PE 5970: Data Mining for Petroleum Engineers: **113102152**

Due Friday, 25th Match **Jiwon Jeon**

Part 1

1. Use the provided Excel spreadsheet that describes a core sample from the Barnett shale.
2. Use a self-organizing map approach to identify petrofacies from the provided dataset. Can this approach be used in a predictive mode – or in other words, can I hold back some data, train my SOM and then characterize the test data?

Your submissions should be your own work. However you may discuss the methodology with your classmates or anyone else. Be judicious in your choice of variables to use. Using all the variables that I provided you with is not the best use of your data.

1. **Selection of Core data**

From the given Barnett shale data, important variables for Self-Organizing Map are selected using a Principal Component Analysis (PCA), which provides a covariance matrix of eigenvectors for each variable and principal component.

Table 1 shows the results of PCA displaying the 1st principal component (PC1) to the 3rd principal component (PC3).

Table 1. PCA results

	PC1	PC2	PC3
Depth	0.951789	-0.29874	-0.04018
Corr_por	-0.00071	0.011772	0.023566
TOC	0.006513	0.010974	0.032161
Quartz	0.014241	0.137374	0.228076
Calcite	-0.0784	-0.19029	-0.48978
Dolomite	0.008984	-0.00708	-0.09545
Illite	0.044373	0.039157	0.209631
Smectite	0.001377	-0.0002	0.000572
Kaolinite	-0.00182	-0.00077	0.005264
Chlorite	-0.01037	-0.00857	-0.02328
Mixed_Clays	0.025821	0.018941	0.171621
Albite	0.005478	0.017168	0.036531
Siderite	-0.01538	-0.01683	-0.0287
Apatite	0.000345	0.023362	-0.00104
Aragonite	0.000101	0.000491	0.002023
Clays	0.059384	0.048738	0.363812
Carbonate	-0.08469	-0.21374	-0.61179
GR	0.271555	0.895347	-0.33535
RHOB	-4.57E-05	-0.0006	-0.0006
RHOBE_Norm	-5.12E-05	-0.0006	-0.0006
TNPH	0.012979	0.050657	0.023258

The result can also be described by biplot which shows the importance of the components in length from the center. A sample biplot is drawn in **Figure 1**.

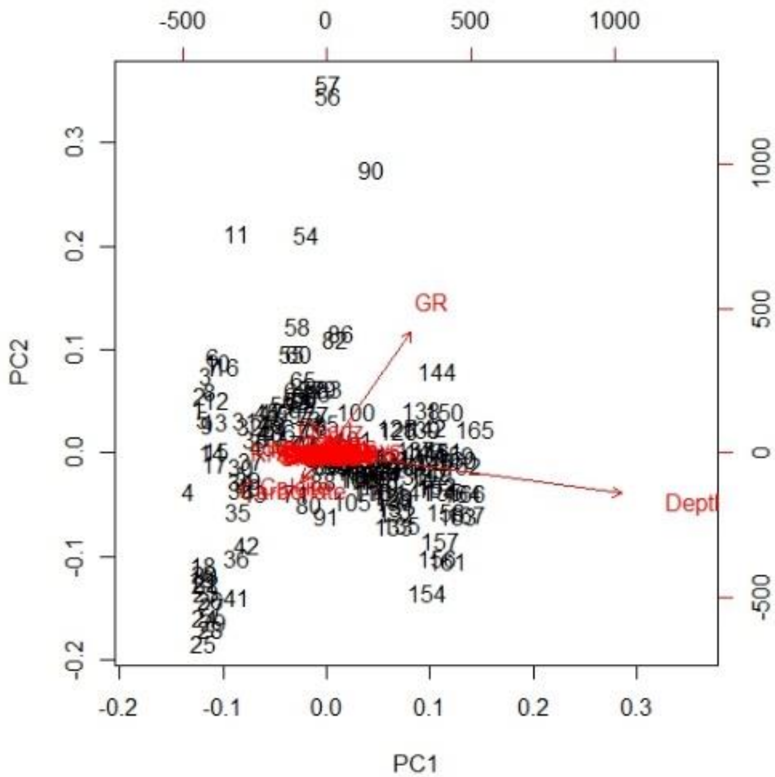

Figure 1. PCA biplot

Considering PC 1 to PC 3, the following variables are selected for SOM analysis;

