Contents lists available at ScienceDirect

International Journal of Coal Geology

journal homepage: www.elsevier.com/locate/coal

Automatic organofacies identification by means of Machine Learning on Raman spectra

Natalia A. Vergara Sassarini [a,b,c,*,1], Andrea Schito [d], Marta Gasparrini [e], Pauline Michel [b], Sveva Corrado [a]

[a] Dipartimento di Scienze, University of Roma Tre, Largo San Leonardo Murialdo 1, 00146 Rome, Italy
[b] IFP Energies nouvelles, 1-4 Avenue du Bois-Préau, 92852 Rueil-Malmaison, France
[c] Sorbonne Université, ED 398 – GRNE, 4 place Jussieu, 75252 Paris, France
[d] Department of Geology and Geophysics, School of Geosciences, University of Aberdeen, Aberdeen AB24 3UE, UK
[e] University of Milan, Earth Sciences Department, via Mangiagalli 34, 20133 Milan, Italy

ARTICLE INFO

Keywords:
Raman spectroscopy
Machine learning
Cluster analysis
Dispersed organic matter
Principal component analysis
Thermal maturity

ABSTRACT

In this study we compare and evaluate different unsupervised clustering algorithms for organofacies discrimination in low maturity dispersed organic matter based on Raman spectroscopic analyses. A total of 1363 Raman spectra were collected from a set of 27 organic-rich samples from the Lower Toarcian shale interval of the Paris Basin sub-surface. Rock-Eval pyrolysis data indicate a type II to type III kerogen with a vitrinite reflectance (R_o%) between 0.45% and 0.65%, and T_{max} between 415 °C and 438 °C. Organic petrographic observations under transmitted light reveal the presence of organofacies composed by amorphous organic matter, opaque, and translucent phytoclasts. An optical classification of organic particles was performed on about 40–60 fragments per sample and used as the ground truth. Raman spectra were obtained for all the classified fragments and principal component analysis was performed to underline the variability among spectra. Unsupervised clustering was then applied on Raman spectra principal components. Three clustering methods were applied to evaluate their effectiveness in predicting number, shape and density of clusters and a contingency matrix was used to quantify their ability to predict different organofacies. Gaussian Mixture Model (GMM) was found to be the best algorithm for organofacies identification showing an accuracy mostly between 80% and 90%. This work outlines how unsupervised clustering of Raman spectra of dispersed organic matter can reduce the uncertainty in thermal maturity assessment and help the classification of highly heterogeneous organofacies when using large datasets for Earth and planetary sciences studies.

1. Introduction

The characterization of dispersed organic matter (OM) in sedimentary successions has classically been used in basin analysis for paleoenvironmental and paleo-thermal reconstructions and hydrocarbon exploration (Vandenbroucke and Largeau, 2007; Allen and Allen, 2013). In recent years, it has also been applied to investigate the deformation sequence and structural style in fold-and-thrust belts (Corrado et al., 1998, 2010, 2021; Toro et al., 2004; Di Paolo et al., 2012; Caricchi et al., 2016; Schito and Corrado, 2018; Aldega et al., 2018; Balestra et al., 2019; Lucca et al., 2019; Atouabat et al., 2020; Muirhead et al., 2020; Gusmeo et al., 2022), to evaluate the role of carbon degassing in regulating the global carbon cycle during the interaction between massive volcanic emplacements and source rocks (Marzoli et al., 1999; Svensen et al., 2004; Iyer et al., 2018), to determine the temperatures achieved during frictional heating along fault planes (Kitamura et al., 2012; Kaneki et al., 2016; Kedar et al., 2020; Muirhead et al., 2021), hydrothermal alteration (Schito et al., 2022), or meteorite impacts (Parnell et al., 2005), and to planetary exploration (Hickman-Lewis et al., 2022; Westall et al., 2021; Bhartia et al., 2021; Farley et al., 2020; Quirico et al., 2020).

In basin analysis studies, a realistic reconstruction of the time-temperature history of the sedimentary infill is pivotal to understand basin evolution and hydrocarbon generation and migration and is

[*] Corresponding author at: Dipartimento di Scienze, University of Roma Tre, Largo San Leonardo Murialdo 1, 00146 Rome, Italy.
E-mail address: amanda.vergara@unipd.it (N.A. Vergara Sassarini).
[1] Present address: Center of Studies and Activities for Space (CISAS) “G. Colombo”, University of Padua, 35,131 Padua, Italy.

https://doi.org/10.1016/j.coal.2023.104237
Received 12 October 2022; Received in revised form 5 April 2023; Accepted 5 April 2023
Available online 8 April 2023

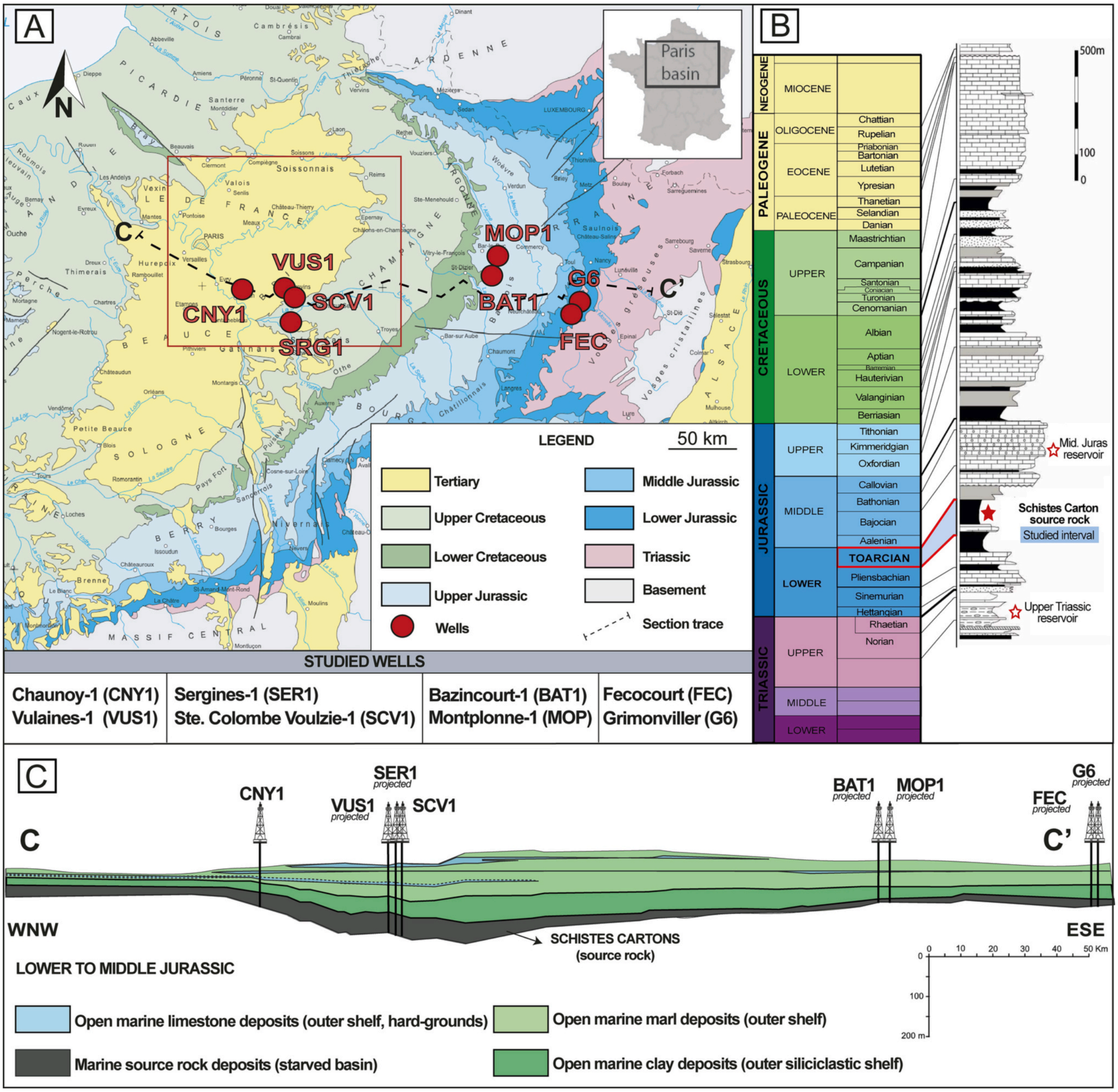

Fig. 1. Geographic and geological setting of the study area in the Paris Basin. The red box depicts the basin depocenter. (A) Geological map of the Paris Basin (modified after Gély and Hanot, 2014) with location of the basin depocenter (red frame) and the eight studied wells. In the upper right frame is the location of the Paris Basin in northern France. The dashed black line represents the trace of the cross-section illustrated in C. (B) Synthetic stratigraphic column of the Paris Basin sedimentary pile. The studied interval is indicated with the solid red star, whereas the main reservoirs are highlighted with empty red stars. (C) WNW-ESE geological cross-section across the Paris Basin illustrating the facies distribution of the Lower to Middle Jurassic sedimentary pile with the projection of the eight wells of interest (modified after Delmas et al., 2002). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

commonly calibrated by means of organic matter thermal maturity (Palumbo et al., 1999; Harris and Peters, 2012; Allen and Allen, 2013). In this context, vitrinite reflectance is considered a highly reliable and reproducible tool for dispersed OM thermal maturity assessment (e.g., McCartney and Teichmüller, 1972; Davis, 1978; Tissot, 1984; Izart et al., 2016; among others), despite some limitations (Schito et al., 2016; Corrado et al., 2020; Luo et al., 2020; Nirrengarten et al., 2020; Sorci et al., 2020). Dispersed OM is a mixture of heterogeneous organic materials derived primarily from the decomposition products of plant material, bacteria and algae, whose geochemistry both reflects the original depositional environment and the burial history experienced by the sedimentary succession. Reflected light optical microscopy allows to characterize complex assemblages of organic constituents (macerals), while vitrinite reflectance measurements consent to determine thermal maturity degree. Nevertheless, this last method is semi-quantitative, time-consuming, and strongly dependent on the operator skill to correctly identify the single organic matter constituents. In this context, Raman spectroscopy has recently received attention as a tool to evaluate thermal maturity of organic matter in source-rocks to complement and enhance the more traditional approaches (see Henry et al., 2019 and Schito et al., 2023 as the most recent reviews), being a non-destructive, high-resolution technique (e.g., Beny-Bassez and Rouzaud, 1985; Jehlička and Bény, 1992; Wopenka and Pasteris, 1993), that returns information on the molecular structure of the OM.

Table 1
Rock-Eval pyrolysis and vitrinite reflectance data (from Corrado et al., 2022; Vergara Sassarini, 2022). For each well: sample name and depth, R_o% with standard deviation T_{max} and Total Organic Content (TOC).

Well	Sample name		Depth (m)	[1]R_o (%)	[2]T_{max} (°C)	[3]TOC (%)
BAZINCOURT1	**BAT1**	**bat1–1**	930	0.62 ± 0.05	431	1.04
		bat1–2	950	0.61 ± 0.06	432	1.04
CHAUNOY1	**CNY1**	**cny1–1**	2000	0.55 ± 0.06	429	2.00
		cny1–2	2006	0.63 ± 0.07	430	1.68
		cny1–3	2008	0.62 ± 0.07	431	1.26
		cny1–10	2033	0.60 ± 0.07	432	3.44
		cny1–12	2046	0.59 ± 0.05	433	4.21
		cny1–14	2053	0.59 ± 0.04	433	5.28
FECOCOURT	**FEC**	**fec-1**	0	0.49 ± 0.07	417	9.39
		fec-2	0	0.43 ± 0.06	421	13.89
		fec-3	0	0.48 ± 0.07	415	9.19
GRIMONVILLER	**G6**	**g6–1**	7.25	0.49 ± 0.04	423	7.93
		g6–2	14.9	0.54 ± 0.04	425	4.63
		g6–3	21.7	0.52 ± 0.07	421	9.53
MONTPLONNE 1	**MOP1**	**mop1–1**	948	0.54 ± 0.07	432	1.14
		mop1–2	1018	0.64 ± 0.07	420	7.21
		mop1–3	1026	0.62 ± 0.07	419	6.52
SERGINES 1	**SRG1**	**ser1–1**	1913.4	0.66 ± 0.05	435	0.86
		ser1–2	1916	0.65 ± 0.07	433	0.99
STE COLOMBE VOULZIE 1	**SCV1**	**stcv1–1**	2210	0.63 ± 0.06	438	1.16
		stcv1–2	2254	0.62 ± 0.07	437	3.61
		stcv1–3	2270	0.64 ± 0.07	437	6.11
		stcv1–4	2285	0.65 ± 0.06	435	6.04
VULAINES 1	**VUS1**	**vul1–1**	2136	–	433	1.55
		vul1–2	2138	0.58 ± 0.05	430	1.87
		vul1–3	2150	0.65 ± 0.05	433	1.33
		vul1–4	2154	0.62 ± 0.07	432	1.31

[1] Vitrinite reflectance (R_o %) from Corrado et al., 2022; Vergara Sassarini (2022).

[2] T_{max} calculated through Rock-Eval 6 from Corrado et al., 2022; Vergara Sassarini (2022).

[3] Total Organic Content (TOC) from Corrado et al., 2022; Vergara Sassarini (2022).

Many studies have focused on the correlation between Raman-derived parameters and thermal maturity on both continental and marine derived OM (see Henry et al., 2019 for a review). Some studies have found good correlations for the range of maturity crucial to oil and gas exploration (>0.5 R_o%; e.g., Cheshire et al., 2017; Sauerer et al., 2017–2021; Al-Hajeri et al., 2020–2021). However, Raman technique application still needs to be optimized for the immature stage of hydrocarbon generation (<0.5–0.6 Ro%; Dow, 1977; Tissot, 1984) where thermal maturity can be challenging to assess (e.g., Henry et al., 2019 and references therein; Karg and Sauerer, 2022) and, thus, it is still tied to optical petrography classification (Hinrichs et al., 2014; Wilkins et al., 2014; Lünsdorf, 2016; Henry et al., 2018).

In the past few years, only a handful of works have discussed about the use of Machine Learning (ML) techniques to quicken and automate the optical characterization of organic matter (Mlynarczuk and Skiba, 2017; Skiba and Młynarczuk, 2018; Wang et al., 2019; Lei et al., 2021; Wang et al., 2022). Most of these works are based on the study of coals, where maceral characterization relies upon the analysis of images acquired through optical petrography using supervised ML techniques, the efficiency of which depends on the availability and quality of a suitable training set (i.e., classified dataset or ground truth) (Mlynarczuk and Skiba, 2017; Skiba and Młynarczuk, 2018; Wang et al., 2019; Lei et al., 2021; Wang et al., 2022).

The use of supervised ML on OM characterization by means of Raman spectroscopy, on the other hand, has been tested only in a few works (Schito et al., 2019–2021). Relying on a classified dataset these authors demonstrate the potential of supervised ML on Raman spectra for organofacies identification. Nevertheless, the application of the learning algorithms depends, as well, on the availability of a consistent classified dataset.

Till date, no attempts on the application of unsupervised ML techniques (i.e., unknown ground truth) on Raman spectra have been made. In this study, an automatic classification of organoclasts (amorphous organic matter, translucent and opaque phytoclasts), based on unsupervised learning/clustering techniques applied on Raman spectra is proposed for the first time on low diagenesis dispersed OM in which high organofacies heterogeneity often prevents a correct thermal maturity assessment.

