[PS]Comparison of Clustering Techniques to Define Chemofacies in Mississippian Rocks in The STACK Play, Oklahoma*

David Duarte[1], Rafael Pires de Lima[1], Roger Slatt[1], and Kurt Marfurt[1]

[1]The University of Oklahoma, Norman, Oklahoma (dduarte@ou.edu)

Abstract

From reducing uncertainty in well correlations to identifying target zones with low and high reservoir quality, chemostratigraphy has demonstrated to be an excellent tool in geosciences. Chemofacies, an analog of lithofacies, are characterized by a signature composition of 30 elements obtained with X-Ray fluorescence (XRF) spectroscopy. However, a standard definition of the chemofacies is ambiguous due to the different techniques available for clustering analysis. We aim to create a methodology for clustering Mississippian strata with a similar elemental composition in the Anadarko Basin. The chemofacies are used for well correlations, paleoenvironment interpretations, identify landing zones and refine a sequence stratigraphic framework.

To address the issue of chemofacies clustering ambiguity, we used different unsupervised learning techniques in over 1000 analyses of XRF spectroscopy, acquired for Mississippian strata in 4 cores located in The STACK Play, Oklahoma. We lead with different questions to define a methodology for defining the chemofacies. The first issue we deal with is the selection of the elements to be clustered. Sometimes chemofacies are based only on the elements that have been used in the literature as a proxy for any geological parameter. For example, Sr, Mg and Ca as carbonates proxies. But with the objective to incorporate information that might escape traditional geological inference, we also used principal component analysis (PCA) as a preprocessing step before clustering the elements. The next challenge we address is to analyze which clustering algorithm can better represent Mississippian strata. We compare the results of Hierarchical cluster analysis (HCA), K-means, Self-organizing map (SOM), and Density-based spatial clustering (DBSCAN). Then, we used PCA to geologically constrain the clusters and define the chemofacies. Finally, the chemofacies were compared with thin sections and well logs. The analysis we performed allowed us to define the most appropriate workflow that honors the geology embedded in the lithofacies. The selection of unsupervised learning algorithm is based both in the resulted chemofacies and the clustering objectives. We propose that the segmentation of massive gravity flows facies from hemipelagic facies can be achieved with two clusters. However, more clusters are necessary if the objective is to identify lower and higher reservoir quality intervals within these two main clusters.

Abstract

Chemostratigraphy has been used as a key tool in geoscience for the identification of target zones with low and high reservoir quality. By grouping zones with similar elemental composition (defined as chemofacies) as determined by X-Ray fluorescence (XRF) spectroscopy, the uncertainty in well correlations can be reduced.

However, the challenge remains in obtaining a unified and unambiguous definition for chemofacies, since different clustering algorithms produce different results. To address these different clustering results, we have used different unsupervised learning techniques in over 1,000 analyses of XRF spectroscopy acquired for Mississippian strata in 4 cores from the STACK play, Oklahoma. Our analysis resulted in the identification of chemofacies that can be used for well correlations, paleoenvironmental interpretations, identification of landing zones, and refining sequence stratigraphic frameworks.

We started by selecting the elements to be clustered based on the information provided by principal components analysis (PCA). This guarantees our methodology will consider all the available chemical information, and not just the one commonly associated to geological proxies. We tested different cluster algorithms: hierarchical cluster analysis (HCA), K-means, and density-based spatial clustering (DBSCAN) using the principal components that represent 80% of the data variance. We then analyzed which cluster algorithm better represent Mississippian strata from a geological standpoint. Finally, the chemofacies were compared with information from thin sections and well logs. This approach has allowed us to define an appropriate workflow that honors the geology embedded in the lithofacies.

