PE 5970: Data Mining for Petroleum Engineers: 113102152

Due Monday, 28th March Jiwon Jeon

I have provided you with about 25 different well logs. Chose 10 of these and use the well logs to identify electrofacies. You can use as many logs as you like, however PE, RHOB, NPHI, GR and Deep Resistivity are a must. You can import the logs as .LAS files to Techlog, Petra, IP or Petrel to export the logs in a log versus depth format.

Your submissions should be your own work. However you may discuss the methodology with your classmates or anyone else. You can use any one of the clustering techniques discussed in class or any other unsupervised classification technique of your choice.

Your submission should include a list of the wells you have chosen, a brief description of your unsupervised classification technique and a discussion of your results. Note that your results will be more meaningful if your wells are physically located close to one another.

1. Selection of Wells

The given well logs were recorded from the eight counties in the Barnett Shale region as highlighted in **Figure 1**.

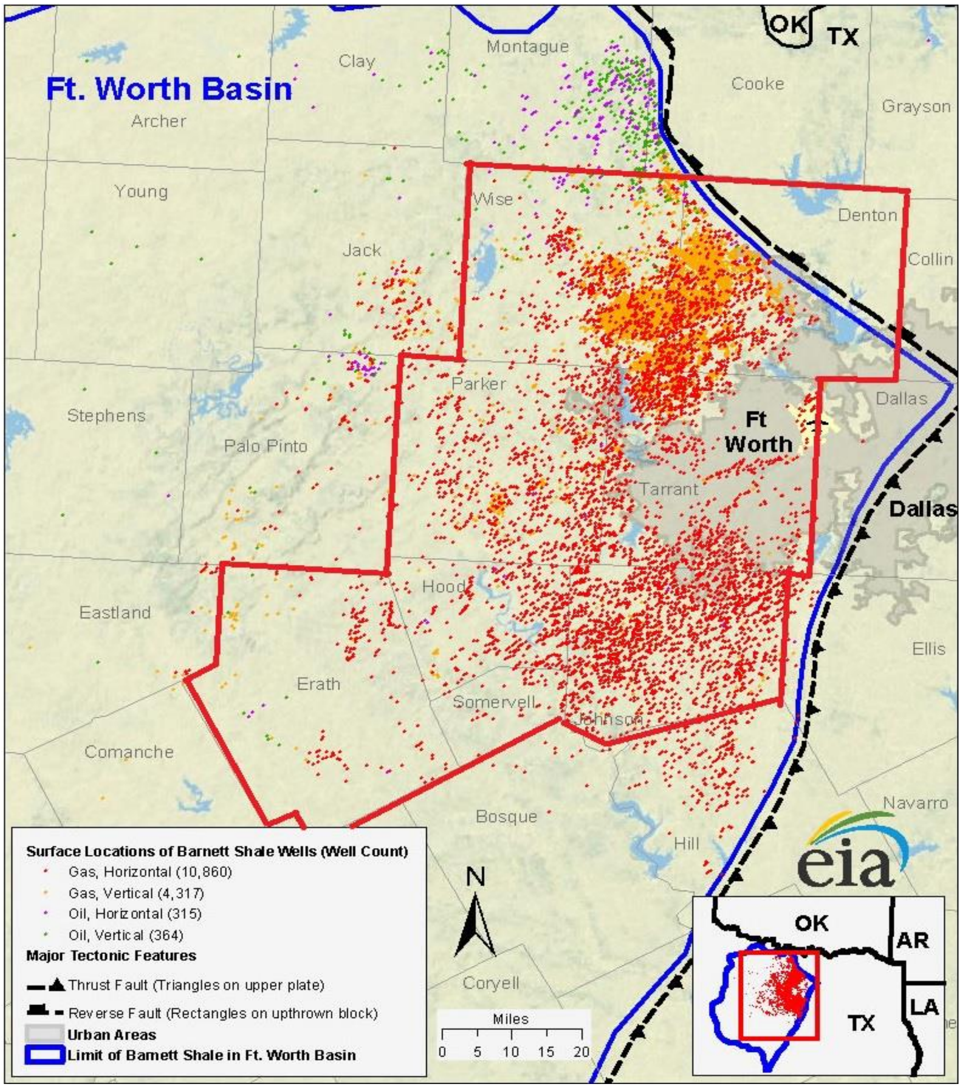

Figure 1. Barnett Shale Wells (HDPI, USGS, Pollastro et al (2007))

Out of these eight counties, the following counties are classified as "Core" area where the shale is thicker and production is concentrated;

- Denton
- Wise
- Tarrant
- Johnson

The other four counties, Parker, Hood, Somervell, and Erath are considered as "Non-Core" areas, however, they are also active in production.