2. Geological setting

The Paris Basin is the largest onshore sedimentary basin in France (110.000 km^2). It is a Mesozoic and Cenozoic intracratonic basin (Brunet and Le Pichon, 1982; Pomerol, 1989; Perrodon and Zabek, 1991), developed on a deformed Paleozoic basement interpreted as the northern branch of the Variscan thrust belt (Delmas et al., 2002; Guillocheau et al., 2000; Gonçalvès et al., 2003). Its evolution started in Late Permian-Early Triassic with the rifting stage related to the Tethys

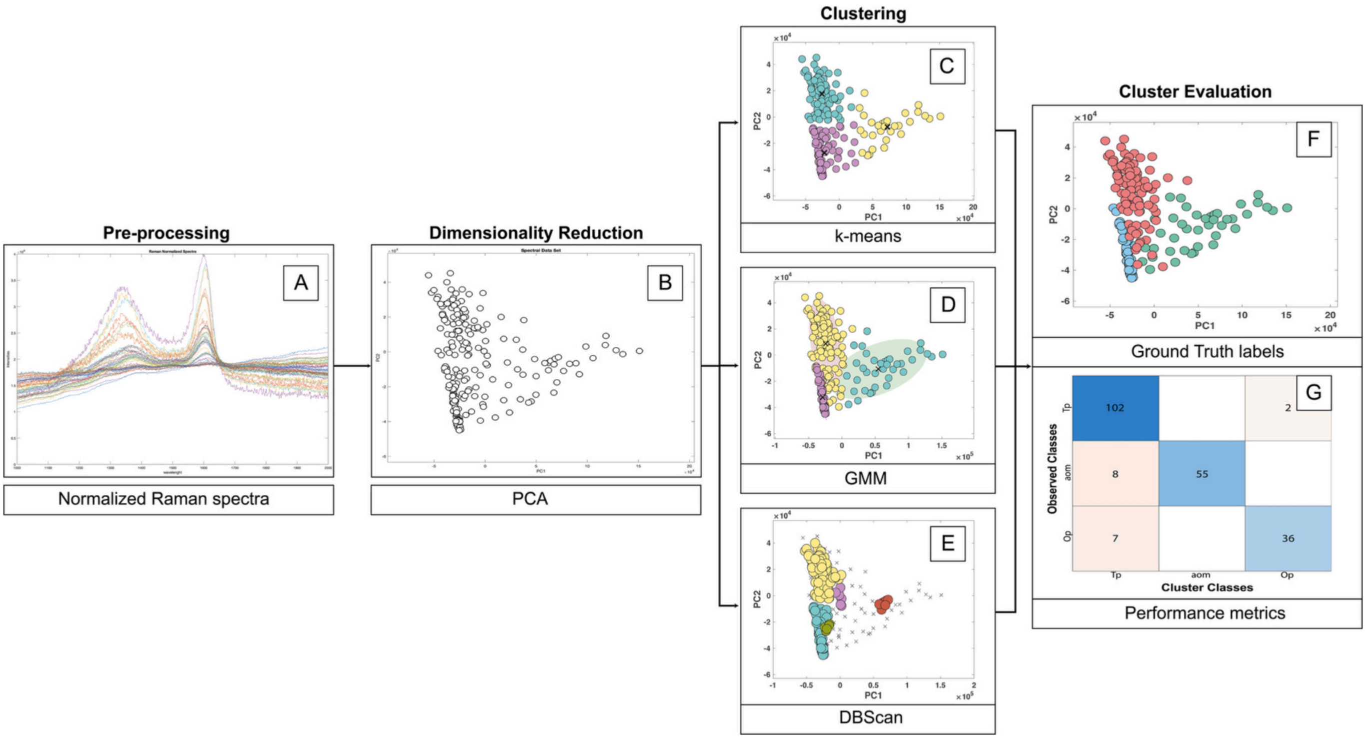

Fig. 2. Workflow of the unsupervised learning routine from raw Raman spectra for evaluation. Example from well SER1. (A) Plot of normalized Raman spectra; in the x-axis the Raman shift (cm^{-1}), in y-axis the Intensities; (B) Scatter plot of Raman spectra after PCA. In the x-axis is the first Principal Component (PC1) whereas in the in y-axis is the second Principal Component (PC2); (C) Clustering output for k-means algorithm; (D) Clustering output for GMM algorithm; (E) Clustering output for DBSCAN algorithm; (F) Classified dataset or ground truth. AOM = Amorphous Organic Matter (blue circles); Tp = Translucent phytoclasts (red circles), Op = Opaque phytoclasts (green circles); (G) Contingency matrix illustrating the predicted (x-axis) versus the observed classes (y-axis). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

opening, followed by a period of thermal subsidence during Jurassic to Early Cretaceous times due to the opening and closing of the Tethys Sea and the opening of the Atlantic Ocean (Brunet and Le Pichon, 1982; Perrodon and Zabek, 1991; Guillocheau et al., 2000). Maximum burial depths were reached during Cretaceous times, followed by an uplift event of variable intensity associated with the late phases of the Late Cretaceous and Early Paleogene Africa-Europe convergence (Guillocheau et al., 2000). The central part of the basin is filled up with about 3000 m thick deposits spanning in age from Permian to Quaternary (Fig. 1) and includes two main hydrocarbon reservoir units: the Upper Triassic fluvial sandstones (Chaunoy Formation) and the Middle Jurassic marine carbonates (Delmas et al., 2002). The principal source rock of the basin corresponds to the Lower Toarcian black shales denominated “*Schistes Carton*”, essentially consisting in organic matter-rich marls that experienced oil-window temperatures in the basin depocenter since the end of the Cretaceous (Poulet and Espitalié, 1987; Espitalié et al., 1988; Katz, 1995; Disnar et al., 1996; Delmas et al., 2002). The lateral and vertical oil migration towards both Upper Triassic and Middle Jurassic reservoirs occurred along normal faults mainly in Late Cretaceous times (Delmas et al., 2002).

The “*Schistes Carton*” were deposited in shallow waters characterized by intense activity of pelagic organisms and by normal salinity (Tissot et al., 1971; Espitalié et al., 1988; Fonseca et al., 2021). The analysis of fossil content, as well as the consistent distribution of clay minerals and trace elements of organic matter across the basin, suggests that the shale samples originated from a singular stratigraphic zone (early falciferum Zone) and were deposited under relatively homogenous environmental conditions as a uniform layer. (Tissot et al., 1971; Disnar et al., 1996). Kerogen analysis shows that the “*Schistes Carton*” are a type-II, III and mixed II-III source rock (Vergara Sassarini, 2022), while Rock-Eval pyrolysis shows generally high oil contents, up to 60 kg of hydrocarbons per ton of shale and TOC values up to 10% (Corrado et al., 2022; Vergara Sassarini, 2022; Espitalié et al., 1987).

3. Materials and methods

3.1. Materials

Eight wells from the Paris Basin crossing the “*Schistes Carton*” were selected. Wells are distributed along a ~ 500 km long WNW-ESE transect (Fig. 1) in which the sampled interval occurs at progressively shallower depths going from the depocenter (~2300 m) to the eastern margin of the basin (0 m; Table 1). In total, 27 samples were selected for analyses. Samples are characterized by a range of maturities going from the immature to the onset of the oil-window stage (0.43 to 0.66 R_o%; T_{max} between 415 and 438 °C; Table 1; Corrado et al., 2022; Vergara Sassarini, 2022).

Raman spectroscopy analyses were performed on isolated kerogen, extracted by means of the traditional technique described in Traverse (2007). Each sample (*ca* 30 g) was cleaned, dried, and gently ground. The carbonate fraction was dissolved by a gradual addition of HCl (concentrated at 37 wt%) on the dried powder. The mixture was then neutralized (pH = 7) by flushing with deionized water. Digestion of the remaining silicate fraction was carried out by rinsing with an equal parts mixture of distilled water and HF concentrated at 50 wt% for 48 h. Kerogen isolates plus any acid-resistant minerals were neutralized by the addition of deionized water and preserved in a H_2O-HCl solution. After the removal of part of the H_2O-HCl mixture, three drops of OM solid residue were extracted and diluted in 25 ml of distilled water. Lastly, a portion of the obtained mixed solution was deposited and dried on thin section slides for Raman spectroscopy analyses.

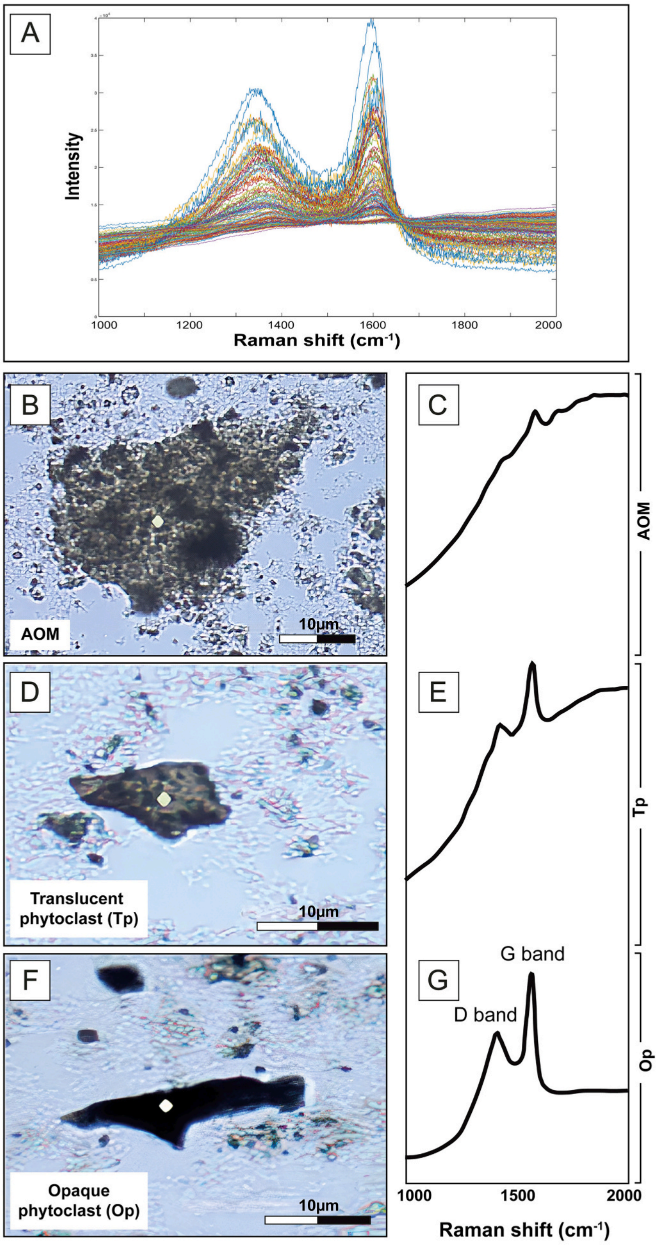

Fig. 3. Spectra classification under Transmitted light petrographic observation. (A) Normalized Raman spectra of the measured organofacies: AOM, Tp and Op; (B) AOM in transmitted light; (C) Typical Raman spectra of AOM; (D) Tp in transmitted light; (E) Typical Raman spectra of Tp; (F) Op in transmitted light; (E) Typical Raman spectra of Op.

3.2. Raman spectroscopy

Raman spectroscopy was performed on isolated kerogen using a Jobin Yvon micro-Raman LabRam system in a backscattering geometry. Data were gathered using a 600 grooves/mm spectrometer grating and CCD detector in the first order Raman spectrum range (700–2300 cm^{-1}). A Neodimium-Yag green laser (532 nm) was used as the light source and the laser power was adjusted by optical filters (i.e., 1% of the optical laser power) to <0.4 mW. An integration time of 25 s and 2 repetitions for each measurement was used to record Raman backscattering. This setup, coupled with the use of a green laser source and optical filters, allowed lessening the fluorescence background.

Raman spectra of OM is defined by a first order region (1000–1800 cm^{-1}) and a second order region (2400–3500 cm^{-1}). In low maturity OM bands in the second order region are weak, thus, they were not detected. The first order Raman region is composed by two main peaks: the disordered (D) band at 1350 cm^{-1} and the graphite (G) band at 1585 cm^{-1} (Tuinstra and Koenig, 1970). The G-band is the only band occurring in graphite in the first order region and has been interpreted as the in-plane vibration of the carbon atoms in the graphite sheets. The D-band appears in disordered amorphous OM and has been interpreted as: (i) the effect of a double-resonant Raman scattering process (Pócsik et al., 1998; Thomsen and Reich, 2000; Reich and Thomsen, 2004; Pimenta et al., 2007); (ii) the ring breathing vibration in the graphene sub-unit (PAHs) (Castiglioni et al., 2001; Di Donato et al., 2004; Negri et al., 2004; Lünsdorf, 2016; Rebelo et al., 2016); (iii) aromatics with six rings or more (Li et al., 2006). In dispersed OM additional bands appear depending on the degree of the coalification; however, the number, origin and nomenclature of these bands are still under debate (for a complete review refer to Henry et al., 2019).

3.3. Raman spectroscopy and transmitted-light petrography

Raman spectroscopy analyses were performed on the selected 27 samples and 1363 spectra were obtained (40 to 60 spectra per sample). Optical observations were performed under the built-in transmitted light device using magnification objectives at 20×, 50× and 100×. Raman analyses were performed at 50×, with a numerical aperture of 0.75, to avoid differences derived from the laser spot size (2 μm). A focal length of 80 cm was used for the monochromator unit with a slit of 200 μm and a confocal hole of 200 μm.

Spike removal and normalization relatively to the mean intensities of the Raman spectra were performed as pre-processing steps before Principal Component Analysis (PCA) and clustering analysis. Given the high dimensionality of the spectral dataset, in which each observation is composed by a 1021 rows per 2 columns matrix, a reduction of the feature space (i.e., independent variables, in this case the intensities row) is required to improve the performance of ML algorithms (Bellman, 1957). In this work, the spectral dataset is described by a n_i x m_i matrix, where the n_i rows represent the number of measurements and the m_i columns represent the feature space defined as the frequency of Raman spectra (here, 1021 for each spectrum). PCA was performed using the intensities of each spectrum (y-axis of the matrix) since the Raman shift (cm^{-1}) values (x-axis) are the same in all spectra. PCA is a conventional dimensionality reduction technique that projects a multivariate (or high-dimensional) dataset into a lower-dimensional coordinate system that captures the maximum amount of variation in the dataset, ensuring a minimum information loss (Jolliffe, 1986; Duda et al., 2001). Once the whole dataset has been reduced, a representative number of components is selected to proceed with the clustering analysis. For a more detailed explanation of PCA, see Supplementary Materials, Text 1.

3.4. Unsupervised Clustering methods

Clustering, also known as cluster analysis, aims to identify natural groupings or clusters within multidimensional data based on a chosen measure of "similarity" (i.e., Euclidean distance, probability distribution, etc.), so that data items within a cluster are very similar to each other but dissimilar to data items in other groups (Jain et al., 2000; Grira et al., 2005). Clustering is usually performed when no information is available about the belonging of data items to predefined classes (i.e., non-labeled data) (Jain and Dubes, 1988) and, thus, is traditionally considered as a type of unsupervised learning technique (Ghahramani, 2003; Grira et al., 2005). In this study, three different types of clustering methods (i.e., partitioning, probabilistic-based and density-based) were applied to the acquired dataset through different clustering algorithms: k-means, Gaussian Mixture Models (GMM) and Density-Based Spatial Clustering of Application with Noise (DBSCAN).

The different cluster algorithms were applied to the spectral dataset. It should be noted that cluster analysis was performed grouping together

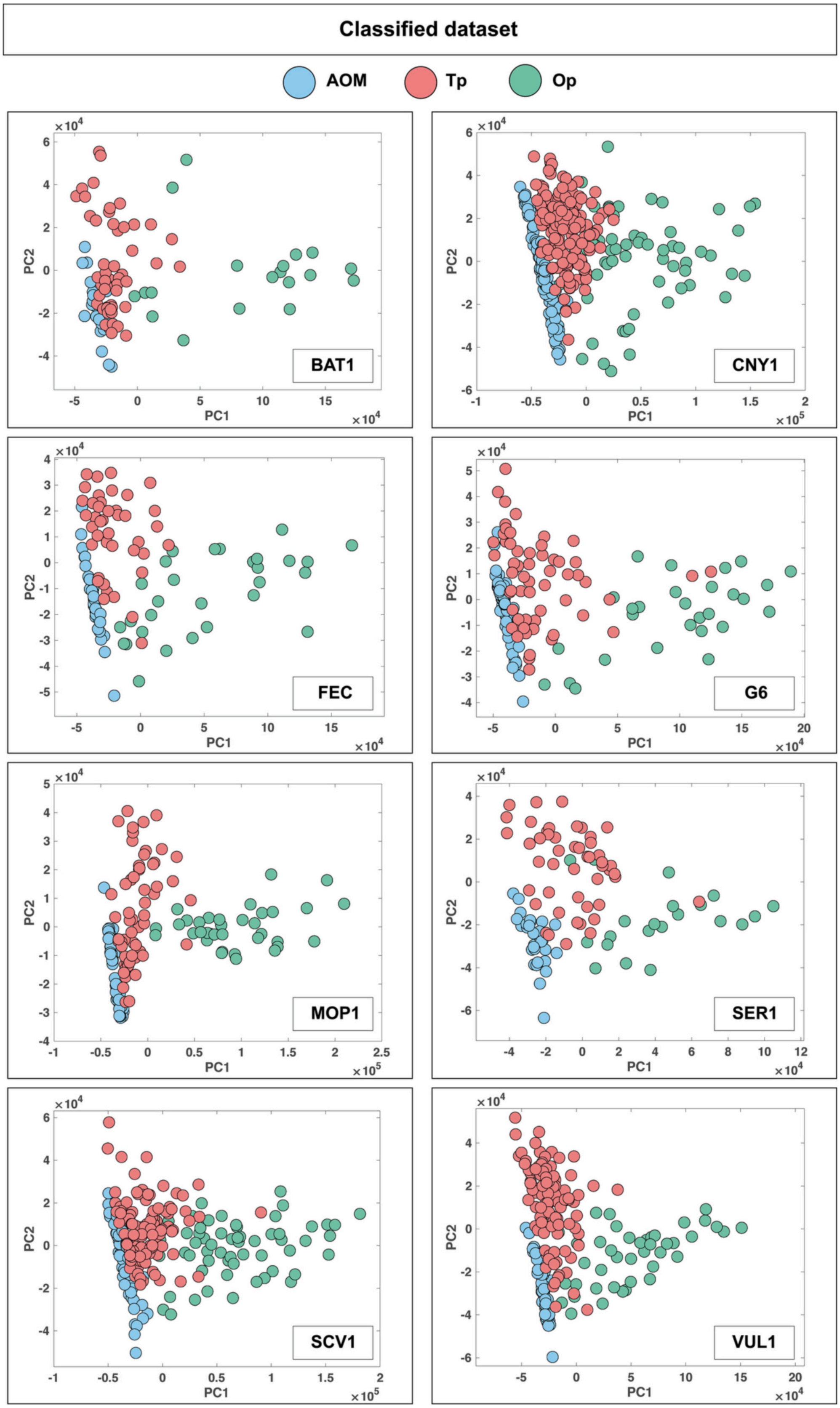

Fig. 4. PCA and optical classification of the measured Raman spectra for each well. AOM = amorphous organic matter (blue circles); Tp = translucid phytoclasts (red circles); Op = Opaque phytoclasts (green circles). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

