A spectral approach for the clustering of source rocks

V. I. Makri[a,b], D. Pasadakis[c]

[a] Institute of GeoEnergy (IG)-FORTH, Chania, Greece.
[b] Technical University of Crete, School of Mineral Resources Engineering, Chania, Greece.
[c] Università della Svizzera italiana, Institute of Computing, Lugano, Switzerland.

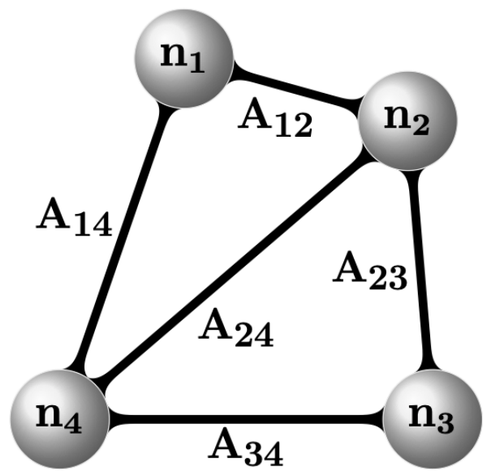
- S. White and P. Smyth, *A spectral clustering approach to finding communities in graphs*, in Proceedings of the 2005 SIAM International Conference on Data Mining (SDM), pages 274-285, 2005.
- D. Pasadakis, C. L. Alappat, O. Schenk, and G. Wellein, *Multiway p-spectral graph cuts on Grassmann manifolds*, Machine Learning 111, pages 791-829, 2022.
- K. Peters, C. Walters, and J. Moldowan, *Geochemical correlation and chemometrics*, In The Biomarker Guide, pages 475-482, Cambridge University Press, 2004.

Spectral multiway clustering

For a graph $\mathcal{G}(V, E, \boldsymbol{A})$ with n vertices:

- Adjacency: $\boldsymbol{A} \in \mathbf{R}^{n\times n}$, degree $\boldsymbol{D} \in \mathbf{R}^{n\times n}$.
- Graph Laplacian matrix:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} = \begin{bmatrix} \sum_j \boldsymbol{A}_{1j} & -\boldsymbol{A}_{12} & 0 & -\boldsymbol{A}_{14} \\ -\boldsymbol{A}_{12} & \sum_j \boldsymbol{A}_{2j} & -\boldsymbol{A}_{23} & -\boldsymbol{A}_{24} \\ 0 & \boldsymbol{A}_{23} & \sum_j \boldsymbol{A}_{3j} & -\boldsymbol{A}_{34} \\ -\boldsymbol{A}_{14} & -\boldsymbol{A}_{24} & -\boldsymbol{A}_{34} & \sum_j \boldsymbol{A}_{4j} \end{bmatrix}$$



➥ Normalized variant: $\boldsymbol{L}_{\text{rw}} = \boldsymbol{D}^{-1}\boldsymbol{L}$, with

$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, $\lambda_1 = 0$ for a connected graph.

Identify k clusters using the smallest k eigenvectors of $\boldsymbol{L}_{\text{rw}}$.

Rayleigh quotient minimization

$$\min_{\boldsymbol{U}\in\mathbb{R}^{n\times k}} F(\boldsymbol{U}) = \operatorname{Tr}(\boldsymbol{U}^\intercal \boldsymbol{L}\boldsymbol{U}),$$
subject to $\boldsymbol{U}^\intercal\boldsymbol{U} = \boldsymbol{I}$.

Graph construction

- Input data $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, n: source rock samples, m: n-alkane peak areas.
- Connectivity $\boldsymbol{G}$ via a k-nearest neighbors (kNN) routine.
- Similarity $\boldsymbol{S}$ via a Gaussian kernel.

➜ $\boldsymbol{A} = \boldsymbol{G} \odot \boldsymbol{S}$ for a sparse graphical representation of the **source rocks**.

Spectral graph coordinates

	$\boldsymbol{u}_1$	...	$\boldsymbol{u}_k$
$\boldsymbol{U}_1$	u_{11}	...	u_{1k}
$\vdots$	$\vdots$	...	$\vdots$
$\boldsymbol{U}_n$	u_{n1}	...	u_{nk}

✔ Direct multiway clustering.

✔ Global information.

Maximize modularity function

$$Q = \sum_i \left(\boldsymbol{D}_{ii} - \left(\sum_j \boldsymbol{D}_{ij} \right)^2 \right)$$

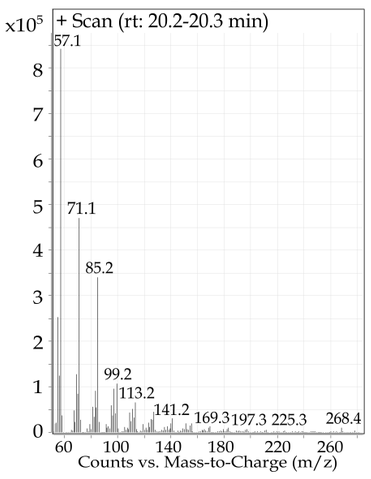
➜ High values: clear community structure.