Table 2. Selected variables

	PC1	PC2	PC3
Corr_por	-0.00071	0.011772	0.023566
TOC	0.006513	0.010974	0.032161
Quartz	0.014241	0.137374	0.228076
Calcite	-0.0784	-0.19029	-0.48978
Illite	0.044373	0.039157	0.209631
Mixed_Clays	0.025821	0.018941	0.171621
Clays	0.059384	0.048738	0.363812
Carbonate	-0.08469	-0.21374	-0.61179

Although Depth and Gamma Ray have high eigenvalues which mean a high impact, they are eliminated from the analysis in order to focus on petrophysical properties. Likewise, Corrected porosity and TOC are included in spite of low eigenvalues.

2. Identification of Petrofacies

For the purpose of predictive modes, 125 data points out of 167 total Barnett shale data points are sorted and scaled to be analyzed in SOM.

a. Number of Clusters

8 x 8 SOM model is chosen by confirming little number of grey grids in "counts" plot. This indicates that most data points of the selected are utilized to determine the grid. 500 iterations are designed since "Training progress" is stabilized around 300 iterations and lasts until 500.

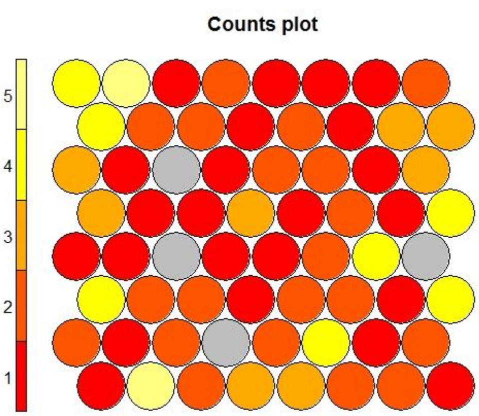

Figure 2. Counts plot

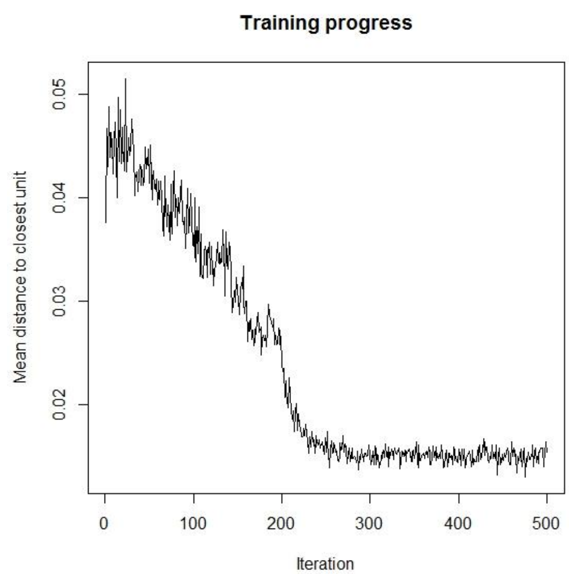

Figure 3. Training progress

To find the optimum number of clusters, the eight(8) variables are plotted to compare the property distribution in SOM model. SOM model scales all data in order to remove the unit impact of each property. Thus, the data will be analyzed only in dimensionless scale. **Figure 4** shows the result implying that the first and third properties have similar distribution, while second and fifth, fourth and eighth, and sixth and seventh properties are similar, respectively. Therefore, the number of clusters can be determined as four(4).