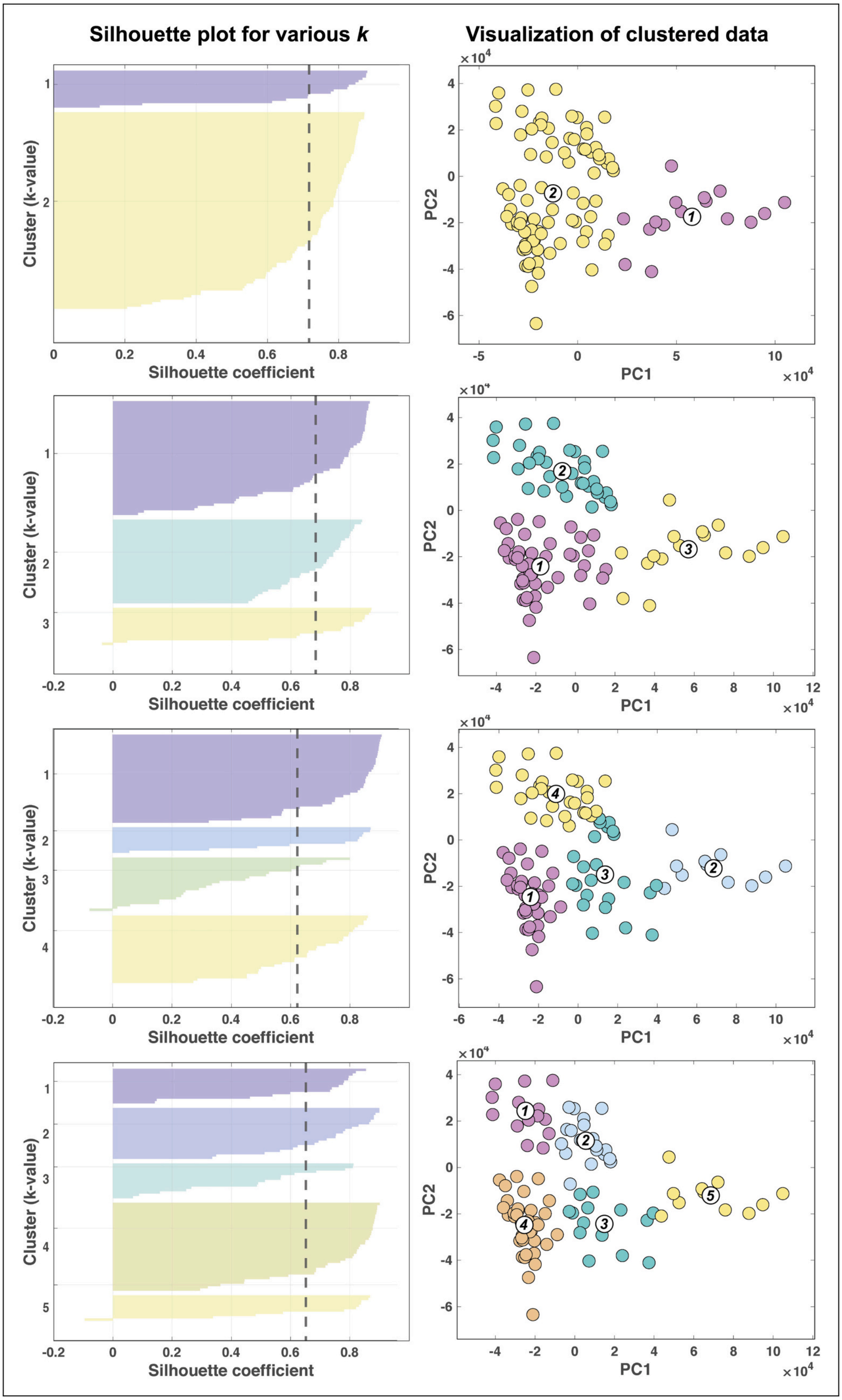

Fig. 5. Silhouette plots and relative cluster outputs for different *k*-values. Left column: Silhouette plots for 2 to 5 clusters. The dashed line represents the average Silhouette score for each *k*-value. Right column: visualization of clustering according to each *k*-value.

Table 2
Silhouette coefficients calculated for various *k*-values (1 to 7) for the studied wells.

Well	Average silhouette coefficient vs number of clusters (*k*)					
	k = 2	*k* = 3	*k* = 4	*k* = 5	*k* = 6	*k* = 7
BAT1	0.90	0.68	0.65	0.61	0.51	0.51
CHY1	0.76	0.54	0.58	0.59	0.57	0.56
FEC	0.81	0.64	0.66	0.62	0.64	0.64
G6	0.83	0.71	0.60	0.59	0.60	0.58
MOP1	0.79	0.76	0.69	0.73	0.67	0.66
SER1	0.72	0.68	0.62	0.65	0.65	0.62
STCV1	0.73	0.69	0.56	0.57	0.55	0.50
VUL1	0.76	0.70	0.70	0.63	0.60	0.60

samples for each well rather than for each sample to increase the amount of data. This can be done since samples derived from the same stratigraphic interval (i.e., same depths) and thus thermal maturity of the samples within a single well is almost the same (Table 1).

3.4.1. Partitioning methods: k-means and Silhouette Score

k-means is an iterative, centroid-based algorithm designed to partition data items within a dataset into a specified number of k clusters (Eq. 3 in Supp. Mat. 1; Omran et al., 2007; Nasraoui and N'Cir, 2019). Each cluster is represented by its center (i.e., centroid), which corresponds to the arithmetic mean of data points assigned to the cluster. k-means iteratively recalculates the position of the cluster's centroids based on the distance (here, Euclidean) between a data item and the cluster centroid until cluster assignments do not change, or the maximum number of iterations (i.e., 1000) is reached. (Han et al., 2012; Aggarwal and Reddy, 2014). The resulting *k* clusters are, then, as compact and as separate as possible. Given that centroids are calculated using the arithmetic mean, it forms spherical-like shaped clusters.

k-means requires the user to select the desired *k* number of clusters. However, there are methods to evaluate the best *k* value in unsupervised scenarios. One statistical method to assess the optimal *k*-value consists in the calculation of the Silhouette Score (Eq. 4 in Supp. Mat. 1; Rousseeuw, 1987; Anitha and Patil, 2019; Arima et al., 2008). Silhouette Score values close to 1 indicate perfect clustering with the highest compactness and well separated clusters, while values close to −1 indicate bad clustering. Values near to zero denote overlapping.

k-means algorithm is a "crisp" or "hard" clustering, since it performs hard assignments of points to clusters (e.g., assuming equal variance for objects within a cluster) in contrast to fuzzy or "soft" methods where each data item is defined by a degree of membership to every cluster center (Han et al., 2012).

3.4.2. Probabilistic model-based methods: Gaussian Mixture Model (GMM) and BIC

Probabilistic model-based clustering methods are based on finite mixtures and provide an example of a "soft" clustering approach. In a mixture, clusters are represented by a set of probability distributions (e.g., Gaussian distributions) Everitt et al., 1981; Titterington et al., 1985; McLachlan and Basford, 1988; McLachlan and Peel, 2000; Frühwirth-Schnatter and Frèuhwirth-Schnatter, 2006). A GMM can be represented as the weighted sum of a number of Gaussian component densities (i.e., the number of clusters) (Eq. 6 in Supp. Mat. 1). Each Gaussian in the mixture (i.e., a cluster) is characterized by its mean, covariance, and weight (Eq. 8–10 in Supp. Mat. 1). The goal of GMM is to iteratively find the best combination of these parameters to define a number of clusters. GMM is particularly useful since it allows to cluster complicated geometrical shapes by combining the number of clusters (*K*) and the covariance matrix (Σ_j) which defines cluster's geometry (shape and orientation). Both, the optimal (*K*) and Σ_j, are calculated by applying the Bayesian Information Criterion (BIC, Schwarz, 1978). The covariance matrix can be: 1) A full covariance matrix that can be independently oriented in any direction; 2) a diagonal matrix type where the clusters are forced to be oriented along the coordinate axes. Both type of covariant matrices can have an unshared structure where each Gaussian distribution can independently differ in size and orientation, or a shared structure where all distributions are of the same size and orientation (Murphy, 2012). For a detailed explanation of GMM and BIC, refer to Supp. Mat. 1.

3.4.3. Density-based methods: Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Clusters can be defined as dense regions in the data space, separated by sparser regions. The density of an object can be measured by the number of objects close to it in a region of the data space defined as a neighborhood (Tan et al., 2013). The goal of DBSCAN (Density-Based Spatial Clustering of Applications with Noise; Ester et al., 1996) is to find objects characterized by dense neighborhoods. DBSCAN quantifies the density of an object by defining the radius (Eps) of a neighborhood for every object and counting the number of objects within that Eps (Tan et al., 2013; Schubert et al., 2017). In addition, DBSCAN determine whether a neighborhood is dense or not by imposing a user-specified density threshold (MinPts). Objects in the dataspace which contains at least Minpts objects within the defined Eps radius are assigned as *core points,* while objects positioned at the edge of a neighborhood delimited by Eps are assigned as *border points*. Following these definitions, clusters are formed by connecting core points and their neighborhoods to form dense regions as clusters separated by areas of lower density (Ester et al., 1996; Han et al., 2012; Schubert et al., 2017) DBSCAN also allows the identification of outliers or noise. Noise points are any data item that is neither a *core point* nor a *border point* (Ester et al., 1996). For a detailed explanation of DBSCAN, refer to Supp. Mat. 1.

3.5. Performance metrics

After a clustering algorithm is applied to a set of data items, a fundamental step is to quantitively compare the obtained partitions with the ground truth (i.e., dataset classified by means of optical observations) to establish its robustness. Given the prior knowledge of the ground truth, extrinsic or supervised evaluation metrics (evaluation indexes) were used to evaluate the clustering quality by comparison with the known labels. To evaluate the goodness of clustering a *contingency matrix* must be built. This matrix contains four terms: *True Positives* (TP), *False Positives* (FP), *False Negatives* (FN) and *True Negatives* (TN) (Palacio-Niño and Berzal, 2019). After the calculation of the contingency matrix, three different evaluation indexes were used: Random Index (RI), Adjusted Random Index (ARI), F-measure. For details on the different evaluation methods used in this work see Supp., Mat. 1.

3.6. Raman spectra deconvolution

Deconvolution enables the assessment of various Raman parameters (e.g., peak position, width, area and intensity) for individual bands (Beyssac et al., 2002; Rahl et al., 2005; Lahfid et al., 2010; Hinrichs et al., 2014; Wilkins et al., 2014; Chen et al., 2017; Sauerer et al., 2017; Schopf et al., 2005; Li et al., 2006; Guedes et al., 2010; Bonoldi et al., 2016; Ferralis et al., 2016; Schito et al., 2017; Schito and Corrado, 2020). Raman deconvolution uses a two-Lorentzian-bands fitting performed by a modified version of the PeakFit program designed by O'Haver (2015). This simplified method allows to evaluate basic Raman parameters avoiding errors derived from a complex multiple fitting. A quadratic baseline was subtracted to correct high fluorescing spectra with control points between 750 and 850 cm^{-1} and between 1880 and 1920 cm^{-1}.

3.7. General workflow

The flowchart presented in Fig. 2 illustrates the workflow from preprocessing of the raw Raman spectra and clustering till clustering

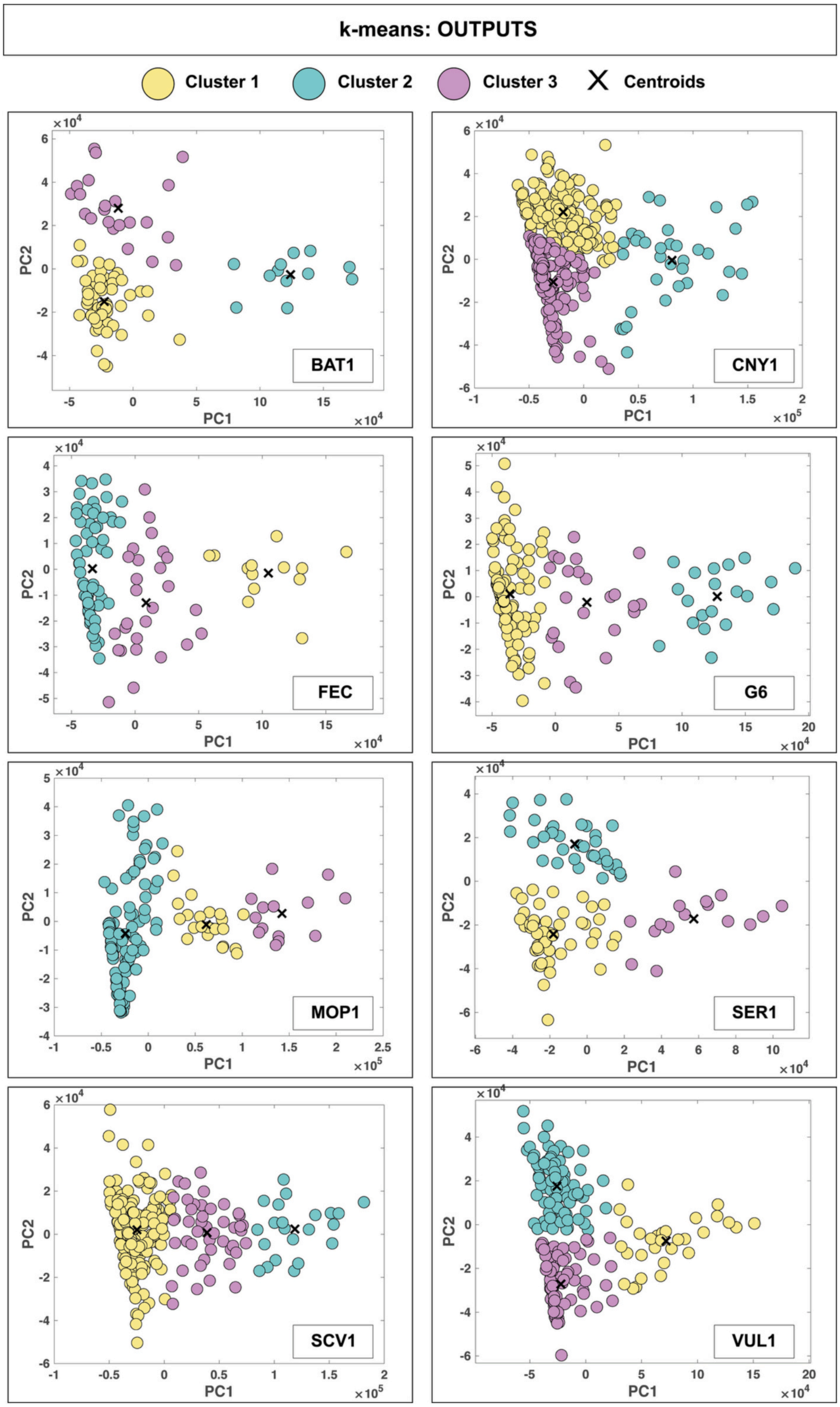

Fig. 6. k-means algorithm outputs for the studied wells. The three different identified classes are illustrated in yellow (cluster 1), green (cluster 2) and purple (cluster 3), while "x" indicates cluster centroid.

evaluation. PCA routine, cluster analysis and spectra deconvolution were implemented with MATLAB™ statistic tool.

4. Results

4.1. Organofacies identification and Raman spectra classification

While performing Raman analyses, organic particles were classified under transmitted light observations based on petrographic information (e.g., texture, color, and shape of particles) following Tyson (1995)

Table 3
BIC values calculated for different *K* and covariance matrix for each studied well.