Data background

For this study, a set of 83 rock samples from Western Greece potential source rocks was used. These samples are of multiple geological ages, located in the so-called Ionian geotectonic zone.
They were treated in the lab with the following experimental prodecure:

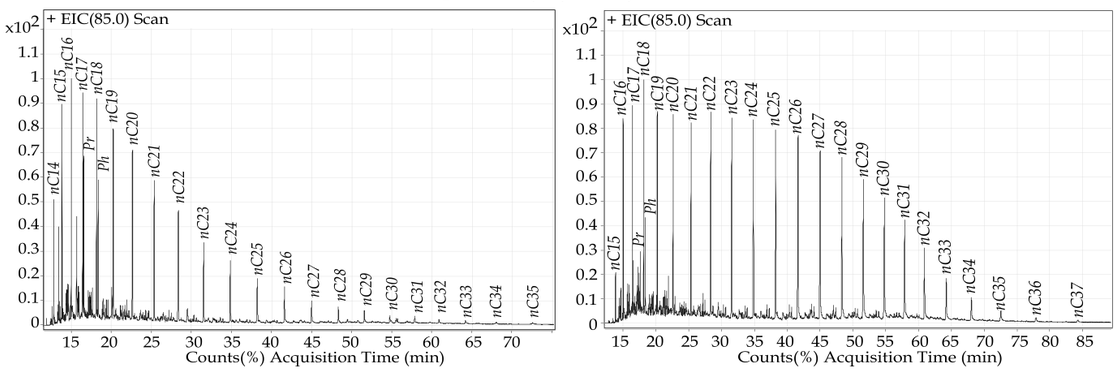
Experimental procedure

nC19 mass spectrum



- Soxhlet extraction using a DCM:MeOH mixture (90:10 v/v).
- De-asphaltening by n-pentane (7ml) and filtation through Teflon syringe filters (0.45μm).
- SARA fractionation technique on silica-alumina chromatographic column for the separation of maltenes into saturated, aromatic and NSO fractions.
- GC-MS analysis on Agilent 7890A Gas Chromatograph (HP-5MS UI (60m x 250μm x 0.25μm column)) coupled to an Agilent 5975E Mass Spectrometer.
- Identification of peak areas of n-alkanes and isoprenoids by their relative retention times and mass spectra.
- Peak areas of nC15-nC35 were used for the multiway spectral clustering.

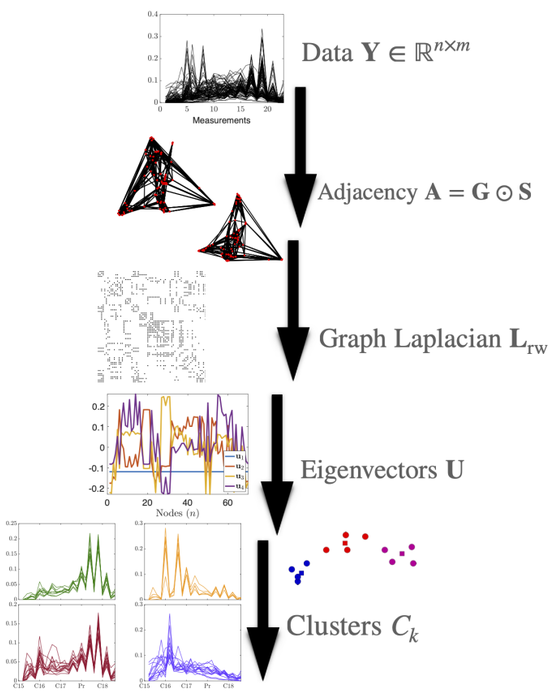
GC-MS chromatograms (m/z 85) of saturated fractions