Location and Data
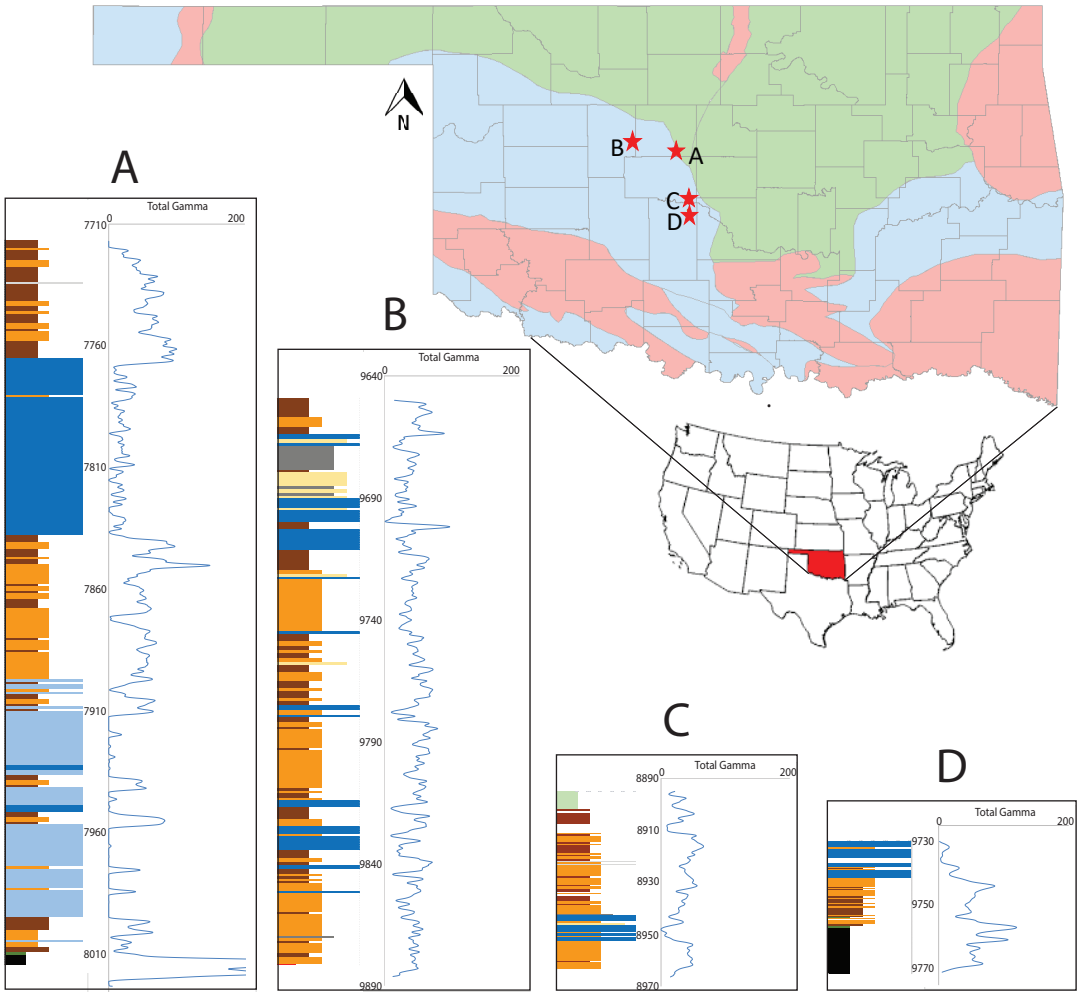
Figure 1. Location, facies classification (legend in Figure 2), and gamma ray response of the four cores used in this research.

Lithological Description and HHXRF Data: Core A

Legend:
- Massive calcite-cemented siltstone
- Wackstones and packstones
- Argillaceous mudstone
- Bioturbated mudstone
- Calcite cement
- Calcareous siltstone
- Interbedded mudstone
- Post-Strata
- Glauconitic sandstone
- Woodford

Figure 2. Core A elemental profile with gamma ray (CGR) profile and facies classification. Elemental composition information is based on hand-held X-Ray fluorescence data. This core contains 300 samples at 1.0 ft resolution. Three formations are present in this core, from bottom to top: Woodford shale, Osage, and Meramec formation.