Well	Number of mixture components (K)						Best covariance matrix ($\sum_j$)
	K = 2	*K* = 3	*K* = 4	*K* = 5	*K* = 6	*K* = 7	
BAT1	4309.31	4293.82	4305.75	4330.49	4300.75	4315.94	Full-unshared
CHY1	14,897.11	14,754.05	14,759.71	14,781.28	14,790.32	14,790.62	Full-unshared
FEC	4731.91	4661.17	4676.52	4680.69	4697.66	4706.57	Full-unshared
G6	6895.85	6809.58	6808.86	6826.34	6836.32	6864.72	Full-unshared
MOP1	6692.00	6582.70	6597.50	6598.60	6611.50	6627.50	Full-unshared
SER1	4676.18	4633.63	4643.00	4659.93	4679.53	4654.90	Full-unshared
STCV1	10,402.01	10,383.89	10,400.58	10,422.90	10,449.27	10,472.17	Full-unshared
VUL1	9659.42	9513.51	9517.75	9530.20	9544.35	9564.29	Full-unshared

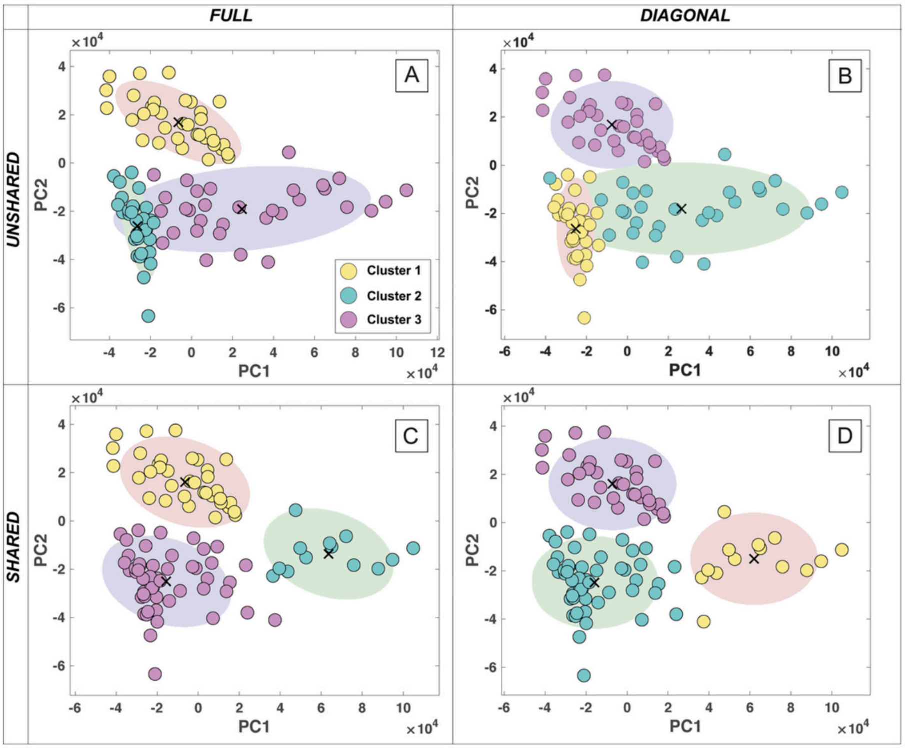

Fig. 7. Different covariance matrix types (shared, unshared) and structures (full, diagonal) with their relative outputs. (A) Full-unshared covariance matrix. Ellipses may be independently rotated with respect to the main axes and can independently assume any position and shape; (B) Diagonal-unshared covariance matrix. The major and minor axes of the ellipses are parallel or perpendicular to the x and y axes, but clusters can independently assume any shape and position; (C) Full-shared covariance matrix. Ellipses may be independently rotated with respect to the main axes, but all ellipses are the same size and have the same orientation; (D) Diagonal-shared covariance matrix. Ellipses admit no rotation and share size and orientation.

indications. Classification was later used to evaluate the quality and performance of the unsupervised clustering. Three categories were recognized (Fig. 3 B-G): (i) AOM (Fig. 3 B); (ii) translucent phytoclasts (Tp; Fig. 3 D); and (iii) opaque phytoclasts (Op; Fig. 3 F). Optical analyses point out a predominance of AOM in the 27 kerogen samples. AOM is light yellow to brown colored and characterized by irregular shapes. Most of the AOM Raman spectra are overwhelmed by fluorescence, while some of them are characterized by a wide G-band and an incipient D-band (Fig. 3 C). Phytoclasts are abundant in most of the samples. Tp are transparent, filter-like orange to dark brown fragments and characterized by an irregular to polygon-shaped (generally elongated) appearance. Their corresponding Raman spectra show a high to moderate fluorescent background, a large and prominent G-band and often a well-developed D-band (Fig. 3 E). Op are opaque dark brown to black particles, usually characterized by lath-shaped fragments with sharp angular outlines. Their Raman spectra display the lowest fluorescence with respect to AOM and Tp and are characterized by well-defined G and D-bands (Fig. 3 G).

Based on these observations, all 1363 spectra were classified into one of the three different categories.

4.2. Raman spectra pre-processing and PCA

The obtained spectra were spike-cleaned and normalized prior to cluster analysis. Additionally, a dimensionality reduction through PCA was carried out reducing the n-dimensional dataset (1021 independent variables) to a 2-dimensional dataset. The first 2 two principal components (PC) explain between the 96% and the 98% of the variance in the original matrix (refer to Supp. Mat. 2). The scatter plot of the PC1 against the PC2 (score plot) is a useful tool to evaluate the distribution of a set of data. Fig. 4 shows the score plot of the entire dataset in the PC space pointing out that data items naturally distribute forming a series of clusters (Fig. 4). By coloring each data item according to their optical classification, it can be observed that similar organofacies tend to cluster with a general trend AOM-Tp-Op from low to high PC1 values (Fig. 4).

In the AOM cluster (blue dots in Fig. 4), data mainly spread along PC2 with a tight spread along PC1. The Tp cluster shows the highest variability along the PC2, whereas the spread along PC1 increases at larger PC2 values. Op's cluster displays the largest spread along both axes.

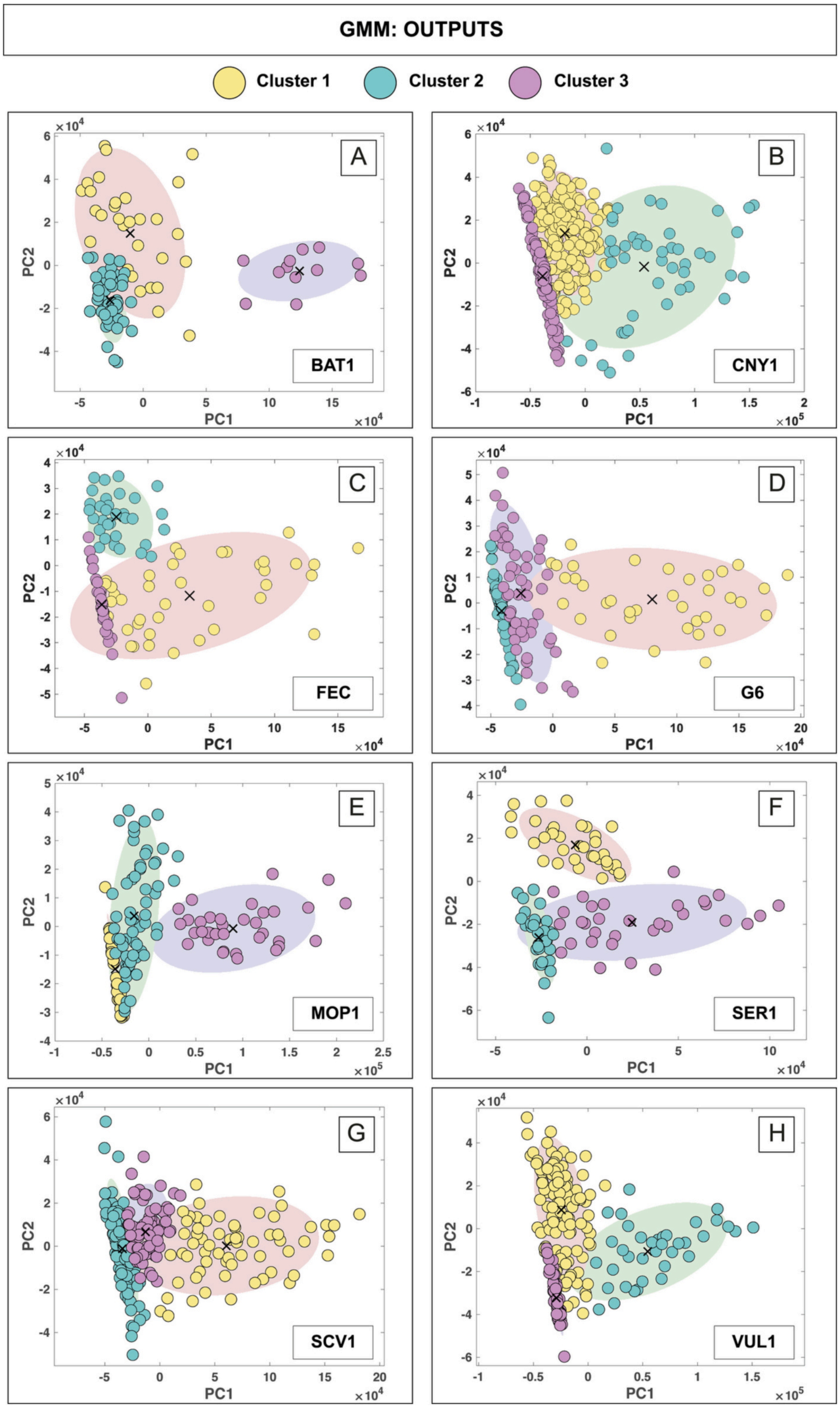

Fig. 8. GMM clustering outputs for the eight different analyzed wells. Ellipses show the density contours given by the covariance matrix type and structure.

4.3. *Cluster analysis*

4.3.1. *k-Means clustering*

The main limit of k-means algorithm is that it requires the user to define the number of clusters (*k* value) to fit the objective function (Eq. 3 in Supp. Mat. 1, please read this section for further information on this method). In this study, the *k* value could have been arbitrarily set to match the number of observed classes (i.e., Tp, Op and AOM categories; Fig. 4). However, in order to test a fully unsupervised scenario, a set of *k* values was evaluated by analyzing the *silhouette score*. Given a range of *k* values ($k = 1 \ldots n$), a silhouette plot (Fig. 5) measures the similarity of points within a cluster by using a silhouette coefficient and the size of the silhouette plot for each value of *k*. In general, an optimal *k* value should be characterized by the combination of ahigh silhouette score and by homogeneous clusters in thickness and size that can be evaluated trough visual inspection of the silhouette plots.

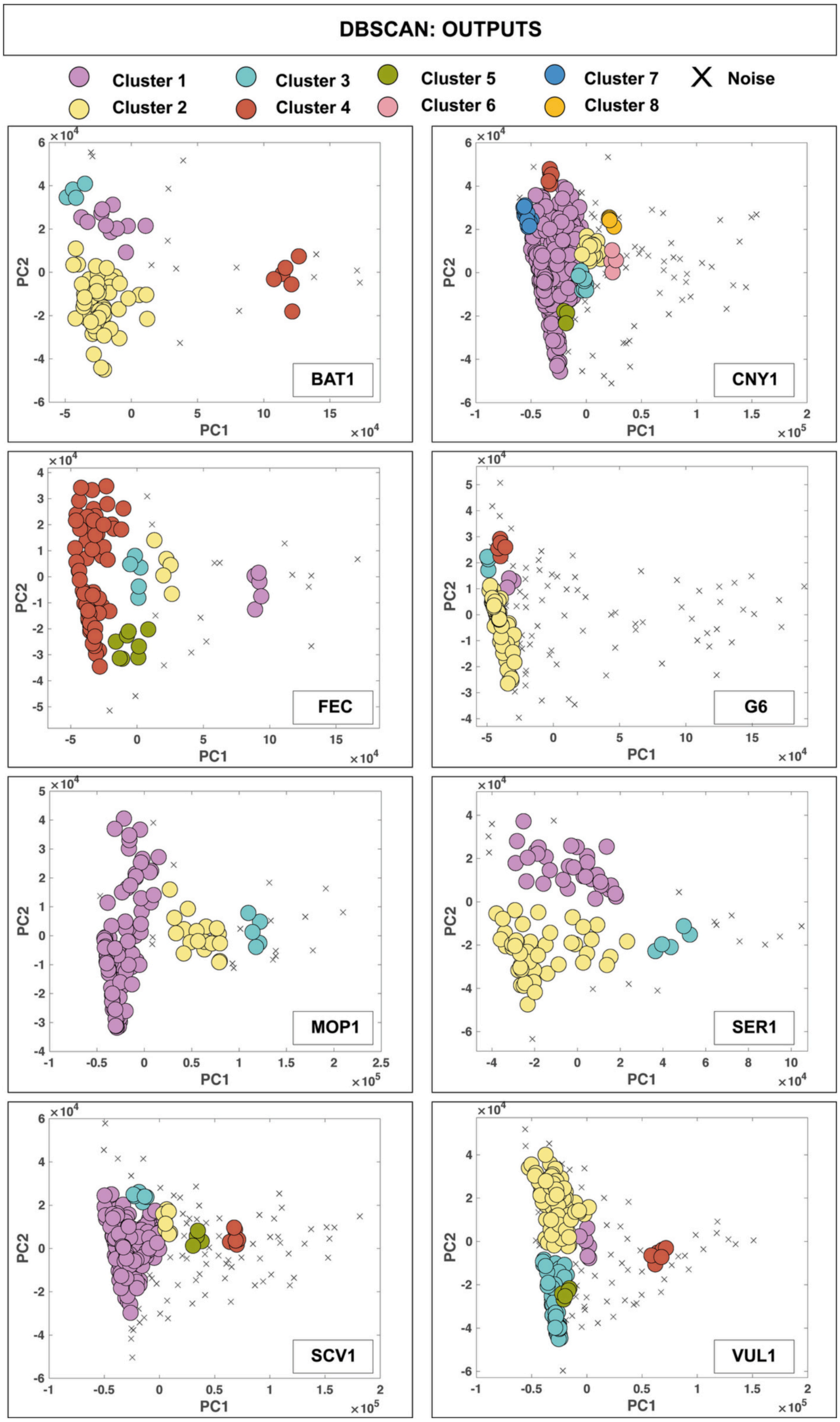

Fig. 9. DBSCAN algorithm outputs for the studied wells. The number of clusters vary according to the initial parameters Eps and MinPts.

We analyzed the *silhouette score* for a series of values of k from 2 to 7 since this is the number where average *silhouette score* reach a plateau (Table 2). As shown in Table 2, average silhouette scores going from 2-to7-*k* generally tend to decrease with the highest average silhouette scores corresponding to 2 and 3-cluster partitions.

A visual inspection of silhouette plots (see example from SER1 well in Fig. 5) outlines that the 2-*k* option generates irregular partitions, with a prominent and thick main cluster, and a secondary narrow one, the 3- to 4-*k* partitions generate clusters with similar thickness and hence similar sizes and the 5-*k* partition generates irregular clusters in thickness and size. Given that, 3-cluster partitions show higher average silhouette scores (Table 2) with the most regular partition, the silhouette analysis suggests that the optimal *k*-value for all wells is 3 (Supp. Mat. 3).

Fig. 6 displays the outputs of the k-mean clustering using $k = 3$. This

Table 4
DBSCAN parameters calculated for each well.

Well	MinPts	Eps	[1]Total Points	[2]Noise Points	[3]Number of Clusters
BAT1	4	1.36E+04	93	14	4
CHY1	4	5.25E+03	324	59	8
FEC	4	1.22E+04	104	17	5
G6	4	4.16E+03	157	75	4
MOP1	4	1.12E+04	149	15	3
SER1	4	1.20E+04	101	16	3
SCV1	4	5.60E+03	227	72	5
VUL1	4	6.07E+03	210	53	5

[1] Total number of spectra for the given well.
[2] Number of spectra considered as noise by DBSCAN for each well.
[3] Number of clusters calculated by DBSCAN for a given Eps and MinPts value.

figure shows that the position of the centroid for the three clusters locates at increasing PC1 values but at similar PC2 values for FEC, G6, MOP1 and SCV1 wells, while in BAT1, CNY1, SER1 and VUL1 wells the first and second centroids are aligned at about the same PC1 value.

4.3.2. *GMM clustering*

In Gaussian Mixture Models both the number and geometry of clusters can vary and require the user to define them through two input arguments: the number of components (clusters - K) and the covariance matrix type and structure (Σ_j). Choosing an optimal number of components (K) is critical to prevent overfitting or underfitting, while the covariance matrix constrains the geometric features (i.e., volume, shape, and orientation) of the clusters. Here, GMM were fitted using all possible covariance types (full and diagonal) and structure (shared and unshared) for a number of K values (Table 3). BIC (see Sup. Mat. 1 for more information) was applied for a number of components from 2 to 7 selecting the best values for both inputs, K and Σ_j. Table 3 shows the outputs of the BIC application to a mixture of 2 to 7 K components in terms of BIC scores and type and structure of the covariance matrix for the eight analyzed wells. The covariance evaluation through the BIC indicates that the optimal Σ_j for $2 \geq K \geq 7$ is always represented by a full-unshared matrix (Table 3, Supp. Mat. 4), which allows clusters to independently assume different shapes and sizes and to be rotated with respect to the coordinate axes (i.e., PC1 and PC2) as shown in the example in Fig. 7.

The best K for the dataset is denoted by the lowest BIC value, which represents either a better fit or a lower number of parameters. From Table 3, it is possible to observe that the best number of components for all the studied wells corresponds to $K = 3$ (Table 3, Supp. Mat. 4), except for well G6. In this case, BIC scores are very similar so that $K = 4$ score is almost the same than the one for $K = 3$ (Table 3). In this case the choice of the less complex model (K = 3) is advisable.

A full display of the GMM outputs for the eight analyzed wells are shown in Fig. 8. Comparing with k-means results it can be observed that (i) GMM always depicts the narrower (higher density) clusters towards the lowest PC1 values; (ii) a second less compact cluster is always present and (iii) the third cluster moving towards the highest PC1 values, always shows the lowest density of points (i.e., high variability).

4.3.3. *DBSCAN*

DBSCAN requires two user-specified parameters to initialize: *Eps* and *MinPts* (see methods section and Supp. Mat. 1). *Eps* is a parameter that depends on the distance (here, Euclidean) between data items. Hence, if the chosen *Eps* value is too small, a large part of the data will not be clustered and considered as noise. In contrast, for a too high value of *Eps*, clusters will merge and most of the items will fall in the same cluster. The *MinPts* parameter represents, instead, the minimum number of points to form a cluster. Accordingly, *MinPts* should be neither too large to allow small clusters to be generated, nor too small, in order to reduce the influence of noise. Here, as suggested in Ester et al. (1996), *MinPts* can be set to 4 for all 2-dimensional datasets. The *Eps* parameter was determined using the k-dist plot (see Supp. Mat. 1, 5).

The obtained outputs for all the studied samples are illustrated in Fig. 9. In Table 4, the input parameters and the obtained number of clusters for each sample are shown. Fig. 9 shows that the number of clusters is generally higher than expected (i.e., 3 cluster; see Fig. 4), going from a minimum of three (MOP1 and SER1) up to eight clusters (CNY1; Table 4). Moreover, data items mostly cluster towards the negative PC1 values, while moving towards the positive PC1 values, most of the data items are considered as noise ($\sim$10 to $\sim$50% of the data) or generate small-sized clusters (Fig. 9 and Table 4).

Table 5
Number of predicted versus observed classes and the calculated Random index (RI) for k-means, GMM and DBSCAN algorithms.

