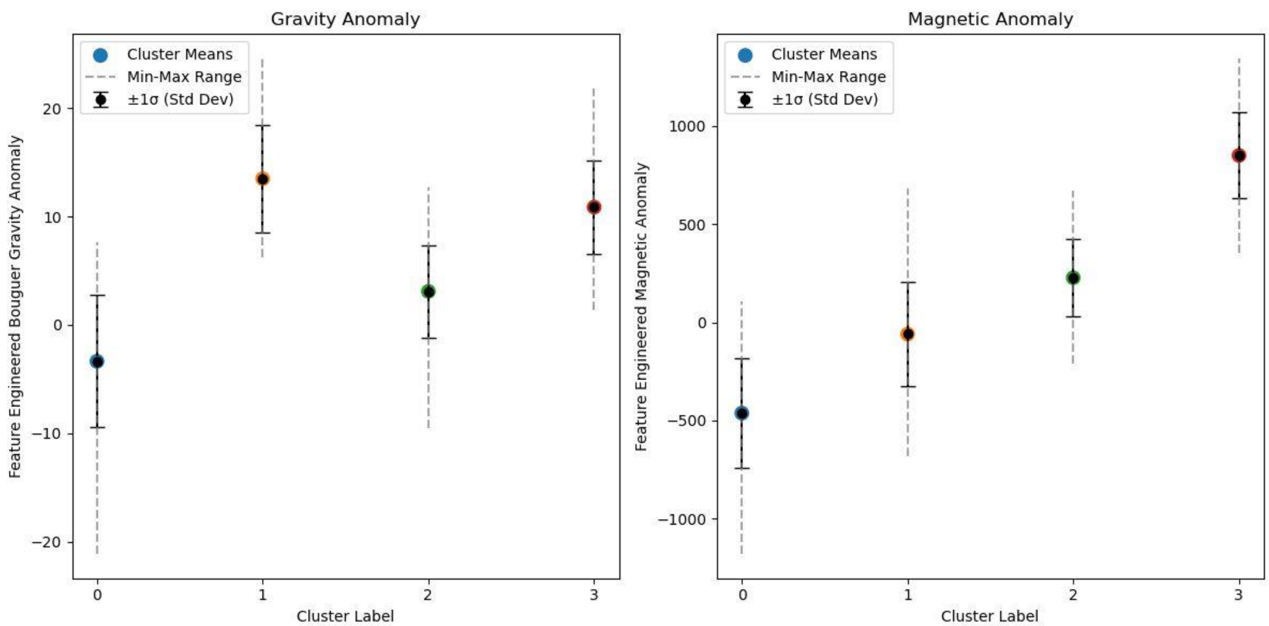


Figure 12: *Cluster-wise statistical distribution of Bouguer gravity and magnetic anomalies. Error bars indicate ±1 standard deviation, while vertical dashed lines show the full min–max range of values in each cluster. Clear differences in means and spread across clusters demonstrate the ability of the SOM + K-Means method to identify geologically meaningful patterns in airborne geophysical data.*

The statistical distribution of clusters based on gravity and magnetic anomalies reveals meaningful separations in the data. Clusters with tighter standard deviations represent geophysical coherent zones, while wider spreads may point to heterogeneous or transitional regions. The combination of SOM for feature mapping and K-Means for clustering allows for a balance between topological preservation and statistical robustness. This facilitates clearer interpretation of subsurface variability, which can inform geological modelling, anomaly classification, or even exploration targeting.

U-MATRIX REPRESENTATION

The Unified Distance Matrix (U-Matrix) is used in Self-Organizing Maps (SOM) to visualize the distances between neighboring neurons in the grid. It indicates boundaries between clusters by showing areas of high dissimilarity. High values in the U-Matrix represent significant differences between adjacent weight vectors, suggesting cluster separation.