Comparison of Clustering Techniques to Define Chemofacies in Mississippian Rocks in The Stack Play, Oklahoma

David Duarte, Rafael Pires de Lima, Roger Slatt, and Kurt Marfurt

Methods

Proxies

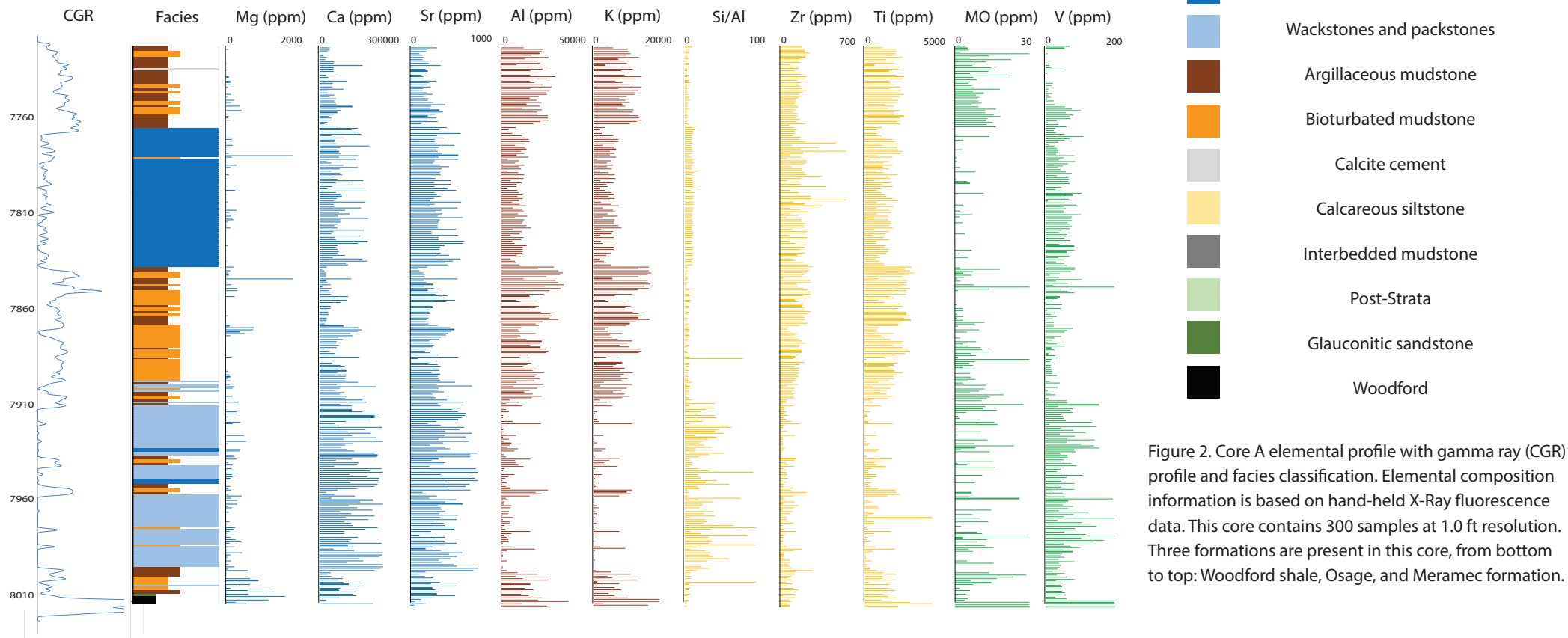| Elemental Proxy	Paleoenvironment Interpretation
Calcium (Ca)	Carbonate Source
Strontium (Sr)	Carbonate Source
Magnesium (Mg)	Dolomite Source
Aluminum (Al)	Clay minerals and Feldspars
Potassium (K)	Clay minerals and Feldspars
Si/Al	Quartz
Titanium	Continentally derived
Zirconium (Zr)	Continentally derived
Molybdenum (Mo)	Bottom water anoxia
Vanadium (V)	Bottom water anoxia

Figure 3. List of proxies used for clustering analysis and their paleoenvironment interpretation.