Well	Clustering algorithm	Translucent phytoclasts			AOM			Opaque phytoclasts			Total RI
		observed	*predicted*	*RI*	*observed*	*predicted*	*RI*	*observed*	*predicted*	*RI*	
BAT1	k-Means	53	22	0.62	21	59	0.59	19	12	0.92	0.71
	GMM	53	30	0.58	21	51	0.66	19	12	0.92	0.72
	DBSCAN	–	–	–	–	–	–	–	–	–	–
CHY1	k-Means	155	151	0.75	105	139	0.79	64	34	0.91	0.82
	GMM	155	180	0.90	105	97	0.97	64	47	0.94	0.94
	DBSCAN	–	–	–	–	–	–	–	–	–	–
FEC	k-Means	41	27	0.54	34	64	0.69	29	13	0.85	0.69
	GMM	41	32	0.89	34	32	0.98	29	40	0.89	0.92
	DBSCAN	–	–	–	–	–	–	–	–	–	–
G6	k-Means	60	23	0.65	68	114	0.71	29	21	0.92	0.76
	GMM	60	52	0.83	68	67	0.94	29	38	0.89	0.89
	DBSCAN	–	–	–	–	–	–	–	–	–	–
MOP1	k-Means	59	22	0.50	52	109	0.61	36	16	0.86	0.66
	GMM	59	64	0.94	52	47	0.96	36	36	0.97	0.96
	DBSCAN	59	19	0.46	52	105	0.58	36	6	0.88	[*]0.641
SER1	k-Means	49	36	0.83	31	49	0.82	21	16	0.93	0.86
	GMM	49	36	0.83	31	33	0.94	21	32	0.85	0.87
	DBSCAN	49	32	0.81	31	48	0.79	21	5	0.84	[*]0.843
SCV1	k-Means	111	160	0.67	58	22	0.65	58	45	0.84	0.72
	GMM	111	72	0.78	58	92	0.83	58	63	0.95	0.86
	DBSCAN	–	–	–	–	–	–	–	–	–	–
VUL1	k-Means	104	88	0.89	63	91	0.85	43	31	0.93	0.89
	GMM	104	117	0.92	63	55	0.96	43	38	0.96	0.95
	DBSCAN	–	–	–	–	–	–	–	–	–	–

[*] RI calculated for the only DBSCAN wells with a number of clusters equal to the ground truth number of clusters (i.e., 3).

Table 6
ARI, F_2 measure, Precision (Pr) and Recall (R) calculated for each well for the different tested clustering algorithms.

External Cluster Validation					
Well	Clustering Algorithm	ARI	Pr	R	F_2-measure
BAT1	k-Means	0.17	0.76	0.67	0.62
	GMM	0.23	0.71	0.67	0.45
	DBSCAN	–	–	–	–
CHY1	k-Means	0.32	0.79	0.70	0.70
	GMM	0.73	0.93	0.87	0.95
	DBSCAN	–	–	–	–
FEC	k-Means	0.32	0.63	0.55	0.53
	GMM	0.69	0.90	0.90	0.78
	DBSCAN	–	–	–	–
G6	k-Means	0.33	0.71	0.63	0.62
	GMM	0.61	0.81	0.84	0.73
	DBSCAN	–	–	–	–
MOP1	k-Means	0.34	0.55	0.50	0.45
	GMM	0.82	0.95	0.94	0.94
	DBSCAN	0.35	0.55	0.44	0.41
SER1	k-Means	0.44	0.84	0.80	0.79
	GMM	0.52	0.81	0.84	0.73
	DBSCAN	0.34	0.85	0.71	0.71
SCV1	k-Means	0.32	0.45	0.48	0.47
	GMM	0.30	0.80	0.84	0.64
	DBSCAN	–	–	–	–
VUL1	k-Means	0.55	0.86	0.82	0.82
	GMM	0.76	0.94	0.90	0.96
	DBSCAN	–	–	–	–

4.4. *Clustering evaluation*

The quality of the clustering can be evaluated by comparing the unsupervised clustering with the ground truth (i.e., organofacies classification). It should be noted that this can only be done when the clustering algorithm is able to identify the number of clusters matching the ground truth classes (i.e., 3). This is the case for k-means and GMM methods, but only for the wells MOP1 and SER1 when working with DBSCAN. In these cases, the three clusters are generally distributed along then PC1 axis resembling the real distribution shown in Fig. 4. Thus, moving from low to high PC1 values, the first cluster is associated with the AOM group, the second with the Tp group and the third with the Op group.

Three different extrinsic metrics (RI, ARI and F2-measure) were used to compare clustering predictions against the ground truth (for details, see Supp. Mat. 1). For these indices, the value 1 indicates high quality of the partition against the ground truth and values closer or lower than 0 indicates no agreement. RI (Eq. 14 in Supp. Mat. 1) values have been estimated for each one of the observed classes (Tp, AOM and Op) against the classes predicted by k-means, GMM and DBSCAN methods. The average RI calculated for each sample is shown as Total RI in Table 5. GMM and k-means show average RI values above 0.5, meaning that >50% of the predicted data items were assigned to their true class.

However, GMM show an overall better performance with an average RI above 0.7 for all the analyzed samples. Additionally, GMM reveal high RI values when predicting Tp (RI between 0.58 and 0.94) and AOM (RI > 0.8). While both GMM and k-means show a good performance, when classifying Op data items, with RI values included between 0.66 and 0.97 and 0.59–0.86.

DBSCAN shows a high performance for SER1 and MOP1 wells, even though the majority of the Op are recognized as noise (see Fig. 9). Since noise points are not classified, they are not included in the RI calculation. This, along with a significant number of misclassifications (False Positives and False Negatives; for details refer to Supp. Mat. 1), accounts for the high RI values for the Op class in SER1 and MOP1 wells.

RI computes the similarity between two clusters by counting pairs of data items that are assigned in the same or in different clusters in the predicted and true datasets. Nevertheless, RI does not consider the possibility of agreement by chance between two clustering (Morey and Agresti, 1984; Santos and Embrechts, 2009). To take account of "randomness" in the dataset, the ARI (Eq. 15 in Supp. Mat. 1) was computed for all the studied samples (Table 6). ARI is equal to 0 for random classification (independently of the number of clusters) and is equal to 1, when clustering perfectly matches observations.

As shown in Table 6, calculated ARI scores are lower than RI values. This is particularly evident for k-means clustering, where ARI values are <0.5 (except for VUL1 sample) with a minimum value of 0.17. On the other hand, GMM show ARI values >0.5 for most of the studied samples, with a minimum ARI value of 0.23 for BAT1 well. DBSCAN shows ARI values included between 0.35 (MOP1) and 0.34 (SER1).

Both RI and ARI give equal weight to false positives (FP) and false negatives (FN). In classification problems, assigning two similar data items to different clusters (i.e., FN or type-II error) is sometimes worse than allocating pairs of dissimilar data items within the same cluster (i.e., FP or type-I error). Here, F_2-measure (Table 6; Eq. 16–18 in Supp. Mat. 1) was used to incorporate the influence of FN's by applying a weight $\beta = 2.0$. This modification increases the influence of Recall (Eq. 18 in Supp. Mat. 1). GMM show the best performance with F_2-measure values >0.5 for most of the samples (apart from BAT1), and with values >0.9 for CHY1, MOP1 and VUL1. In fact, precision (Pr, Eq. 17 in Supp. Mat. 1), which indicates the fraction of correct positive predictions (TP), is >0.8. This means that when the model makes a prediction for a certain class, it is correct 80% of the times. Recall indicates which proportion of actual positives (TP + FN) was correctly identified. Here, $R > 0.8$ for most of the samples, indicating that the model correctly identifies the 80% of the observed classes. These results suggest that the influence of FN's is minor, and so the resulting F_2-measure is still high, indicating an optimal prediction performance.

According to ARI and F_2-measure metrics, the overall best performance is achieved when GMM clustering is applied.

5. Discussion

5.1. *Choosing the best clustering algorithm*

When evaluating clustering performance between different algorithms, it is critical to evaluate their capacity to distinguish and accurately represent the cluster characteristics in terms of data distribution, shape, size, density, number of clusters, among others (Tan et al., 2013). Here, we defined three critical aspects the algorithms must meet to accurately represent the studied dataset: number of clusters, cluster shape (e.g., globular, elliptical, rectangular) and orientation (parallel or rotated to the coordinate axis), and cluster distribution (i.e., dispersion or density of data items in the PC space).

Fig. 4 suggests that organofacies in the studied samples (i.e., AOM, Tp, and Op) generally cluster with an elliptical shape, with varying length and wideness ratio and rotated with respect to the main axis (i.e., PC1 and PC2). In terms of cluster distribution, AOM (Fig. 4, in blue color) appears as a compact and narrow cluster dispersed mostly along the PC2. Tp (Fig. 4, in red color) shows varying densities and is dispersed mostly along the PC2, mimicking the AOM distribution, tough with a relatively large spread along the PC1. Op (Fig. 4, in green color) appears as a low-density cluster with a major spread along the PC1 and a scattered variation along the PC2. This means that the challenge of the algorithm is to define three elliptic clusters, mostly overlapping, rotated with respect to the score plot axes and with different densities.

DBSCAN fails when establishing the number of clusters, since its determination strongly relies on the selection of the initial parameters (MinPts and Eps). These parameters, combined, are a function of density and determine whether a cluster can be defined: if the minimum number of points (MinPts) meets the minimum distance between two points (Eps), then such points are considered as a cluster. One of the drawbacks of this density-based approach is that DBSCAN does not work well for datasets with varying densities or datasets with very sparse data. If Eps threshold is low enough, areas characterized by both higher and lower

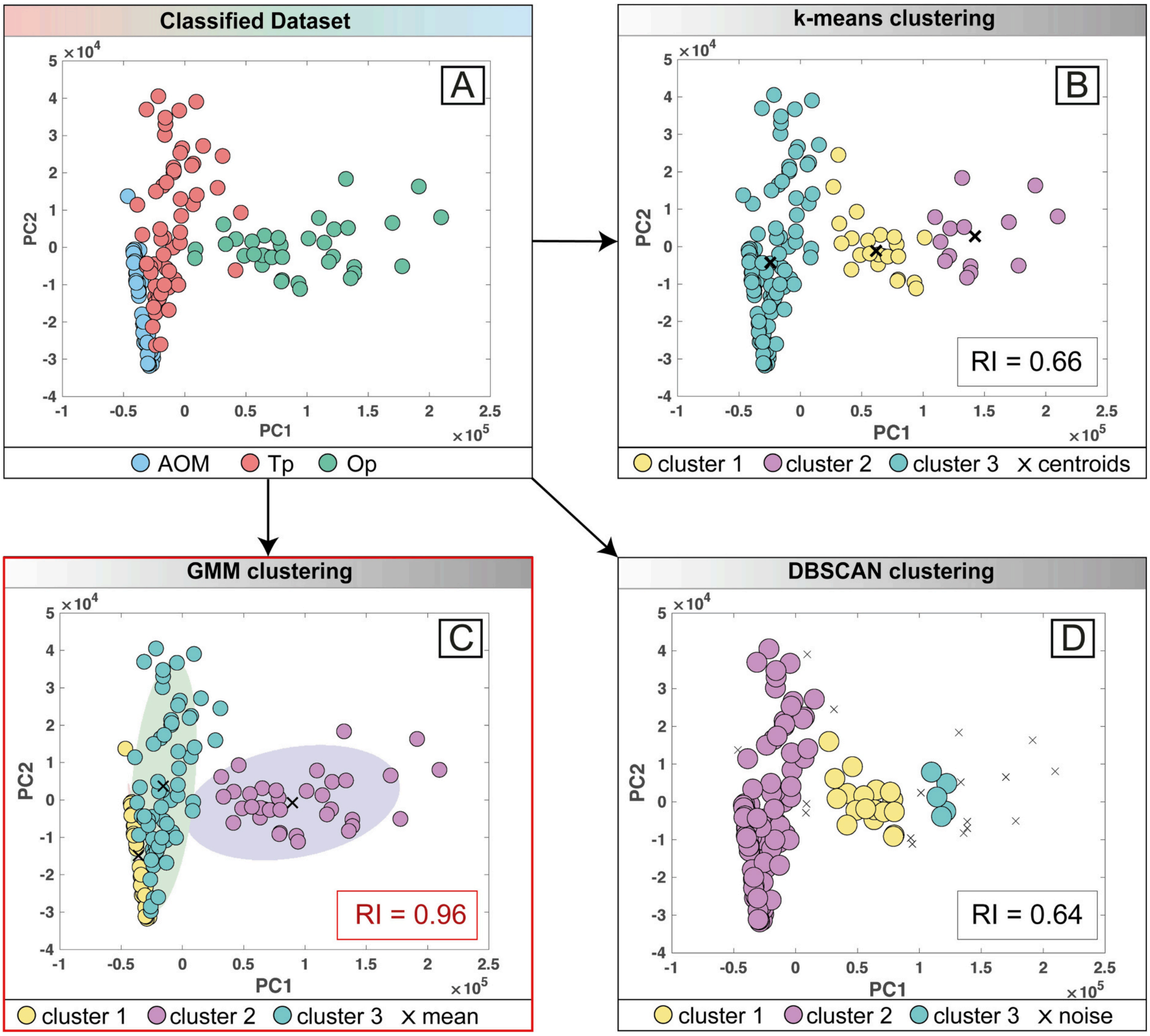

Fig. 10. Example of the clustering outputs for the tested algorithms (k-means, GMM and DBSCAN) in comparison with the classified dataset applied to the well MOP1. The boxes contain the Random Index (RI) calculated for each one of the tested algorithms. (A) Manually classified dataset (ground truth) illustrating the classified organofacies into AOM, Tp and Op; (B) k-means clustering output; (C) GMM clustering output. Ellipses show the density; (D) DBSCAN clustering output.

densities will become a single cluster, while if the Eps threshold is high enough most of the points will be marked as noise. As shown in Fig. 4, the studied dataset is characterized by clusters with varying densities going from a highly compact AOM cluster to a sparse Op cluster. This implies that a MinPts-Eps combination cannot then be chosen appropriately for all clusters, and so the intrinsic cluster structure cannot be characterized by global density parameters. For this reason, nor the shape or cluster distribution could be correctly detected (example in Fig. 10 D).

k-means can determine three partitions through the silhouette analysis, correctly matching the number of classes given by the classified dataset. However, in terms of cluster shape, the fixed distance norm (i.e., Euclidean) imposes a fixed globular geometrical shape of the clusters regardless of the actual data distribution. Consequently, k-means generally fails to predict the distribution of the AOM class and tends to mix AOM and Tp classes (Fig. 6, example in Fig. 10 B), since these two classes slightly overlap k-means algorithm is a “crisp” or “hard” clustering that does not allow overlapping partitions since each data item is assigned to exactly one cluster. In contrast, in fuzzy or “soft” methods (e.g., GMM) the assignment of a data item to a cluster relies on its degree of membership to the cluster center (Han et al., 2012).

GMM algorithm better establishes both the number and shape of clusters for the studied dataset (Fig. 8, example in Fig. 10 C). In fact, the use of a full-shared covariance matrix allows for ellipsoid-shaped clusters rotated with respect to the main axis. This turns to be particularly useful in accurately representing the AOM distribution and the dispersion of the Op class (Fig. 8, example in Fig. 10 C).

Accordingly, mixture models based on Gaussian distributions represent the best clustering algorithm for the studied dataset. This algorithm shows a high performance due: to i) the use of a Bayesian criteria to forecast the expected number of clusters ii) the ability of a full-unshared covariance matrix to assume different shapes and orientations, and to create overlapping partitions and iii) the ability to identify classes with different densities. GMM clustering was able to correctly recognize the natural distribution of the Raman spectra of different organic components in a reduced feature space, with an accuracy >70% for most of the samples and an accuracy >80% for Tp which are the target organofacies for thermal maturity assessment.

5.2. *Correlations against thermal maturity: R_o% and T_{max}*

A set of correlations between R_o% and T_{max} (Table 1) previously measured on the same samples analyzed in this study (Corrado et al., 2022; Vergara Sassarini, 2022) and one of the most widely used Raman

Table 7
Raman parameter (ΔD-G) calculated for only Tp and all classes together.

Well	Sample	[1]ΔD-G (Tp)			[2]ΔD-G (all)		
		ΔD-G	*s.d ΔD-G*	[3]*n*	*ΔD-G*	*s.d ΔD-G*	*n*
BAT1	bat1_1	231.60	7.72	9	235.32	8.01	38
	bat1_2	234.46	7.77	19	233.33	10.96	36
CNY1	cny1_1	234.20	6.79	22	230.95	16.12	32
	cny1_2	232.57	4.25	27	232.79	9.80	36
	cny1_3	242.07	6.61	42	241.42	11.92	61
	cny1_10	235.47	6.52	31	239.38	9.57	40
	cny1_12	232.83	5.72	29	233.19	11.05	40
	cny1_14	235.49	6.87	22	235.60	14.19	33
FEC	fec_1	225.49	7.12	12	229.31	15.83	33
	fec_2	220.64	10.76	7	230.42	15.33	19
	fec_3	221.03	6.36	8	229.26	13.93	19
G6	g6_1	225.56	7.71	16	228.66	20.90	34
	g6_2	220.41	13.17	16	227.52	15.78	31
	g6_3	222.99	19.92	8	228.82	16.96	19
MOP1	mop1_1	233.37	4.47	19	234.72	12.32	35
	mop1_2	234.40	9.02	17	238.13	11.05	30
	mop1_3	235.07	5.35	15	235.07	13.91	30
SER1	ser1_1	239.05	5.20	17	242.20	7.69	46
	ser1_2	237.33	5.09	19	240.71	9.78	40
STCV1	stcv1_1	233.90	6.46	13	239.39	13.02	44
	stcv1_2	239.05	6.77	22	240.75	10.97	45
	stcv1_3	236.28	7.15	19	239.37	12.58	54
	stcv1_4	237.56	5.52	15	234.21	10.68	42
VUL1	vul1_1	234.71	8.72	29	235.05	15.43	40
	vul1_2	229.95	7.90	29	228.39	16.07	42
	vul1_3	234.12	5.94	22	234.37	14.57	40
	vul1_4	230.16	7.49	26	230.34	14.90	47

[1] Distance between D and G peaks (ΔD-G) for translucid phytoclasts (Tp).