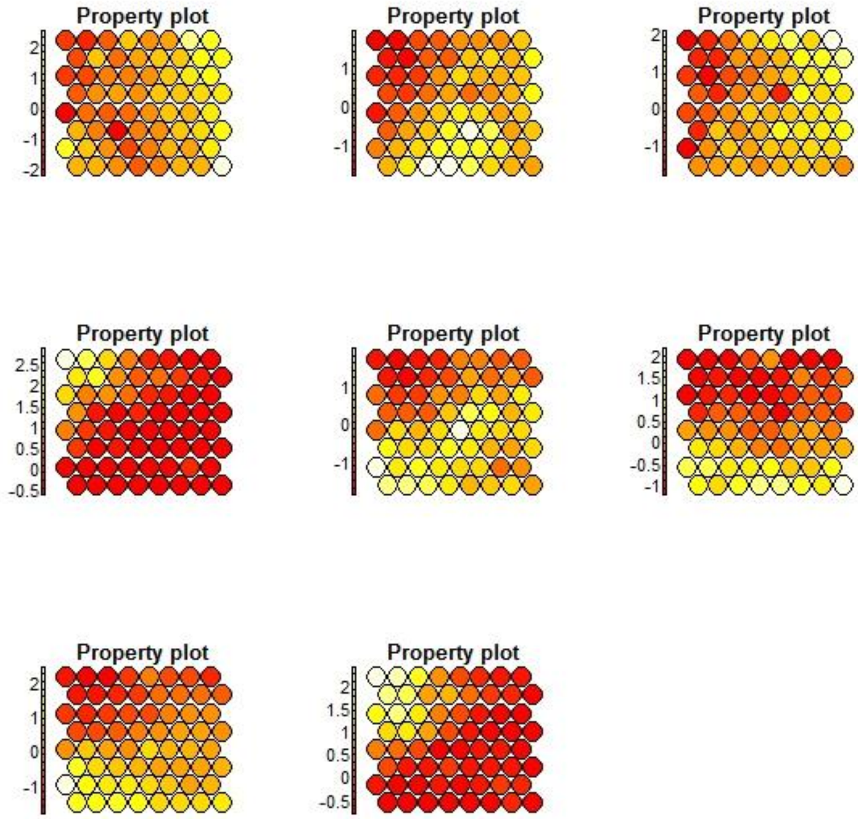

Figure 4. Distribution of Properties

The number of clusters can also be determined by comparing Sum-of-Squares Between (SSB) clusters and Sum-of-Squares Within (SSW) all clusters as shown in **Figure 5**.

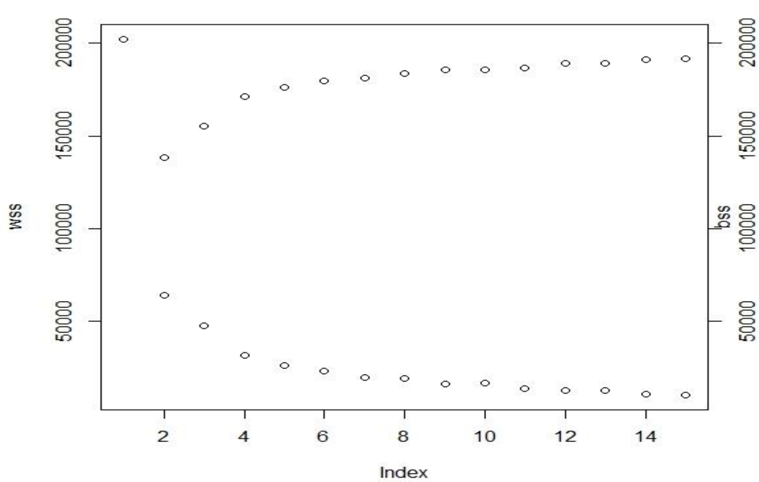

Figure 5. SSB and SSW

b. Determination of Petrofacies

By running SOM model, the following four(4) clusters are mapped with eight(8) variables (petrophysical properties);