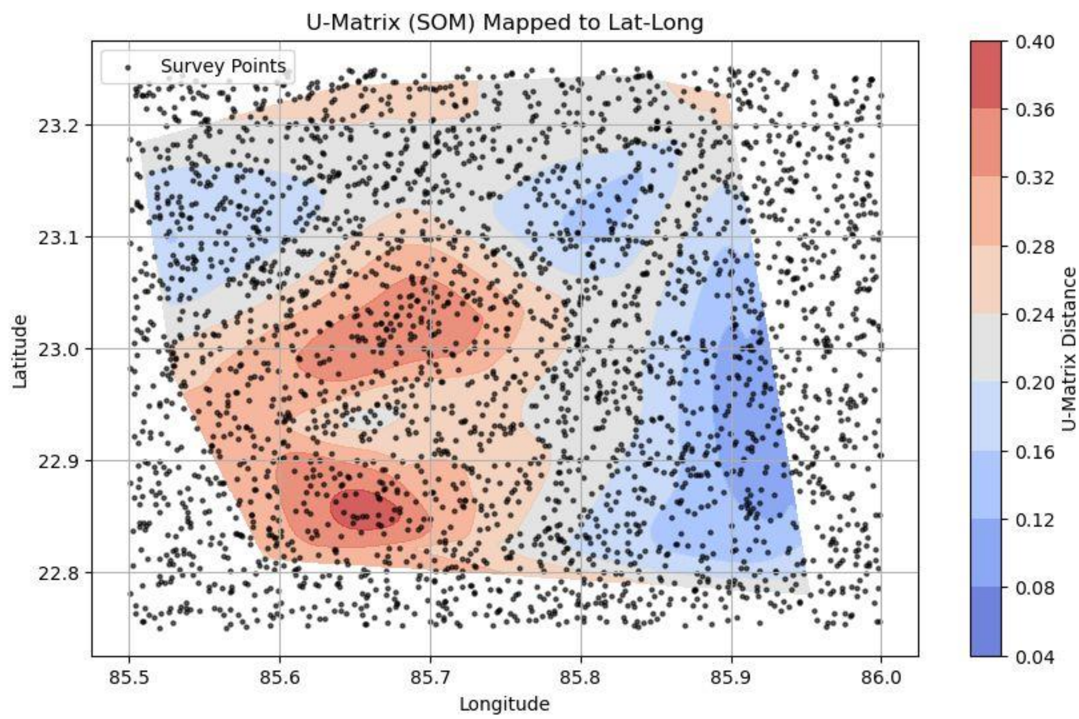


Figure 13: *U-Matrix representation of SOM neurons mapped across the study area. High-distance zones (darker regions) indicate boundaries between clusters, revealing potential fault zones or lithological transitions across the subsurface.*

When integrated with geospatial coordinates such as latitude and longitude, the U-Matrix becomes an essential tool for mapping the geographical spread of subsurface anomalies. This representation not only highlights spatial variations in magnetic or gravity anomalies but also aids geoscientists in correlating these patterns with actual geological features such as lithological boundaries, mineralized zones, or tectonic structures. Thus, it bridges the gap between unsupervised clustering and real-world geological interpretation.

SPATIAL MAPPING OF CLUSTERS

The spatial cluster map visualizes variations in subsurface geophysical patterns geographically. Each cluster is color-coded, representing distinct geophysical characteristics identified by SOM and grouped using K-Means. Dense clusters suggest clear geological boundaries, such as different lithological units, mineralized zones, or fault structures. This map helps geoscientists identify regions for further exploration or detailed surveying, combining coordinates with geophysical anomaly clustering for georeferenced interpretation of subsurface structures.