Selection of Number of Clusters

Figure 4. Elbow method used to choose the appropriate number of clusters.

Comparison of Input Data

Figure 5. Results of clustering analysis using k-means for different input data. Raw is the original data from the HHXRF; scaled refers to the Z-score scaled data; scaled-PCA are the data resulted from principal component analysis; and scaled-proxies are the data only with the selected proxies. Every set of input data (raw, scaled, scaled-PCA, scaled-proxies) is used with all the measurements in the core (left) and only the measurements in the upper bench (right). Colors in every column represent the same cluster. However, colors in panels Raw, Scaled, Scaled-PCA, and Scaled-Proxies are not correlated.

Comparasion of Clustering Algorithms

Figure 6. Results of clustering analysis using k-means, Ward, and DBSCAN for two different input data (scaled and scaled-PCA). Scaled is all the Z-score scaled data scaled-PCA refers to the data resulted from principal component analysis. For this comparision of methods all the data from core A was used. Colors in every column represent the same cluster. However, colors in panels Raw, Scaled, Scaled-PCA, and Scaled-Proxies are not correlated.

Validation of Results: Thin Sections

Figure 7. Thin sections used to compare the results of the clustering analysis from core A. The letter on the thin section corresponds to the letter on the Figures 5 and 6. A) 7741.2 ft. B) 7822.5 ft. C) 7836.7 ft. D) 7844.2 ft. E) 7898.9 ft. F)7916.1 ft. G) 7954.3 ft. H) 7972.2 ft.

Validation of Results: Data Integration

Figure 8. Mineralogical composition used to compare the results of the clustering analysis. Note that the scale from mineralogical composition is not in the same scale than the lithological description or the HHXRF data.
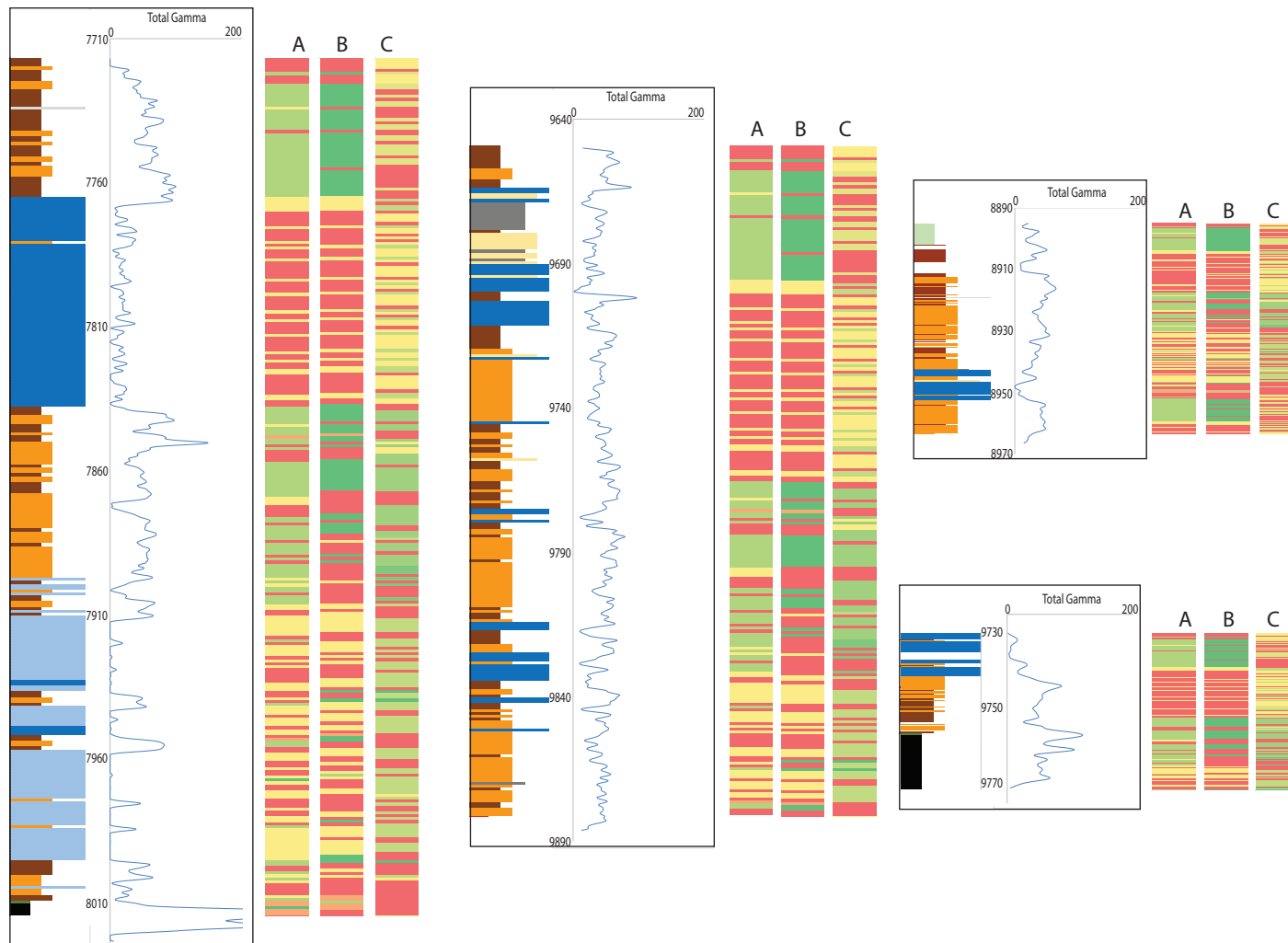
Combining Cores