Well logs from the "Core" areas were primarily chosen and investigated using IP to export PE, RHOB, NPHI, GR and Deep Resistivity in a log versus depth format. 9 wells fell into consideration, the wells from Parker County were therefore explored to ensure enough dataset of 10 wells; total 16 wells were selected.

In order to establish a reliable dataset from the upper and lower Barnett Shale region, logs from the 16 wells were examined and cleaned up with the limitations considering general properties of shale formation;

Table 1. Log Limitations for Well Selection

Log	Unit	Limitation
PE	[B/E] or [dimensionless]	2.0 ~ 5.0
RHOB	[G/C3]	2.2 ~ 2.8
NPHI	[V/V]	≤ 0.3
GR	[GAPI]	100 ~ 200
AT90	[OHMM]	≤ 30

The following 10 wells in **Table 2** were finally selected for classification;

Table 2. List of Selected Wells

Well No.	Original Well No.	County
1	421213165800	Denton
2	421213207700	Denton
3	422513036300	Johnson
4	423673405000	Parker
5	42367340940009	Parker
6	42367343850009	Parker
7	42367344380009	Parker
8	42367348830009	Parker
9	42497365690009	Wise
10	42497368240009	Wise

2. Unsupervised Classification Technique – K-Means Clustering

The electrofacies of the chosen 10 wells were identified by K-Means Clustring technique.

a. K-Means Clustering

In K-Means Clustering, we partition the observed data into a pre-determined number of clusters. The clusters satisfy the following properties;

1. Each data point in the observation belongs to at least one of the clusters.
2. The clusters are not overlapping: no data point belongs to more than one cluster.

Based on these properties, K-Means Clustering aims to minimize the *within-cluster variation* as small as possible. The *within-cluster variation* means how the data points within a cluster differ from each other, and can be mathematically expressed by *within-cluster-sum-of-squares*; sum of distance of each point to the centroid of the cluster where the point is included.

$$Min \sum_{i=1}^{k} \sum_{x \epsilon S_i} \|x - \mu_i\|^2$$

where S_i = i-th cluster

μ_i = centroid (mean) of the points in S_i

The algorithm to solve this equation is an iterative refinement process;

1. Determine the number of clusters and compute the cluster centroid.
2. Randomly assign observed data points to the cluster whose centroid provides the least *within-cluster-sum-of-squares.*

   $$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \le j \le k\}$$

   where m_i, m_j = mean(centroid) of each cluster

   x_p is assigned to only one $S_i^{(t)}$
3. Calculate the new means to be the centroids of the data points in the new clusters.

   $$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$
4. Iterate the process until the assignment no longer changes.

b. Optimal Number of Clusters

The number of clusters are specified by determining Sum-of-Squares Between (SSB) clusters and Sum-of-Squares Within (SSW) all clusters. SSB quantifies the total difference between all clusters while SSW quantifies the differences between data point and the cluster centroid over all clusters. The following approach provides a way to find the optimal number of clusters;

1. Randomly choose number of clusters and compute SSB and SSW;
   For ‘g’ groups,

$$SSB = \sum_{l=1}^{g} n_l (\overline{x}_l - \overline{x})' (\overline{x}_l - \overline{x})$$

$$SSW = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (x_{lj} - \overline{x}_l)' (x_{lj} - \overline{x}_l)$$

2. Repeat 1 for a different number of clusters
3. Plot SSB and SSW vs number of clusters and find the number where SSB and SSW significantly changes (elbow of the plot).

3. Identification of Electrofacies

By plotting SSW and SSB, the optimal number of clusters was determined as 4.