[2] ΔD-G for the three different spectral classes: Tp, AOM and opaque phytoclasts (Op).

[3] Number of spectra used for deconvolution and calculation of ΔD-G parameter. Due to the high degree of fluorescence in some AOM spectra, deconvolution was not possible and hence some spectra were not considered for correlations.

parameters, ΔD-G (i.e., distance between D and G peaks), were obtained with the aim of verifying the effect of GMM prediction on the degree of correlation and standard deviation (Table 7 and Fig. 11).

Table 7 shows the ΔD-G derived from GMM predictions for Tp class and for the three predicted classes combined (AOM + Tp + Op), while Fig. 11 shows the correlation and relative standard deviation before (Fig. 11 B and D) and after (Fig. 11 A and C) GMM prediction. Fig. 11 A-B shows ΔD-G correlate against R_o% with a an R-square (R^2) value of 0.685 (*p*-value <0.00002) for Tp (Fig. 11 A) and 0.491 (p-value <0.00007) for all classes (Fig. 11 B), while regression for ΔD-G against T_{max} (Fig. 11 C, D) return an R^2 value of 0.504 (p-value <0.00003) for Tp (Fig. 11 A) and 0.345 (p-value <0.0001) for all classes (Fig. 11 B). R^2 values are not particularly high when compared with previous works (Schmidt et al., 2017; Schito et al., 2017; Li et al., 2020) and this can be attributed to the very narrow thermal maturity interval here considered (0.43–0.66%Ro). However, of particular interest is the observed increase in the degree of correlation (R^2) after the implementation of GMM that changes from 0.49 to 0.68 in Fig. 11 (A and B) and from 0.34 to 0.50 in Fig. 11 (C and D). The increase in the degree of correlation is accompanied by a significant reduction in the standard deviation, highlighting the crucial role played by GMM in minimizing uncertainties in thermal maturity assessment, particularly in the context of complex organofacies.

5.3. *Towards an automatic identification of organic facies and future applications*

One of the main limitations on the application of Raman spectroscopy to organic petrographic studies is to combine the view enhancement offered by oil-immersion microscopy and the quality of Raman spectra (i.e., avoid interferences from oil). Alternatively, transmitted light observations have shown to be a viable option (Henry et al., 2018: Schito et al., 2019, this study) although they still require acid attack concentration of kerogen for quality observations. In addition, petrographic studies in transmitted light are extremely time-consuming and require a skilled operator to be reliable and accurate.

Raman analyses are quick and do not require a particular sample preparation, yet the quality of the observations is still dependent on the operator skill and experience. In this regard, finding an automatic method to minimize the operator intervention could enhance the potential of Raman spectroscopy to be applied as a routinely fast method by the industry or for big-data analysis as in the case of planetary sciences. In this context, a machine learning approach would be a possible solution to overcome such limits.

The successful use of different ML approaches (e.g., pattern recognition, neural networks, image analysis processing, semantic segmentation) based on the analysis of a collection of petrographic images (oil-immersion) of coal samples for the characterization and identification of different maceral groups (i.e., vitrinite, inertinite and liptinite) and minerals in coals (Mlynarczuk and Skiba, 2017; Skiba and Młynarczuk, 2018; Wang et al., 2019, 2022; Lei et al., 2021) demonstrates the great potential of this approach as a tool to automate the qualitative and quantitative characterization of the complex and heterogeneous structure of hard coal.However, the success of such methodologies is bound to the existence of a good quality, large, classified image training set (i.e., ground truth) required to source the learning algorithms. Besides, methodologies based solely on image analysis does not furnish any information about the chemical properties of carbonaceous materials (i.e., thermal maturity). A supervised ML approach to Raman interpretation has been recently applied by Schito et al. (2021) that worked on polished petrographic plugs and highlighted how the use of the air-immersion objective usually assembled on Raman apparatus allowed maceral discrimination only for a small portion of the spectral dataset. However, the implementation of a supervised machine learning routine based on multivariate statistical analysis (PLS-DA) allowed these authors to classify almost the entire dataset, increasing the number of recognized vitrinite fragments used for thermal maturity assessment. Still, this previous work was based on the knowledge of the ground truth for a portion of the analyzed dataset.

In the present work, we are making a step further, showing that the potential of an unsupervised approach is that ground truth knowledge is not required for classification, relying solely on the ability of the chosen algorithm to represent real data and on the initial tuning of the selected algorithm parameters (i.e., number of clusters, covariance matrix, Eps and MinPts). Here, the application of a probabilistic-based clustering (GMM) on the whole dataset of Raman spectra allowed to assign data items within a cluster showing how the relative position and geometry of the clusters on the score plot (shape and distribution) might give an indication of the type of organic material. Our results confirm previous observations from Schito et al. (2021) suggesting that, on the score plot, the path of increasing aromatization can be followed moving on the PC1 (in this study AOM always lie at the lowest values, while Op at the highest). Moreover, our results show how cluster density is clearly largest for the Op group, which is generally composed by material with different origins (i.e., charring, detritic OM, graphitic carbon etc.). In this regard, the GMM method could be also applied in studies concerning specific macerals group as already accomplished through image analysis within the inertinite group (Mlynarczuk and Skiba, 2017; Skiba and Młynarczuk, 2018).

While it is true that in this study we analyzed concentrated OM that required a time-consuming preparation, future research may explore the application of the GMM method to polished pellets used in organic petrography under reflected light, or even to unpolished rock chips, with the goal of optimizing and streamlining the analytical process. Moreover, considering the resemblance of the aims of previous image analysis-based works and the present study, it is suggested that the combined use of ML applied to image analysis and unsupervised ML on

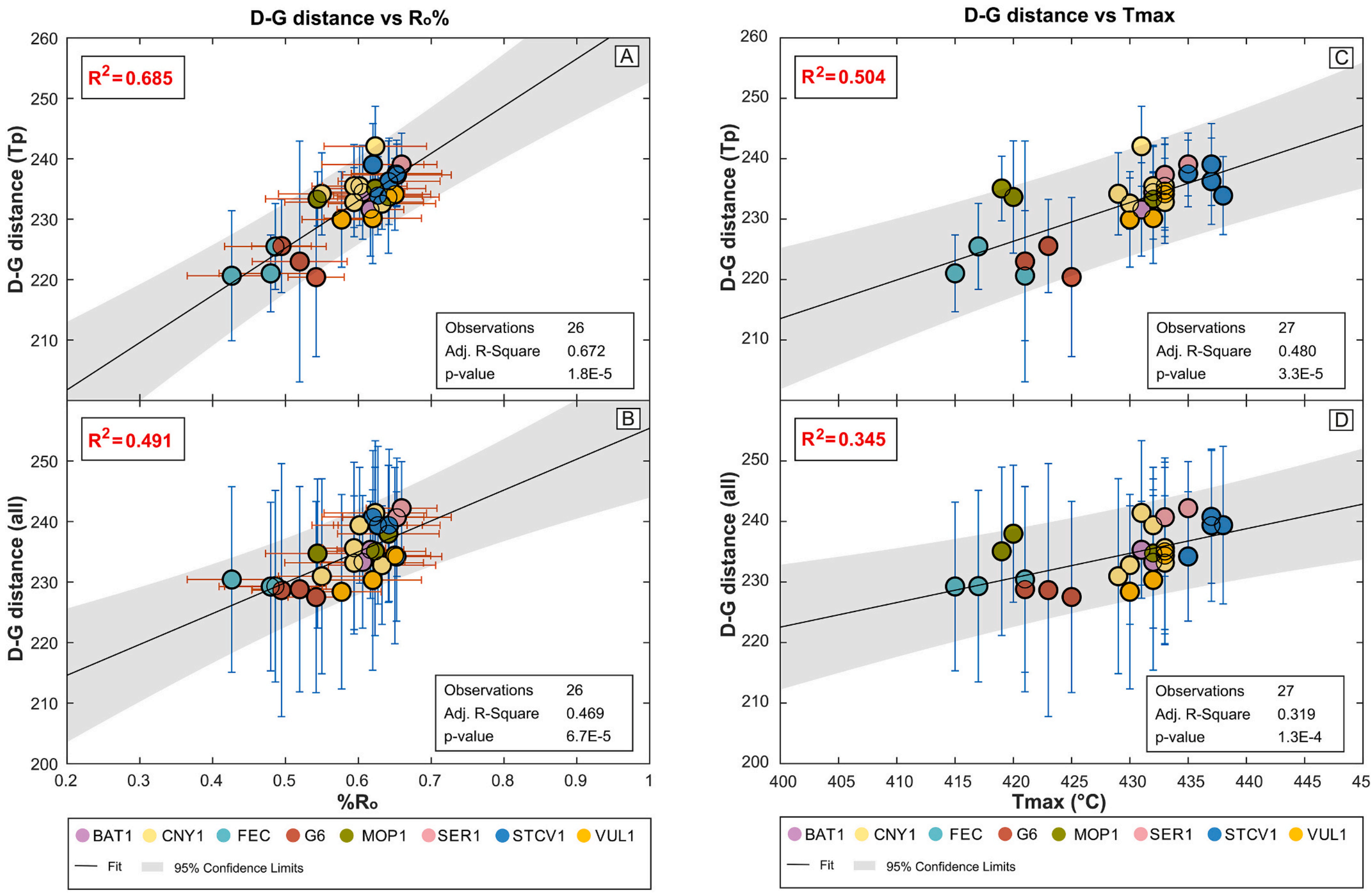

Fig. 11. Scatter plot and linear prediction with 95% confidence interval of ΔD-G versus vitrinite reflectance (R_o%) and T_{max}. (A, C) Correlations derived from deconvolution of Tp spectra, illustrating the general trend of parameter evolution with increasing R_o% and T_{max}, respectively. (B, D) Correlations derived from deconvolution of all spectra together (AOM, Tp and Op) spectra, illustrating the general trend of parameter evolution with increasing R_o% and T_{max}, respectively. R_o% for sample VUL1_1 was not available (Table 1) and so it was not considered for the correlations with vitrinite (figure continued on next page).

Raman spectra could represent a robust and reliable method towards the automation of OM characterization for thermal maturity assessment.

Compared to other widely used spectroscopic techniques for the characterization of organic matter, such as the FTIR (D'Angelo et al., 2010; D'Angelo et al., 2011; Bao et al., 2018), in which the material recognition depends on the analysis of many absorption bands and so easier to be detected on a score plot, Raman spectroscopy relies on the analysis of only the D-and-G-bands. The existence of a clustering method capable of detecting the variation of organofacies on a score plot, as proven in the present work, is therefore of high value and confirms the suitability of Raman spectroscopy in organic petrographic studies.

6. Conclusions

In this work transmitted light petrographic observations and Raman spectroscopy analyses coupled with three different unsupervised learning clustering algorithms (k-means, GMM and DBSCAN) were performed on samples collected from Schistes Carton" interval of the Paris Basin. Vitrinite reflectance and Rock-Eval pyrolysis analyses collocate the selected samples in the immature interval to the onset of the oil-window stage (0.45 to 0.65 R_o%; 415 °C to 438 °C T_{max}).

Cluster analysis was performed on Raman spectra principal components on samples from the same stratigraphic interval (Toarcian "Schistes Carton"). k-means, GMM and DBSCAN algorithms were evaluated in terms of: (1) parameterization, (2) their ability to accurately predict the number of clusters, cluster shape and orientation and cluster density; (3) their accuracy in making predictions.

k-means requires a visual examination of the silhouette plots to determine the optimal number of clusters making its parameterization an operator-dependent task. This algorithm, in addition, fails to predict cluster shape accurately, especially the AOM cluster, compromising the prediction performance.

DBSCAN is highly sensitive to initial parameters, and so its ability to predict number, shape, size and orientation of clusters strongly relies on its parameterization. Furthermore, this algorithm fails when applied to datasets with varying densities.

GMM clustering shows the best performance in terms of ability to:

(1) automatically determining the number of clusters along with the cluster shape/orientation and distribution by means of the usage of the BIC, significantly reducing the operator intervention for the selection of the initial parameters.
(2) successfully recognizing and classifying the different organic facies based on their Raman spectra with an accuracy >70% for most of the wells. In detail, GMM was able to identify the Tp class with an accuracy >80%, excluding one well.

Correlations between the ΔD-G Raman parameter calculated for the predicted classes and T_{max} and R_o% illustrate how the implementation of GMM clustering for organic facies classification reduces the uncertainties on the calculated Raman parameters, improving the quality of the correlations aimed to the assessment of thermal maturity at low diagenetic stages.

This work outlines how the use of Raman spectroscopy coupled with unsupervised clustering algorithms may facilitate and quicken the characterization of highly heterogeneous dispersed OM by maximizing the variability among different Raman spectra. Thus, the implementation of automated methodologies in the study of complex organic materials has the potential to simplify and speed up the required petrographic studies for industrial applications or exploration purposes, particularly in scenarios where large datasets necessitate prompt screening.

CRediT authorship contribution statement

Natalia A. Vergara Sassarini: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft. **Andrea Schito:** Conceptualization, Methodology, Supervision, Software, Investigation, Writing - original draft, Writing - review & editing. **Marta Gasparrini:** Writing – review & editing, Project administration, Supervision, Investigation. **Pauline Michel:** Writing – review & editing, Formal analysis. **Sveva Corrado:** Conceptualization, Supervision, Project administration, Formal analysis, Investigation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

IFP Energies nouvelles (France) is warmly acknowledgment for kindly providing access to samples, laboratory facilities and unpublished database. Dr. Amalia Spina and Prof. Simonetta Cirilli from the University of Perugia are warmly acknowledged for the high-quality kerogen isolate extraction. This research was funded by: MIUR grants to Roma Tre PhD School in Earth Sciences (XXXIV doctoral cycle, 2018–2021) and IFP Energies nouvelles PhD program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.coal.2023.104237.

References

Aggarwal, C.C., Reddy, C.K., 2014. Data Clustering. Algorithms and applications. In: Data mining and Knowledge Discovery series. Chapman & Hall/CRC, London (652 pp).

Aldega, L., Bigi, S., Carminati, E., Trippetta, F., Corrado, S., Kavoosi, M.A., 2018. The Zagros fold-and-thrust belt in the Fars province (Iran): II. Thermal evolution. Marine and Petroleum Geology 93, 376–390. https://doi.org/10.1016/j.marpetgeo.2018.01.005.

Al-Hajeri, M., Sauerer, B., Furmann, A., Amer, A., Akbar, H., Abdallah, W., 2020. Maturity estimation for Type II-S kerogen using Raman spectroscopy–A case study from the Najmah and Makhul Formations in Kuwait. Int. J. Coal Geol. 217, 103317 https://doi.org/10.1016/j.coal.2019.103317.

Al-Hajeri, M., Sauerer, B., Furmann, A., Amer, A., Al-Khamiss, A., Abdallah, W., 2021. Organic petrography and geochemistry of the prolific source rocks from the Jurassic Najmah and Cretaceous Makhul Formations in Kuwait–Validation and expansion of Raman spectroscopic thermal maturity applications. Int. J. Coal Geol. 236, 103654 https://doi.org/10.1016/j.coal.2020.103654.

Allen, P.A., Allen, J.R., 2013. Basin Analysis: Principles and Application to Petroleum Play Assessment. John Wiley & Sons, Chichester, West Sussex, UK.

Anitha, P., Patil, M.M., 2019. RFM model for customer purchase behavior using k-Means algorithm. Journal of King Saud University-Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2019.12.011.

Arima, C., Hakamada, K., Okamoto, M., Hanai, T., 2008. Modified fuzzy gap statistic for estimating preferable number of clusters in fuzzy k-means clustering. J. Biosci. Bioeng. 105 (3), 273–281. https://doi.org/10.1263/jbb.105.273.

Atouabat, A., Corrado, S., Schito, A., Haissen, F., Gimeno-Vives, O., Mohn, G., Frizon de Lamotte, D., 2020. Validating Structural Styles in the Flysch Basin Northern Rif (Morocco) by Means of thermal Modeling. Geosciences 10 (9), 325. https://doi.org/10.3390/geosciences10090325.

Balestra, M., Corrado, S., Aldega, L., Morticelli, M.G., Sulli, A., Rudkiewicz, J.L., Sassi, W., 2019. Thermal and structural modeling of the Scillato wedge-top basin source-to-sink system: Insights into the Sicilian fold-and-thrust belt evolution (Italy). Bulletin 131 (11–12), 1763–1782. https://doi.org/10.1130/B35078.1.

Bao, R., Luan, X., Wu, S., Yu, X., Zheng, L., 2018. Detecting Marine Kerogen from Western Canada Basin using Terahertz Spectroscopy. ACS omega 3 (7), 7798–7802. https://doi.org/10.1021/acsomega.8b00791.

Bellman, R., 1957. Dynamic Programming. Princeton Univ. Press, Princeton, New Jersey, US.

Beny-Bassez, C., Rouzaud, J.N., 1985. Characterization of carbonaceous materials by correlated electron and optical microscopy and Raman microspectroscopy. Scan Electron. Microsc. 1, 119–132.