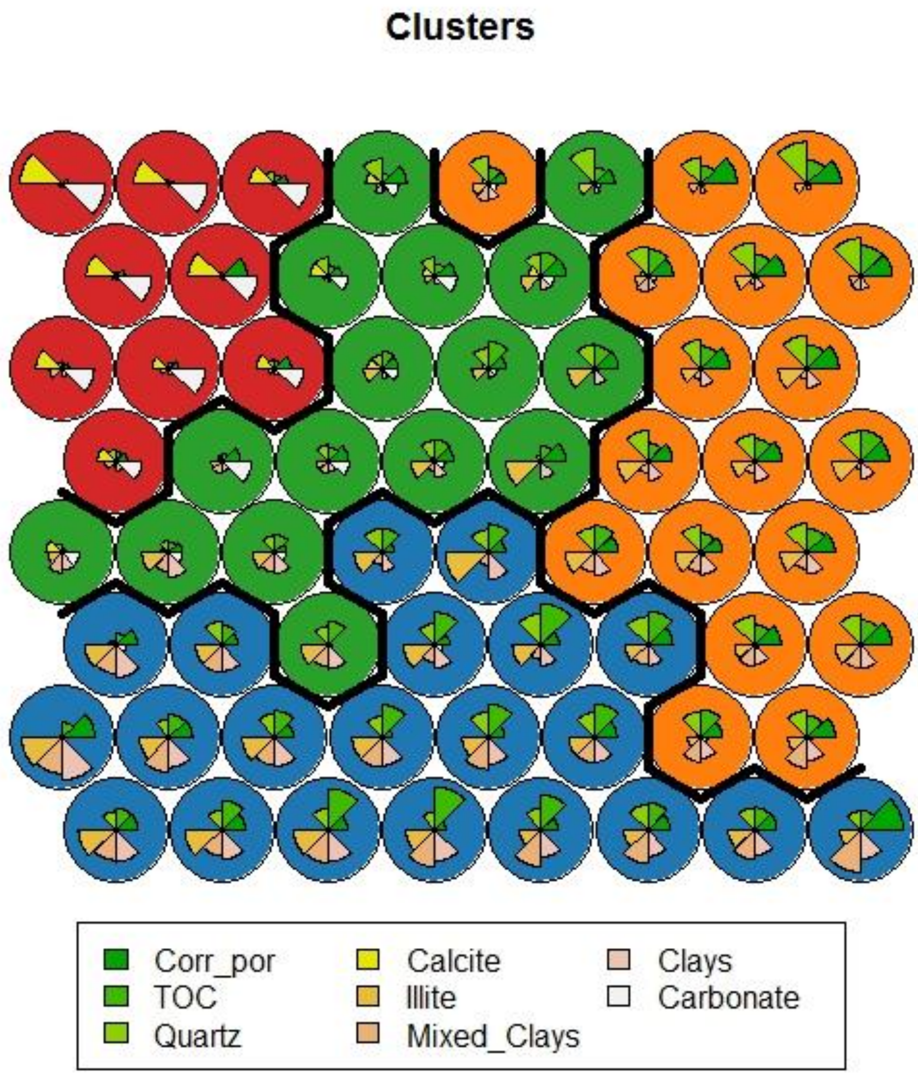

Figure 6. SOM – Clusters

Each grid has the scaled petrophysical properties and corresponded to proper clusters as shown in **Table 3.** Each cluster is characterized by the average properties with minimum and maximum (scaled) values. These clusters are identifying petrofacies as summarized in **Table 4**.