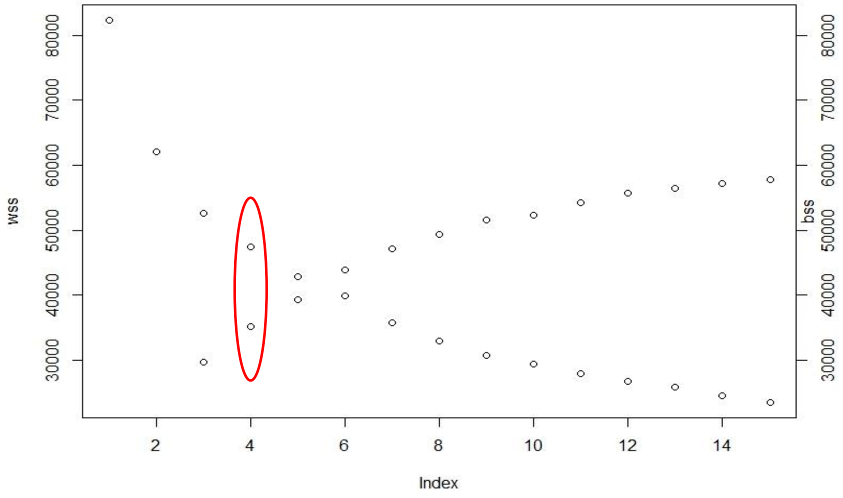

Figure 2. Optimal Number of Clusters

K-Means Clustering resulted in 4 clusters (electrofacies) having different data sizes, and provided the following centroid properties of each cluster;

Table 3. Cluster size and Centroid properties

Cluster	Data size	PEF	RHOB	NPHI	GR	AT90
1	4174	3.307703	2.642775	0.2210295	142.7806	8.480533
2	4783	3.274589	2.613128	0.2154464	114.5304	8.668973
3	2018	3.139967	2.623723	0.1700515	112.9424	16.826018
4	5484	3.587968	2.590638	0.2681393	124.4839	6.221092

The electrofacies along the depth of each well are classified in the attached excel file (Midterm Project_electrofacies_output.xlsx).

In order to visualize the clusters (electrofacies) with the logs used in classification, Principle Component Analysis (PCA) technique was used. PCA in R showed the following contribution of each log to Principle Component 1, 2, and 3;

Table 4. Principle Components of 3D PCA

	PC1	PC2	PC3
PEF	-0.4270553	0.19410176	-0.62687922
RHOB	0.2717705	-0.62041214	-0.64768120
NPHI	-0.6304898	0.06871688	0.05857840
GR	-0.2201486	-0.71340728	0.42325293
AT90	0.5456939	0.25247044	0.07040618

The 3-Dimensional plot of clusters with axis of PC1, PC2, and PC3 is shown as below;

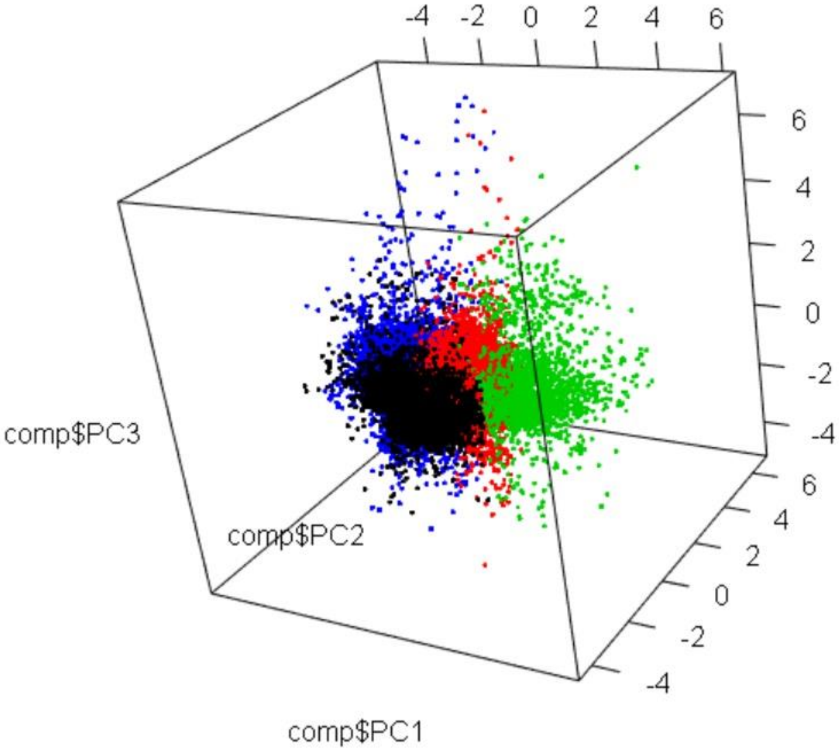

Figure 3. 3D plot of Clusters