Beyssac, O., Goffé, B., Chopin, C., Rouzaud, J.N., 2002. Raman spectra of carbonaceous material in metasediments: a new geothermometer. J. Metamorph. Geol. 20 (9), 859–871. https://doi.org/10.1046/j.1525-1314.2002.00408.x.

Bhartia, R., Beegle, L.W., DeFlores, L., Abbey, W., Hollis, J.R., Uckert, K., Monacelli, B., Edgett, K.S., Kennedy, M.R., Sylvia, M., Aldrich, D., Anderson, M., Asher, S.A., Bailey, Z., Boyd, K., Burton, A.S., Caffrey, M., Calaway, M.J., Calvet, R., Cameron, B., Caplinger, M.A., Carrier, B.L., Chen, N., Chen, A., Clark, M.J., Clegg, S., Conrad, P.G., Cooper, M., Davis, K.N., Ehlmann, B., Facto, L., Fries, M.D., Garrison, D.H., Gasway, D., Ghaemi, F.T., Graff, T.G., Hand, K.P., Harris, K., Hein, J.D., Heinz, N., Herzog, H., Hochberg, E., Houck, A., Hug, W.F., Jensen, E.H., Kah, L.C., Kennedy, J., Krylo, R., Lam, J., Lindeman, M., McGlown, J., Michel, J., Miller, E., Mills, Z., Minitti, M.E., Mok, F., Moore, J., Nealson, K.H., Nelson, A., Newell, R., Nixon, B.E., Nordman, D.E., Nuding, D., Orellana, S., Pauken, M., Peterson, G., Pollock, R., Quinn, H., Quinto, C., Ravine, M.A., Reid, R.D., Riendeau, J., Ross, A.J., Sackos, J., Schaffner, J.A., Schwochert, M., Shelton, O., Simon, R., Smith, C.L., Sobron, P., Steadman, K., Steele, A., Thiessen, D., Tran, V.D., Tsai, T., Tuite, M., Tung, E., Wehbe, R., Weinberg, R., Weiner, R.H., Wiens, R.C., Williford, K., Wollonciej, C., Wu, Y.H., Yingst, R.A., Zan, J., 2021. Perseverance's Scanning Habitable Environments with Raman and Luminescence for Organics and Chemicals (SHERLOC) investigation. Space Sci. Rev. 217, 58. https://doi.org/10.1007/s11214-021-00812-z.

Bonoldi, L., Di Paolo, L., Flego, C., 2016. Vibrational spectroscopy assessment of kerogen maturity in organic-rich source rocks. Vib. Spectrosc. 87, 14–19. https://doi.org/10.1016/j.vibspec.2016.08.014.

Brunet, M.F., Le Pichon, X., 1982. Subsidence of the Paris basin. Journal of Geophysical Research: Solid Earth 87 (B10), 8547–8560. https://doi.org/10.1029/JB087iB10p08547.

Caricchi, C., Corrado, S., Di Paolo, L., Aldega, L., Grigo, D., 2016. Thermal maturity of Silurian deposits in the Baltic Syneclise (on-shore Polish Baltic Basin): contribution to unconventional resources assessment. Ital. J. Geosci. 135 (3), 383–393. https://doi.org/10.3301/IJG.2015.16.

Castiglioni, C., Negri, F., Rigolio, M., Zerbi, G., 2001. Raman activation in disordered graphites of the A 1′ symmetry forbidden k≠ 0 phonon: the origin of the D line. J. Chem. Phys. 115 (8), 3769–3778. https://doi.org/10.1063/1.1381529.

Chen, S., Wu, D., Liu, G., Sun, R., 2017. Raman spectral characteristics of magmatic-contact metamorphic coals from Huainan Coalfield, China. Spectrochim. Acta A: Molec. Biomolec. Spectrosc. 171, 31–39. https://doi.org/10.1016/j.saa.2016.07.032.

Cheshire, S., Craddock, P.R., Xu, G., Sauerer, B., Pomerantz, A.E., McCormick, D., Abdallah, W., 2017. Assessing thermal maturity beyond the reaches of vitrinite reflectance and Rock-Eval pyrolysis: A case study from the Silurian Qusaiba formation. Int. J. Coal Geol. 180, 29–45. https://doi.org/10.1016/j.coal.2017.07.006.

Corrado, S., Di Bucci, D., Naso, G., Giampaolo, C., Adatte, T., 1998. Application of organic matter and clay mineral studies to the tectonic history of the Abruzzo-Molise-Sannio area, Central Apennines. Italy. Tectonophysics 285 (1–2), 167–181. https://doi.org/10.1016/S0040-1951(97)00195-9.

Corrado, S., Invernizzi, C., Aldega, L., D'errico, M., Di Leo, P., Mazzoli, S., Zattin, M., 2010. Testing the validity of organic and inorganic thermal indicators in different tectonic settings from continental subduction to collision: the case history of the Calabria–Lucania border (southern Apennines, Italy). J. Geol. Soc. 167 (5), 985–999. https://doi.org/10.1144/0016-76492009-137.

Corrado, S., Schito, A., Romano, C., Grigo, D., Poe, B.T., Aldega, L., Caricchi, C., Di Paolo, L., Zattin, M., 2020. An integrated platform for thermal maturity assessment of polyphase, long-lasting sedimentary basins, from classical to brand-new thermal parameters and models: an example from the on-shore Baltic Basin (Poland). Mar. Pet. Geol. 122, 104547 https://doi.org/10.1016/j.marpetgeo.2020.104547.

Corrado, S., Gusmeo, T., Schito, A., Alania, V., Enukidze, O., Conventi, E., Cavazza, W., 2021. Validating far-field deformation styles from the Adjara-Trialeti fold-and-thrust belt to the Greater Caucasus (Georgia) through multi-proxy thermal maturity datasets. Mar. Pet. Geol. 130, 105141 https://doi.org/10.1016/j.marpetgeo.2021.105141.

D'Angelo, J.A., Zodrow, E.L., Camargo, A., 2010. Chemometric study of functional groups in Pennsylvanian gymnosperm plant organs (Sydney Coalfield, Canada): implications for chemotaxonomy and assessment of kerogen formation. Org. Geochem. 41 (12), 1312–1325. https://doi.org/10.1016/j.orggeochem.2010.09.010.

Corrado, S., Vergara Sassarini, N.A., Schito, A., Michel, P., Gasparrini, M., 2022. New integrated geochemical and petrographic constraints to paleo-thermal and paleo-environmental reconstructions from organic matter dispersed in the Early Toarcian organic-rich shales of the Paris Basin (France) [abstract]. In: SGI-SIMP Congress;

2022 Sept 19–21; Torino, Italy. Abstract n. 825. https://doi.org/10.3301/ABSGI.2022.02.

D'Angelo, J.A., Escudero, L.B., Volkheimer, W., Zodrow, E.L., 2011. Chemometric analysis of functional groups in fossil remains of the Dicroidium flora (Cacheuta, Mendoza, Argentina): implications for kerogen formation. Int. J. Coal Geol. 87 (2), 97–111. https://doi.org/10.1016/j.coal.2011.05.005.

Davis, A., 1978. The reflectance of coal. In: Analytical Methods for Coal and Coal Products. Academic Press London, pp. 27–81.

Delmas, J., Houel, P., Vially, R., 2002. Paris Basin. Petroleum Potential - Rapport régional d'évaluation pétrolière. In: IFPEN Intern Report (172 pp).

Di Donato, E., Tommasini, M., Fustella, G., Brambilla, L., Castiglioni, C., Zerbi, G., Simpson, C.D., Müllen, K., Negri, F., 2004. Wavelength-dependent Raman activity of D2h symmetry polycyclic aromatic hydrocarbons in the D-band and acoustic phonon regions. Chem. Phys. 301 (1), 81–93. https://doi.org/10.1016/j.chemphys.2004.02.018.

Di Paolo, L., Aldega, L., Corrado, S., Mastalerz, M., 2012. Maximum burial and unroofing of Mt. Judica recess area in Sicily: implication for the Apenninic–Maghrebian wedge dynamics. Tectonophysics 530, 193–207. https://doi.org/10.1016/j.tecto.2011.12.020.

Disnar, J.R., Le Strat, P., Farjanel, G., Fikri, A., 1996. Organic matter sedimentation in the northeast of the Paris Basin: consequences on the deposition of the lower Toarcian Black Shales. Chem. Geol. 131 (1–4), 15–35.

Dow, W.G., 1977. Kerogen studies and geological interpretations. J. Geochem. Explor. 7, 79–99. https://doi.org/10.1016/0375-6742(77)90078-4.

Duda, R.O., Hart, P.E., Stock, D.G., 2001. Pattern Classification. Wiley-Interscience, Inc., Third Avenue New York, NY, US, 688 p.

Espitalié, J., Marquis, F., Sage, L., Barsony, I., 1987. Géochimie organique du bassin de Paris. Revue de l'institut français du pétrole 42 (3), 271–302.

Espitalié, J., Maxwell, J.R., Chenet, Y., Marquis, F., 1988. Aspects of hydrocarbon migration in the Mesozoic in the Paris Basin as deduced from an organic geochemical survey. Org. Geochem. 13 (1–3), 467–481. https://doi.org/10.1016/0146-6380(88)90068-X.

Ester, M., Kriegel, H.-P. Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. Second Int. Conf. On Knowledge Discovery and Data Mining (KDD-96). AAAI Press, pp. 226–231.

Everitt, B.S., Hand, D.J., Everitt, B.S., Hand, D.J., 1981. Mixtures of normal distributions. Finite Mixture Distributions 25–57. https://doi.org/10.1007/978-94-009-5897-5_2.

Farley, K.A., Williford, K.H., Stack, K.M., Bhartia, R., Chen, A., de la Torre, M., Hand, M., Goreva, Y., Herd, C.D.K., Hueso, R., Liu, Y., Maki, J.N., Martinez, G., Moeller, R.C., Nelessen, A., Newman, C.E., Nunes, D., Ponce, A., Spanovich, N., Willis, P.A., Beegle, L.W., Bell, J., Brown, A.J., Hamran, S.-E., Hurowitz, J.A., Maurice, S., Paige, D.A., Rodriguez-Manfredi, J.A., Schulte, M., Wiens, R.C., 2020. Mars 2020 mission overview. Space Sci. Rev. 216 (8), 1–41. https://doi.org/10.1007/s11214-020-00762-y.

Ferralis, N., Matys, E.D., Knoll, A.H., Hallmann, C., Summons, R.E., 2016. Rapid, direct and non-destructive assessment of fossil organic matter via microRaman spectroscopy. Carbon 108, 440–449. https://doi.org/10.1016/j.carbon.2016.07.039.

Fonseca, C., Mendonça Filho, J.G., Lézin, C., Baudin, F., de Oliveira, A.D., Souza, J.T., Duarte, L.V., 2021. Boosted microbial productivity during the Toarcian Oceanic Anoxic Event in the Paris Basin, France: new evidence from organic geochemistry and petrographic analysis. Geol. Soc. Lond., Spec. Publ. 514 https://doi.org/10.1144/SP514-2020-167.

Frühwirth-Schnatter, S., Frèuhwirth-Schnatter, S., 2006. Finite mixture and Markov switching models, 425. Springer, New York. https://doi.org/10.1007/978-0-387-35768-3.

Gély, J.-P., Hanot, F., 2014. Le Bassin parisien: Un nouveau regard sur la géologie. Association des Géologues du Bassin de Paris Bulletin d'information des geologues du bassin de Paris 9, 229–p.

Ghahramani, Z., 2003. Unsupervised learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (Eds.), Advanced Lectures on Machine Learning, Lecture Notes in Computer Science, vol. v. 3176. Springer, Berlin, Heidelberg, pp. 72–112. https://doi.org/10.1007/978-3-540-28650-9_5.

Gonçalvès, J., Violette, S., Robin, C., Pagel, M., Guillocheau, F., de Marsily, G., Bruel, D., Ledoux, E., 2003. 3-D modelling of salt and heat transport during the 248 my evolution of the Paris basin: diagenetic implications. Bulletin de la Société géologique de France 174 (5), 429–439. https://doi.org/10.2113/174.5.429.

Grira, N., Crucianu, M., Boujemaa, N., 2005. Active semi-supervised fuzzy clustering for image database categorization. In: 4th International Workshop on Content-Based Multimedia Indexing, Riga, Latvia, pp. 1–8. https://doi.org/10.1145/1101826.1101831.

Guedes, A., Valentim, B., Prieto, A.C., Rodrigues, S., Noronha, F., 2010. Micro-Raman spectroscopy of collotelinite, fusinite and macrinite. Int. J. Coal Geol. 83 (4), 415–422. https://doi.org/10.1016/j.coal.2010.06.002.

Guillocheau, F., Robin, C., Allemand, P., Bourquin, S., Brault, N., Dromart, G., Friedenberg, R., Garcia, J.-P., Gaulier, J.-M., Gaumet, F., Grosdoy, B., Hanot, F., Le Strat, P., Mettraux, M., Nalpas, T., Prijac, C., Rigollet, C., Serrano, O., Grandjean, G., 2000. Meso-Cenozoic geodynamic evolution of the Paris Basin: 3D stratigraphic constraints. Geodin. Acta 13, 189–245. https://doi.org/10.1016/S0985-3111(00)00118-2.

Gusmeo, T., Schito, A., Corrado, S., Alania, V., Enukidze, O., Zattin, M., Pace, P., Cavazza, W., 2022. Tectono-thermal evolution of Central Transcaucasia: thermal modelling, seismic interpretation, and low-temperature thermochronology of the eastern Adjara-Trialeti and western Kura sedimentary basins (Georgia). J. Asian Earth Sci. 237, 105355 https://doi.org/10.1016/j.jseaes.2022.105355.

Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, Burlington, Massachusetts, Waltham.

Harris, N.B., Peters, K.E., 2012. Analyzing thermal histories of sedimentary basins: methods and case studies–introduction. Analysing thermal Histories of Sedimentary Basins: Methods and Case Studies. SEPM Special Publications 103. https://doi.org/10.2110/sepmsp.103, 218 p.

Henry, D.G., Jarvis, I., Gillmore, G., Stephenson, M., Emmings, J.F., 2018. Assessing low-maturity organic matter in shales using Raman spectroscopy: Effects of sample preparation and operating procedure. Int. J. Coal Geol. 191, 135–151. https://doi.org/10.1016/j.coal.2018.03.005.

Henry, D.G., Jarvis, I., Gillmore, G., Stephenson, M., 2019. Raman spectroscopy as a tool to determine the thermal maturity of organic matter: Application to sedimentary, metamorphic and structural geology. Earth Sci. Rev. 198, 102936 https://doi.org/10.1016/j.earscirev.2019.102936.

Hickman-Lewis, K., Moore, K.R., Hollis, J.J.R., Tuite, M.L., Beegle, L.W., Bhartia, R., Grotzinger, J.P., Brown, A.J., Shkolyar, S., Cavalazzi, B., Smith, C.L., 2022. In situ Identification of Paleoarchean Biosignatures using Colocated Perseverance Rover analyses: Perspectives for in situ Mars Science and Sample return. Astrobiology. https://doi.org/10.1089/ast.2022.0018.

Hinrichs, R., Brown, M.T., Vasconcellos, M.A., Abrashev, M.V., Kalkreuth, W., 2014. Simple procedure for an estimation of the coal rank using micro-Raman spectroscopy. Int. J. Coal Geol. 136, 52–58. https://doi.org/10.1016/j.coal.2014.10.013.

Iyer, K., Svensen, H., Schmid, D.W., 2018. SILLi 1.0: a 1-D numerical tool quantifying the thermal effects of sill intrusions. Geosci. Model Dev. 11 (1), 43–60. https://doi.org/10.5194/gmd-11-43-2018.

Izart, A., Barbarand, J., Michels, R., Privalov, V.A., 2016. Modelling of the thermal history of the Carboniferous Lorraine Coal Basin: Consequences for coal bed methane. Int. J. Coal Geol. 168, 253–274. https://doi.org/10.1016/j.coal.2016.11.008.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice-Hall, Inc., NJ.

Jain, A.K., Duin, R.P.W., Mao, J., 2000. Statistical pattern recognition: A review. IEEE Trans. Pattern Anal. Mach. Intell. 22 (1), 4–37. https://doi.org/10.1109/34.824819.

Jehlička, J., Bény, C., 1992. Application of Raman microspectrometry in the study of structural changes in Precambrian kerogens during regional metamorphism. Org. Geochem. 18 (2), 211–213. https://doi.org/10.1016/0146-6380(92)90132-H.

Jolliffe, I.T., 1986. Principal Component Analysis. Springer-Verlag, New York.

Kaneki, S., Hirono, T., Mukoyoshi, H., Sampei, Y., Ikehara, M., 2016. Organochemical characteristics of carbonaceous materials as indicators of heat recorded on an ancient plate-subduction fault. Geochem. Geophys. Geosyst. 17 (7), 2855–2868. https://doi.org/10.1002/2016GC006368.

Karg, H., Sauerer, B., 2022. Thermal maturity assessment of marine source rocks integrating Raman spectroscopy, organic geochemistry and petroleum systems modeling. Int. J. Coal Geol. 264, 104131 https://doi.org/10.1016/j.coal.2022.104131.

Katz, B.J., 1995. The Schistes Carton—the Lower Toarcian of the Paris Basin. In: Katz, B. J. (Ed.), Petroleum Source Rocks. Springer-Verlag, New York, pp. 51–65. https://doi.org/10.1007/978-3-642-78911-3_4.