Table 3. Details of 8 x 8 SOM model

Grid	Scaled New Test Data								Petro-facies (Cluster)
	Corr_por	TOC	Quartz	Calcite	Illite	Mixed_ Clays	Clays	Carbonate	
1	0.229	0.062	-0.212	-0.635	1.486	1.197	1.413	-0.792	1
2	0.323	0.747	-0.075	-0.635	1.496	0.835	1.209	-0.747	1
3	-0.128	1.802	0.125	-0.635	1.258	1.399	1.393	-0.708	1
4	-1.030	1.941	-0.225	-0.635	0.766	1.736	1.361	-0.688	1
5	-0.385	1.245	0.103	-0.635	0.088	1.718	0.927	-0.664	1
6	0.212	0.557	0.434	-0.635	0.299	1.471	0.874	-0.713	1
7	0.297	0.127	0.141	-0.635	0.467	1.117	0.771	-0.401	1
8	2.540	-0.125	-0.176	-0.635	0.030	2.161	1.120	-0.717	1
9	1.378	-0.502	-1.704	-0.623	1.789	1.646	2.451	-0.621	1
10	0.396	0.187	-0.177	-0.635	0.776	1.544	1.199	-0.713	1
11	-0.223	0.507	0.314	-0.635	1.117	0.968	1.071	-0.715	1
12	-1.089	1.050	0.066	-0.580	0.790	1.137	1.006	-0.628	1
13	-0.381	1.127	0.465	-0.635	0.469	1.303	0.879	-0.675	1
14	0.261	1.135	0.418	-0.563	0.432	0.762	0.641	-0.625	1
15	0.124	0.620	0.527	-0.248	-0.680	0.715	0.263	-0.287	2
16	0.909	0.065	0.631	-0.635	-0.152	1.190	0.509	-0.462	2
17	0.190	-0.756	-1.469	-0.080	0.888	1.007	1.431	-0.021	1
18	-0.438	-0.080	0.355	-0.542	0.506	0.912	0.777	-0.607	1
19	-2.286	0.305	-0.113	-0.433	0.599	0.479	0.549	-0.497	3
20	-0.582	1.082	0.120	-0.626	0.941	0.110	0.470	-0.196	1
21	-0.246	1.839	1.044	-0.635	0.720	-0.104	0.243	-0.599	1
22	0.551	1.195	0.775	-0.561	0.161	0.318	0.234	-0.596	1
23	0.787	0.014	0.817	-0.501	-0.103	0.424	0.100	-0.506	2
24	1.260	0.391	0.951	-0.635	0.440	0.563	0.488	-0.740	2
25	-2.055	-1.470	-1.012	0.729	-0.558	0.345	-0.095	0.463	3
26	-0.714	-0.876	-0.756	-0.008	0.552	0.233	0.619	0.219	3
27	-0.999	-0.235	-0.254	-0.362	0.139	0.276	0.195	-0.023	3
28	-0.834	0.349	0.347	-0.635	0.597	-0.316	0.015	-0.804	1
29	-0.281	0.740	0.286	-0.635	1.974	-0.272	0.847	-0.705	1
30	0.424	0.599	0.485	-0.584	0.745	0.278	0.472	-0.629	2
31	0.173	-0.109	0.670	-0.601	0.489	0.309	0.344	-0.594	2
32	0.858	0.320	0.891	-0.606	0.578	0.053	0.286	-0.683	2
33	-0.992	-1.032	-0.867	0.838	-0.908	-0.797	-0.779	1.184	4
34	-0.219	-1.140	-1.363	-0.333	-0.892	-0.250	-0.817	1.427	3
35	-0.010	-0.629	-0.405	-0.520	-0.747	-0.261	-0.739	0.781	3
36	-0.161	0.026	-0.008	-0.255	0.306	-0.452	-0.134	-0.024	3
37	0.480	-0.520	-1.333	-0.501	1.307	-1.247	-0.109	-0.614	3
38	0.540	-0.241	0.919	-0.635	0.851	-0.303	0.202	-0.788	2
39	0.850	0.127	0.548	-0.607	0.142	-0.249	0.275	-0.713	2
40	0.963	0.856	0.875	-0.584	0.669	-0.569	-0.001	-0.661	2

41	-1.497	-1.395	-1.269	1.532	-0.701	-1.248	-1.017	1.573	4
42	-1.196	-1.308	-1.749	0.729	-1.136	-1.012	-1.246	1.928	4
43	-0.456	-1.173	-1.029	0.905	-1.241	-0.881	-1.175	1.353	4
44	-0.414	-0.275	-0.625	0.527	-0.248	-1.247	-0.804	0.471	3
45	-0.297	0.581	0.017	-0.137	-0.418	-1.248	-0.923	-0.105	3
46	0.421	0.176	0.358	-0.434	0.535	-0.850	-0.212	-0.520	3
47	1.004	0.157	0.796	-0.506	-0.136	-0.384	-0.083	-0.647	2
48	1.294	0.258	1.243	-0.635	0.718	-0.398	0.089	-0.727	2
49	-1.224	-1.489	-1.424	1.777	-1.534	-1.248	-1.437	1.913	4
50	0.574	-1.763	-1.458	2.122	-1.804	-1.075	-1.479	1.856	4
51	-0.756	-0.854	-0.124	0.872	-1.405	-1.080	-1.332	0.688	3
52	0.004	-0.713	-0.119	0.375	-0.944	-1.078	-1.049	0.844	3
53	0.489	0.308	0.223	0.427	0.000	-0.965	-0.485	0.172	3
54	0.603	0.594	0.900	-0.085	-0.689	0.026	-0.537	-0.167	2
55	1.204	0.147	1.287	-0.396	-0.190	-0.698	-0.494	-0.415	2
56	1.319	0.893	1.680	-0.635	-0.867	-0.194	-0.558	-0.708	2
57	-1.560	-1.840	-1.895	2.836	-1.586	-1.248	-1.581	2.391	4
58	-1.755	-1.661	-1.472	2.235	-1.882	-1.204	-1.612	2.113	4
59	-0.868	-1.026	-0.687	1.471	-1.658	-1.247	-1.824	1.675	4
60	0.551	-0.861	0.469	0.672	-1.437	-0.481	-1.144	0.483	3
61	-0.190	-0.293	0.713	-0.137	-0.470	0.120	-0.278	-0.109	2
62	0.033	-0.323	1.444	-0.300	-0.470	-1.247	-0.895	-0.369	3
63	1.821	-0.185	0.617	-0.635	-0.836	-1.247	-0.967	-0.589	2
64	1.550	0.298	2.087	-0.635	-0.792	-1.242	-1.296	-0.534	2