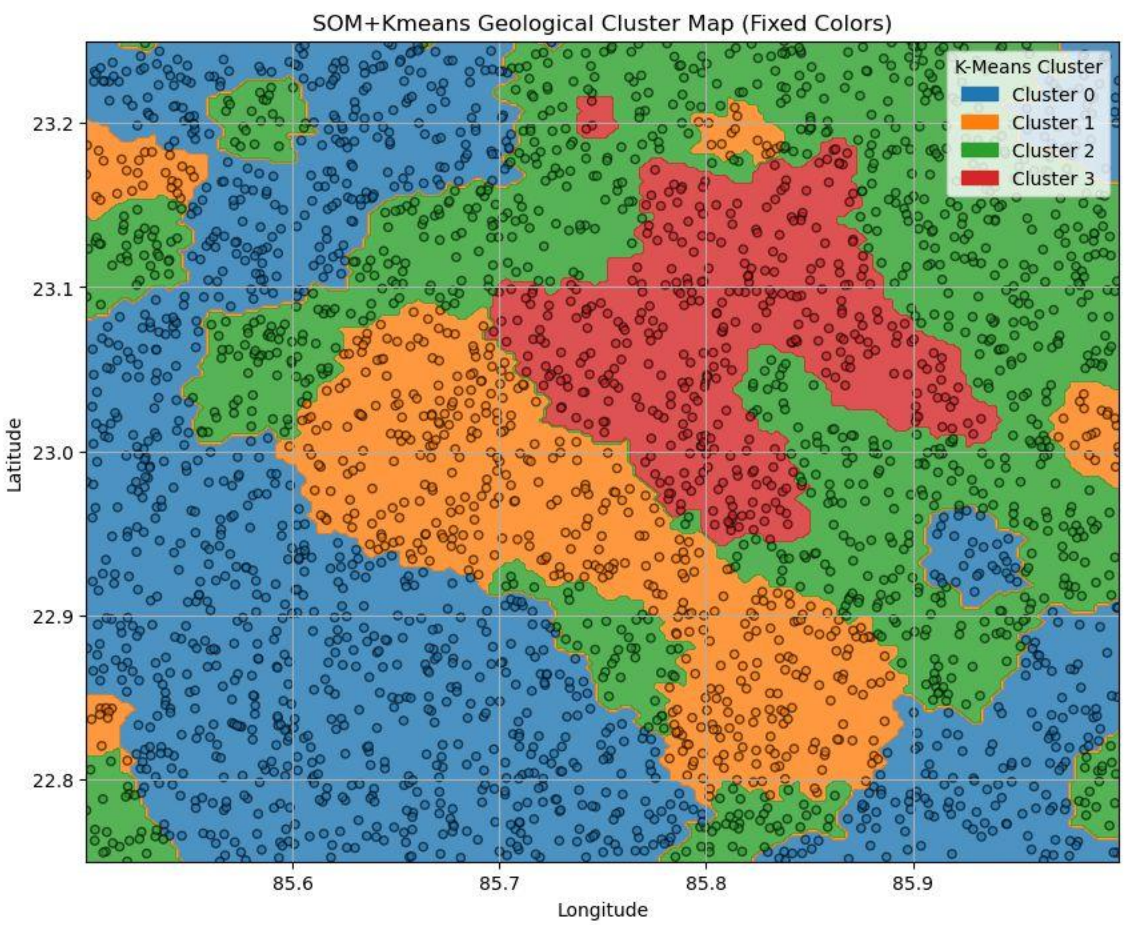


Figure 14: *Geospatial distribution of SOM + K-Means clustered data plotted over latitude and longitude coordinates. Each color-coded region reflects a distinct geophysical signature, providing geological context for anomaly interpretation and potential target zone identification.*

REFERNECES

- Kumar, S., Arasada, R. C., & Rao, G. S. (Year). Multi-Scale Potential Field Data Integration Using Fuzzy C-Means Clustering for Automated Geological Mapping of North Singhbhum Mobile Belt, Eastern Indian Cratons. *Minerals* ***2023***, 13,1-15. DOI/Publisher.
- *https://medium.com/@amirali_62130*
- https://sklearn-som.readthedocs.io/en/latest/#sklearn-som-v-1-1-0
- https://medium.com/machine-learning-researcher/self-organizing-map-som-c296561e2117
- DVIJAYAKUMAR, GSRINIVSA Rao et al. (2024), Cluster analysis of geophysical data for geology differentiation in Agnigundala mineralized belt of Cuddapah Basin, India.
- J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” Computers & Geosciences, vol. 10, no. 2–3, pp. 191–203, 1984.
- Jing Gao & SUNY Buffalo, online Lecture, Clustering Lecture 4: Density-based Methods.
- Google Chrome Browser & AI assistance.