Kedar, L., Bond, C.E., Muirhead, D., 2020. Carbon ordering in an aseismic shear zone: Implications for raman geothermometry and strain tracking. Earth Planet. Sci. Lett. 549, 116536 https://doi.org/10.1016/j.epsl.2020.116536.

Kitamura, M., Mukoyoshi, H., Fulton, P.M., Hirose, T., 2012. Coal maturation by frictional heat during rapid fault slip. Geophys. Res. Lett. 39 (16) https://doi.org/10.1029/2012GL052316.

Lahfid, A., Beyssac, O., Deville, E., Negro, F., Chopin, C., Goffé, B., 2010. Evolution of the Raman spectrum of carbonaceous material in low-grade metasediments of the Glarus Alps (Switzerland). Terra Nova 22 (5), 354–360. https://doi.org/10.1111/j.1365-3121.2010.00956.x.

Lei, M., Rao, Z., Wang, H., Chen, Y., Zou, L., Yu, H., 2021. Maceral groups analysis of coal based on semantic segmentation of photomicrographs via the improved U-net. Fuel 294, 120475. https://doi.org/10.1016/j.fuel.2021.120475.

Li, X., Hayashi, J.I., Li, C.Z., 2006. FT-Raman spectroscopic study of the evolution of char structure during the pyrolysis of a Victorian brown coal. Fuel 85 (12−13), 1700–1707. https://doi.org/10.1016/j.fuel.2006.03.008.

Li, K., Rimmer, S.M., Presswood, S.M., Liu, Q., 2020. Raman spectroscopy of intruded coals from the Illinois Basin: Correlation with rank and estimated alteration temperature. Int. J. Coal Geol. 219, 103369.

Lucca, A., Storti, F., Molli, G., Muchez, P., Schito, A., Artoni, A., Balsamo, F., Corrado, S., Mariani, E.S., 2019. Seismically enhanced hydrothermal plume advection through the process zone of the Compione extensional Fault, Northern. Apennines, Italy. Bulletin 131 (3–4), 547–571. https://doi.org/10.1130/B32029.1.

Lünsdorf, N.K., 2016. Raman spectroscopy of dispersed vitrinite—Methodical aspects and correlation with reflectance. Int. J. Coal Geol. 153, 75–86. https://doi.org/10.1016/j.coal.2015.11.010.

Luo, Q., Fariborz, G., Zhong, N., Wang, Y., Qiu, N., Skovsted, C.B., Suchy, V., Schovsbo, N.H., Morga, R., Xu, Y., Hao, J., Liu, A., Wu, J., Cao, W., Min, X., Wu, J., 2020. Graptolites as fossil geo-thermometers and source material of hydrocarbons: an overview of four decades of progress. Earth Sci. Rev. 200, 103000 https://doi.org/10.1016/j.earscirev.2019.103000.

Marzoli, A., Renne, P.R., Piccirillo, E.M., Ernesto, M., Bellieni, G., De Min, A., 1999. Extensive 200-million-year-old continental flood basalts of the Central Atlantic Magmatic Province. Science 284, 616–618. https://doi.org/10.1126/science.284.5414.616.

McCartney, J.T., Teichmüller, M., 1972. Classification of coals according to degree of coalification by reflectance of the vitrinite component. Fuel 51 (1), 64–68. https://doi.org/10.1016/0016-2361(72)90041-5.

McLachlan, G.J., Basford, K.E., 1988. Mixture models: Inference and applications to clustering, 38. M. Dekker, New York.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, New York (419 p).

Mlynarczuk, M., Skiba, M., 2017. The application of artificial intelligence for the identification of the maceral groups and mineral components of coal. Comput. Geosci. 103, 133–141. https://doi.org/10.1016/j.cageo.2017.03.011.

Morey, L.C., Agresti, A., 1984. The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. Educ. Psychol. Meas. 44 (1), 33–37. https://doi.org/10.1177/0013164484441003.

Muirhead, D.K., Bond, C.E., Watkins, H., Butler, R.W.H., Schito, A., Crawford, Z., Marpino, A., 2020. Raman spectroscopy: an effective thermal marker in low temperature carbonaceous fold–thrust belts. Geol. Soc. Lond., Spec. Publ. 490 (1), 135–151. https://doi.org/10.1144/SP490-2019-27.

Muirhead, D.K., Kedar, L., Schito, A., Corrado, S., Bond, C.E., Romano, C., 2021. Raman spectral shifts in naturally faulted rocks. Geochem. Geophys. Geosyst. 22 (10) https://doi.org/10.1029/2021GC009923 e2021GC009923.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT press, London (1054 p).

Nasraoui, O., N'Cir, C.E.B., 2019. Clustering methods for big data analytics. Techniques, Toolboxes and Applications 1, 91–113.

Negri, F., di Donato, E., Tommasini, M., Castiglioni, C., Zerbi, G., Müllen, K., 2004. Resonance Raman contribution to the D band of carbon materials: modeling defects with quantum chemistry. J. Chem. Phys. 120, 11889–11900. https://doi.org/10.1063/1.1710853.

Nirrengarten, M., Mohn, G., Schito, A., Corrado, S., Gutiérrez-García, L., Bowden, S.A., Despinois, F., 2020. The thermal imprint of continental breakup during the formation of the South China Sea. Earth Planet. Sci. Lett. 531, 115972 https://doi.org/10.1016/j.epsl.2019.115972.

O'Haver, T., 2015. A Pragmatic Introduction to Signal Processing with applications in scientific measurement. https://terpconnect.umd.edu/~toh/spectrum/CurveFittingC.html.

Omran, M.G., Engelbrecht, A.P., Salman, A., 2007. An overview of clustering methods. Intelligent Data Analysis 11 (6), 583–605. https://doi.org/10.3233/IDA-2007-11602.

Palacio-Niño, J.O., Berzal, F., 2019. Evaluation metrics for unsupervised learning algorithms. Preprint at arXiv. https://doi.org/10.48550/arXiv.1905.05667.

Palumbo, F., Main, I.G., Zito, G., 1999. The thermal evolution of sedimentary basins and its effect on the maturation of hydrocarbons. Geophys. J. Int. 139 (1), 248–260. https://doi.org/10.1046/j.1365-246X.1999.00877.x.

Parnell, J., Lee, P., Osinski, G.R., Cockell, C.S., 2005. Application of organic geochemistry to detect signatures of organic matter in the Haughton impact structure. Meteorit. Planet. Sci. 40 (12), 1879–1885. https://doi.org/10.1111/j.1945-5100.2005.tb00151.x.

Perrodon, A., Zabek, J., 1991. Interior Cratonic Basins. Analog Basins: Paris Basin. AAPG Memoir, 51, pp. 663–679.

Pimenta, M.A., Dresselhaus, G., Dresselhaus, M.S., Cancado, L.G., Jorio, A., Saito, R., 2007. Studying disorder in graphite-based systems by Raman spectroscopy. Phys. Chem. Chem. Phys. 9 (11), 1276–1290. https://doi.org/10.1039/B613962K.

Pócsik, I., Hundhausen, M., Koós, M., Ley, L., 1998. Origin of the D peak in the Raman spectrum of microcrystalline graphite. J. Non-Cryst. Solids 227, 1083–1086. https://doi.org/10.1016/S0022-3093(98)00349-4.

Pomerol, C., 1989. Stratigraphy of the Palaeogene: hiatuses and transitions. Proc. Geol. Assoc. 100 (3), 313–324. https://doi.org/10.1016/S0016-7878(89)80051-3.

Poulet, M., Espitalié, J., 1987. Hydrocarbon migration in the Paris Basin. Collection colloques et séminaires-Institut français du pétrole 45, 131–171.

Quirico, E., Bonal, L., Montagnac, G., Beck, P., Reynard, B., 2020. New insights into the structure and formation of coals, terrestrial and extraterrestrial kerogens from resonant UV Raman spectroscopy. Geochim. Cosmochim. Acta 282, 156–176. https://doi.org/10.1016/j.gca.2020.05.028.

Rahl, J.M., Anderson, K.M., Brandon, M.T., Fassoulas, C., 2005. Raman spectroscopic carbonaceous material thermometry of low-grade metamorphic rocks: Calibration and application to tectonic exhumation in Crete. Greece. Earth and Planetary Science Letters 240 (2), 339–354. https://doi.org/10.1016/j.epsl.2005.09.055.

Rebelo, S.L., Guedes, A., Szefczyk, M.E., Pereira, A.M., Araújo, J.P., Freire, C., 2016. Progress in the Raman spectra analysis of covalently functionalized multiwalled carbon nanotubes: unraveling disorder in graphitic materials. Phys. Chem. Chem. Phys. 18 (18), 12784–12796. https://doi.org/10.1039/C5CP06519D.

Reich, S., Thomsen, C., 2004. Raman spectroscopy of graphite. Philos. Trans. R. Soc. London, Ser. A 362 (1824), 2271–2288. https://doi.org/10.1098/rsta.2004.1454.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Santos, J.M., Embrechts, M., 2009. On the use of the adjusted rand index as a metric for evaluating supervised classification. In: *Artificial Neural Networks–ICANN 2009: 19th International Conference, Limassol, Cyprus, September 14-17, 2009, Proceedings, Part II 19*. Springer Berlin Heidelberg, pp. 175–184. https://doi.org/10.1007/978-3-642-04277-5_18.

Sauerer, B., Craddock, P.R., AlJohani, M.D., Alsamadony, K.L., Abdallah, W., 2017. Fast and accurate shale maturity determination by Raman spectroscopy measurement with minimal sample preparation. Int. J. Coal Geol. 173, 150–157. https://doi.org/10.1016/j.coal.2017.02.008.

Sauerer, B., Furmann, A., Fernandes, A., Samara, H., Jaeger, P., Al-Ayed, O., Abdallah, W., 2021. Assessing extreme maturities–Challenging examples from immature Jordanian to overmature Far Eastern unconventional formations. Mar. Pet. Geol. 129, 105103 https://doi.org/10.1016/j.marpetgeo.2021.105103.

Schito, A., Corrado, S., 2018. An automatic approach for characterization of the thermal maturity of dispersed organic matter Raman spectra at low diagenetic stages. Geol. Soc. London Spec. Publ. 484, 107–119. https://doi.org/10.1144/SP484.5.

Schito, A., Corrado, S., 2020. An automatic approach for characterization of the thermal maturity of dispersed organic matter Raman spectra at low diagenetic stages. Geol. Soc. Lond., Spec. Publ. 484 (1), 107–119. https://doi.org/10.1144/SP484.

Schito, A., Corrado, S., Aldega, L., Grigo, D., 2016. Overcoming pitfalls of vitrinite reflectance measurements in the assessment of thermal maturity: the case history of the lower Congo basin. Mar. Pet. Geol. 74, 59–70. https://doi.org/10.1016/j.marpetgeo.2016.04.002.

Schito, A., Romano, C., Corrado, S., Grigo, D., Poe, B., 2017. Diagenetic thermal evolution of organic matter by Raman spectroscopy. Org. Geochem. 106, 57–67. https://doi.org/10.1016/j.orggeochem.2016.12.006.

Schito, A., Spina, A., Corrado, S., Cirilli, S., Romano, C., 2019. Comparing optical and Raman spectroscopic investigations of phytoclasts and sporomorphs for thermal maturity assessment: the case study of Hettangian continental facies in the Holy Cross Mts.(Central Poland). Mar. Pet. Geol. 104, 331–345. https://doi.org/10.1016/j.marpetgeo.2019.03.008.

Schito, A., Guedes, A., Valentim, B., Vergara Sassarini, N.A., Corrado, S., 2021. A Predictive Model for Maceral Discrimination by Means of Raman Spectra on Dispersed Organic Matter: A Case Study from the Carpathian Fold-and-Thrust Belt (Ukraine). Geosciences 11 (5), 213. https://doi.org/10.3390/geosciences11050213.

Schito, A., Muirhead, D., Bowden, S., Parnell, J., 2022. Hydrothermal generation of hydrocarbons in basement rocks, Southern Tuscany. Italian Journal of Geosciences 141 (2), 231–244. https://doi.org/10.3301/IJG.2022.10.

Schito, A., Muirhead, D.K., Parnell, J., 2023. Towards a kerogen-to-graphite kinetic model by means of Raman spectroscopy. Earth Sci. Rev. 104292 https://doi.org/10.1016/j.earscirev.2022.104292.

Schmidt, J.S., Hinrichs, R., Araujo, C.V., 2017. Maturity estimation of phytoclasts in strew mounts by micro-Raman spectroscopy. Int. J. Coal Geol. 173, 1–8. https://doi.org/10.1016/j.coal.2017.02.003.

Schopf, J.W., Kudryavtsev, A.B., Agresti, D.G., Czaja, A.D., Wdowiak, T.J., 2005. Raman imagery: a new approach to assess the geochemical maturity and biogenicity of permineralized Precambrian fossils. Astrobiology 5 (3), 333–371. https://doi.org/10.1089/ast.2005.5.333.

Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X., 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS) 42 (3), 1–21. https://doi.org/10.1145/3068335.

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Stat. 6, 461–464. https://www.jstor.org/stable/2958889.

Skiba, M., Młynarczuk, M., 2018. Identification of macerals of the inertinite group using neural classifiers, based on selected textural features. Arch. Min. Sci. 63 (4) https://doi.org/10.24425/ams.2018.124978.

Sorci, A., Cirilli, S., Clayton, G., Corrado, S., Hints, O., Goodhue, R., Schito, A., Spina, A., 2020. Palynomorph optical analyses for thermal maturity assessment of Upper Ordovician (Katian-Hirnantian) rocks from Southern Estonia. Mar. Pet. Geol. 120, 104574 https://doi.org/10.1016/j.marpetgeo.2020.104574.

Svensen, H., Planke, S., Malthe-Sørenssen, A., Jamtveit, B., Myklebust, R., Eidem, T.R., Rey, S.S., 2004. Release of methane from a volcanic basin as a mechanism for initial Eocene global warming. Nature 429 (6991), 542–545. https://doi.org/10.1038/nature02575.

Tan, P.N., Steinbach, M., Kumar, V., 2013. Introduction to Data Mining, 487–533. Pearson Education limited, Harlow, Essex, UK.

Thomsen, C., Reich, S., 2000. Double resonant Raman scattering in graphite. Phys. Rev. Lett. 85 (24), 5214. https://doi.org/10.1103/PhysRevLett.85.5214.

Tissot, B.P., 1984. Recent advances in petroleum geochemistry applied to hydrocarbon exploration. AAPG Bull. 68 (5), 545–563. https://doi.org/10.1306/AD461336-16F7-11D7-8645000102C1865D.

Tissot, B.P., Califet-Debyser, Y., Deroo, G., Oudin, J.L., 1971. Origin and evolution of hydrocarbons in early Toarcian Shales, Paris Basin, France. AAPG Bulletin 55 (12), 2177–2193. https://doi.org/10.1306/819A3E2E-16C5-11D7-8645000102C1865D.

Titterington, D.M., 1985. Common structure of smoothing techniques in statistics. Int. Stat. Rev./Revue Internationale de Statistique 141–170. https://doi.org/10.2307/1402932.

Toro, J., Roure, F., Bordas-Le Floch, N., Le Cornec-Lance, S., Sassi, W., 2004. Thermal and kinematic evolution of the Eastern Cordillera fold and thrust belt, Colombia. In: Swennen, R., Roure, F., Granath, J.W. (Eds.), Deformation,fluid flow, and reservoir appraisal in foreland fold and thrust belts: AAPG Hedberg Series, no. 1, pp. 79–115. https://doi.org/10.1306/1025687H13114.

Traverse, A., 2007. Paleopalynology, vol. 28. Springer Science & Business Media, Dordrecht, The Netherlands.

Tuinstra, F., Koenig, J.L., 1970. Raman spectrum of graphite. J. Chem. Phys. 53 (3), 1126–1130. https://doi.org/10.1063/1.1674108.

Tyson, R.V., 1995. Sedimentary organic matter: organic facies and palynofacies. Springer, Dordrecht. https://doi.org/10.1007/978-94-011-0739-6.

Vandenbroucke, M., Largeau, C., 2007. Kerogen origin, evolution and structure. Org. Geochem. 38 (5), 719–833. https://doi.org/10.1016/j.orggeochem.2007.01.001.

Vergara Sassarini, N.A., 2022. New Approaches for Sedimentary Basins Thermal Calibration: Towards Integration of Source Rocks Raman Spectroscopy and Carbonate Thermometry. Thesis. University of Roma Tre (262 pp).

Wang, H., Lei, M., Chen, Y., Li, M., Zou, L., 2019. Intelligent identification of maceral components of coal based on image segmentation and classification. Appl. Sci. 9 (16), 3245. https://doi.org/10.3390/app9163245.

Wang, Y., Bai, X., Wu, L., Zhang, Y., Qu, S., 2022. Identification of maceral groups in Chinese bituminous coals based on semantic segmentation models. Fuel 308, 121844. https://doi.org/10.1016/j.fuel.2021.121844.

Westall, F., Hickman-Lewis, K., Cavalazzi, B., Foucher, F., Clodoré, L., Vago, J.L., 2021. On biosignatures for Mars. Int. J. Astrobiol. 20 (6), 377–393. https://doi.org/10.1017/S1473550421000264.

Wilkins, R.W., Boudou, R., Sherwood, N., Xiao, X., 2014. Thermal maturity evaluation from inertinites by Raman spectroscopy: the 'RaMM'technique. Int. J. Coal Geol. 128, 143–152. https://doi.org/10.1016/j.coal.2014.03.006.

Wopenka, B., Pasteris, J.D., 1993. Structural characterization of kerogens to granulite-facies graphite: applicability of Raman microprobe spectroscopy. Am. Min. 78 (5–6), 533–557.