Table 4. Determined Petrofacies for 8 x 8 SOM model

	Petrofacies 1 (blue)		**Petrofacies 2 (orange)**		**Petrofacies 3 (green)**		**Petrofacies 4 (red)**	
	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.
Corr_por	-1.089	2.540	-0.190	1.821	-2.285	0.550	-1.754	0.574
TOC	-0.755	1.941	-0.293	0.893	-1.469	0.580	-1.840	-1.026
Quartz	-1.703	1.044	0.484	2.086	-1.362	1.444	-1.894	-0.687
Calcite	-0.635	-0.080	-0.635	-0.085	-0.519	0.872	0.729	2.836
Illite	0.030	1.973	-0.866	0.850	-1.436	1.306	-1.881	-0.701
Mixed_Clays	-0.316	2.161	-1.247	1.189	-1.247	0.478	-1.248	-0.797
Clays	0.015	2.451	-1.295	0.508	-1.332	0.619	-1.824	-0.779
Carbonate	-0.804	-0.021	-0.787	-0.108	-0.613	1.426	1.184	2.391

c. Prediction of New data

Using SOM model constructed in the previous step, new test data can be characterized with the determined four(4) petrofacies (clusters). The remaining 42 data points out of 167 total data points are examined to predict the petrofacies. In order to compare the properties, this new data set is also scaled.

Using *predict.kohonen* function in R, the scaled test data is predicted and corresponded to the grids of the original training SOM clusters. **Table 5** shows the results of prediction.