Key algorithmic components

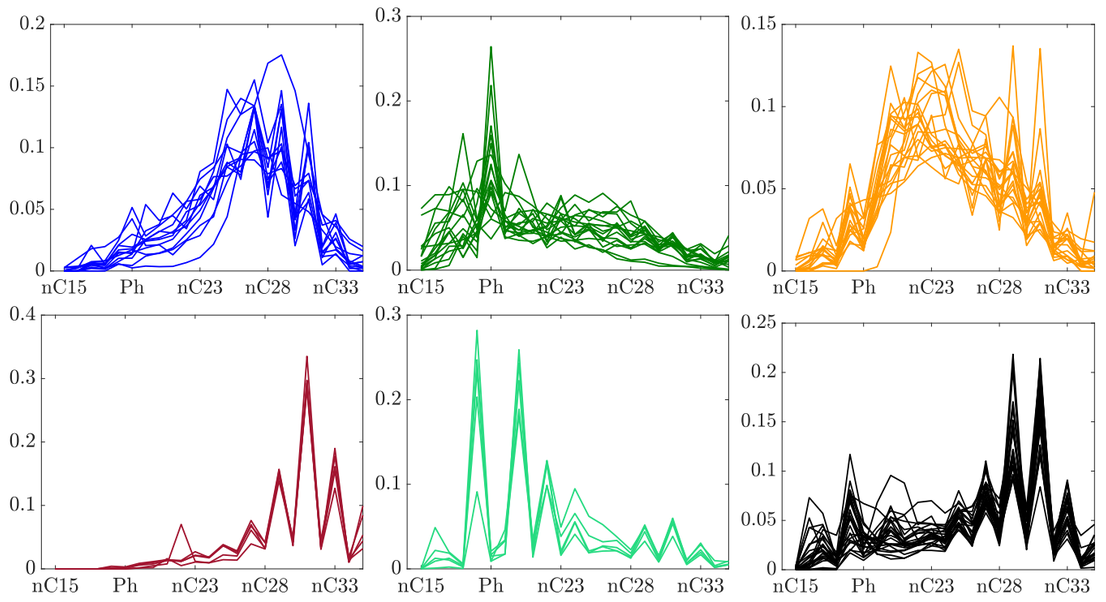
Spectral clustering of source rocks

Input: Data $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ ▷ Source rock measurements
Output: C_k ▷ Source rock clusters
1: Data normalization ▷ Sum of peak areas
2: Build adjacency matrix $\boldsymbol{A}$ ▷ $\boldsymbol{A} = \boldsymbol{G} \odot \boldsymbol{S}$
3: Construct the graph Laplacian $\boldsymbol{L}_{\text{rw}}$
4: Rayleigh quotient min for the first $n/5$ eigenvectors $\boldsymbol{U}$
5: **while** $k \in [2, n/5]$ **do**
6: Cluster $\boldsymbol{U}$ in k groups ▷ Using kmeans
7: Compute the modularity Q_k of the k clusters
8: **end while**
9: Select the clusters C_k for which $Q = \max(Q_k)$



Code available at https://github.com/DmsPas/Spectral-clustering-of-source-rocks.git

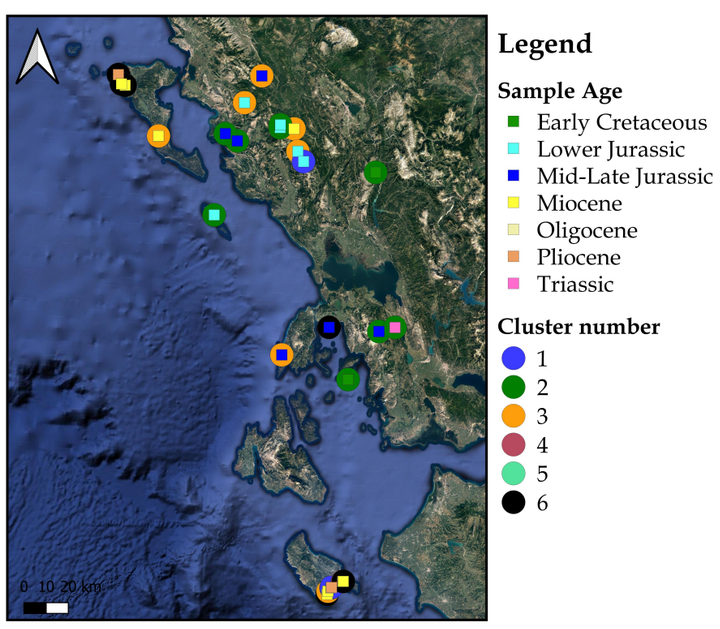
Numerical results



The direct multiway spectral clustering routine identifies 6 clusters in the dataset. The maximum value of modularity reported is $Q = 0.71$. Inspecting the chromatograms of each cluster we observe clear distinctions between the groups.

Interpretation and conclusions

Map distribution



Conclusions

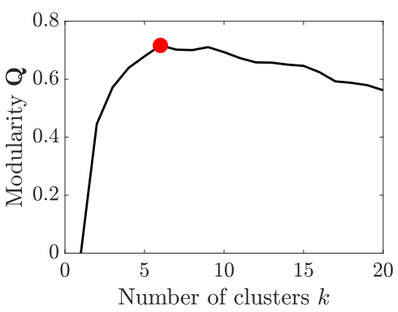
- The 6 "naturally existing" groups showcase distinct concentration profiles of the n-alkanes of the samples.
- A geographical view of the clusters is shown on the map.
- Robust algorithm that can handle datasets of varying size and complexity.
- Higher accuracy compared to traditionally employed methods in Geochemistry (like hierarchical clustering and PCA)
- The reliability of this technique to support the classification of source rocks is proved.
- This technique may be employed in studies involving chromatographic analytical data such as oil fingerprinting, oil-oil and oil-source correlations, as well as bio-informatics.

Acknowledgements

This study is supported by Hellenic Petroleum S.A. (HELPE).

Estimating the number of clusters

Maximizing the modularity of the graph results in high quality clusters and leads to high intra-cluster and low inter-cluster connectivity.



12th FORTH RETREAT 2022