Figure 9. Comparison of K-means (A), Ward (B), and DBSCAN (C) in all the four cores. All the scaled elements were used in the three scenarios. The colors correlate between the well but not between the algorithms. For example, the red color in all the column A represents the same cluster, but the red color in column A does not represent the same cluster in columns B or C.

Discussion

Data normalization is a very important initial step that sometimes is overlooked by geoscientist practitioners. Without normalization, features with larger scales (e.g. ppm of Ca) can dominate features with smaler scales and variance (e.g. ppm of Mo). This can cause the clusters to be completely determined by very few features, which is usually not desired.

Using all the variables have similar results than when using selected elements (proxies) if the elements are selected properly. While using principal components makes it easier to identify the change between formations, using the scaled elements or proxies makes it easier to identify the changes in lithology.

We note a visual correspondence between the cluster results of K-means and Ward. Although not completely different from the first two, DBSCAN clusters present a slightly different characheristic, indicating that the method is sensitive to different aspects of the data.

Unlike K-means and Ward clustering methods, DBSCAN automatically assigns outliers for samples lying too far from the others. This can help the interpreter find patterns in the data that could be overlooked when using only more traditional methods (such as K-means).

Future Work

In this study, we focused on the use of unsupervised learning techniques to compare the different clustering results. We performed a qualitative analysis between the clusters based mostly on our expertise and geological knowledge. We intend to improve our analysis using more quantititative metrics (such as the computation of Mahalanobis distance within clusters)

For our next steps, we plan to use supervised learning methods and evaluate whether the prior labeling of geologically defined facies improves our classification results.

The results we present here are not extensive and can likely be improved if more data are available.

Conclusions

There is not a standard method for clustering XRF data that can be applied for every scenario.

Defining the clustering objective must be the first step in any unsupervised learning analysis.

Scaling the data is an important processing step when dealing applying machine learning techniques.

Segmenting the data in similar rock types helps in the detection of high and low reservoir quality rock within and specific interval.

Using all the variables have similar results than when using selected elements (proxies) if the elements are selected properly.

Principal component analysis is useful when trying to identify large changes in the data (e.g. different formations).

Acknowledgments