Table 5. Predicted Grid of New Test Data

Test Data	Scaled New Test Data								Predicted Grid
	Corr_por	TOC	Quartz	Calcite	Illite	Mixed_ Clays	Clays	Carbonate	
1	0.695	0.470	0.887	-0.545	0.192	1.010	0.517	-0.689	16
2	1.042	-0.198	1.150	-0.545	0.716	1.069	0.922	-0.726	24
3	-0.584	0.664	1.545	-0.313	0.266	-1.249	-0.585	-0.233	62
4	-0.4273	0.108	1.019	-0.545	-0.215	1.949	0.712	-0.631	6
5	1.084	0.644	1.625	-0.545	-0.833	1.025	-0.218	-0.532	56
6	0.107	0.724	0.325	-0.545	0.695	1.421	1.087	-0.731	6
7	-0.238	1.186	0.298	-0.545	0.517	0.276	0.441	-0.511	20
8	-2.526	-2.206	-1.712	4.199	-3.000	-1.264	-2.730	3.348	57
9	-0.259	1.533	-0.140	-0.545	0.140	1.803	0.884	-0.710	5
10	-0.521	-0.312	-0.693	-0.545	1.020	-0.633	0.599	-0.568	36
11	-0.385	1.955	0.430	-0.545	0.255	-0.354	0.081	-0.925	21
12	1.483	-0.145	-0.342	-0.545	1.491	0.937	1.484	-0.836	1
13	0.832	0.731	-0.614	-0.250	0.370	0.085	0.381	-0.244	30
14	0.832	0.704	-0.491	-0.545	0.810	1.715	1.432	-0.773	10
15	-0.080	1.814	-0.649	-0.545	0.695	0.555	0.742	-0.658	14
16	-0.070	0.918	0.694	-0.545	-0.194	0.012	-0.270	-0.113	54
17	0.979	-0.285	-0.834	-0.545	2.088	1.010	1.874	-0.815	1
18	-0.290	-0.861	-1.211	-0.219	-1.105	-0.427	-1.140	1.9606	34
19	0.086	0.577	0.615	-0.545	0.224	1.274	0.674	-0.605	6
20	0.401	0.985	-1.044	-0.376	0.475	-1.763	-0.075	-0.637	37
21	-0.437	0.510	-1.220	0.858	-0.519	-0.853	-0.668	0.509	44
22	-0.164	-0.901	-1.220	-0.200	-0.571	0.159	-0.458	1.279	34
23	0.454	-0.620	-0.412	-0.250	0.789	-0.280	0.637	-0.191	36
24	-0.322	0.242	-0.570	-0.094	0.601	-0.809	0.149	-0.280	36
25	-0.500	1.828	-1.044	-0.545	1.491	-1.440	0.869	-0.710	29
26	0.275	-0.051	-0.456	-0.545	1.575	0.834	1.424	-0.715	1
27	-1.340	0.563	-0.623	1.065	0.391	-0.545	0.186	0.583	44
28	-2.589	-1.550	-1.607	3.653	-2.581	-1.763	-2.325	2.845	57
29	-0.868	-0.045	0.465	0.927	-0.791	-0.824	-0.735	0.677	51
30	-1.057	-1.517	-1.405	-0.037	-1.628	-0.941	-1.785	1.850	42

31	1.672	-0.794	0.764	-0.476	-0.309	0.629	-0.038	-0.720	23
32	2.679	0.650	1.642	-0.432	-0.236	-0.691	-0.518	-0.574	55
33	0.023	-0.399	1.282	0.582	-1.011	-1.528	-1.583	0.714	52
34	0.874	0.363	2.222	0.043	-0.948	-1.058	-1.350	0.043	64
35	0.601	-0.827	-0.272	-0.050	-0.655	-0.295	-0.758	0.949	35
36	-0.374	-0.740	0.404	-0.069	-0.225	0.218	-0.008	0.080	61
37	0.044	-1.664	1.616	-0.056	0.004	-0.324	-0.293	-0.459	61
38	-0.049	-1.309	0.175	0.419	-0.026	-0.075	-0.060	0.164	61
39	0.821	-0.332	0.553	-0.545	0.548	-0.031	0.269	-0.621	38
40	0.359	-1.035	-0.763	-0.000	-0.037	1.025	0.517	-0.202	17
41	-1.393	-1.135	0.008	0.695	-0.173	0.0859	-0.225	0.441	25
42	-0.868	-0.239	-0.395	0.200	-0.299	0.056	-0.060	-0.029	27

By matching the predicted grid number of new test data with the corresponding cluster numbers of the training SOM model in **Table 3**, the petrofacies of the new test data are identified, which results in the characterization of the new test data. However, this prediction is only based on scaled data. **Table 6** shows the predicted petrofacies.

Table 6. Predicted Petrofacies of new test data

Test Data	Predicted Grid	Predicted Petrofacies	Test Data	Predicted Grid	Predicted Petrofacies
1	16	2	22	34	3
2	24	2	23	36	3
3	62	3	24	36	3
4	6	1	25	29	1
5	56	2	26	1	1
6	6	1	27	44	3
7	20	1	28	57	4
8	57	4	29	51	3
9	5	1	30	42	4
10	36	3	31	23	2
11	21	1	32	55	2
12	1	1	33	52	3
13	30	1	34	64	2
14	10	1	35	35	3
15	14	1	36	61	2
16	54	2	37	61	2
17	1	1	38	61	2
18	34	3	39	38	2
19	6	1	40	17	1
20	37	3	41	25	3
21	44	3	42	27	3