Adaptive Hierarchical Clustering for Petrographic Image Analysis

Andrea Pasini
Politecnico di Torino
Torino, Italy
andrea.pasini@polito.it

Elena Baralis
Politecnico di Torino
Torino, Italy
elena.baralis@polito.it

Paolo Garza
Politecnico di Torino
Torino, Italy
paolo.garza@polito.it

Davide Floriello
ENI S.p.A.
San Donato Milanese (MI), Italy
Davide.Floriello@eni.com

Michela Idiomi
ENI S.p.A.
San Donato Milanese (MI), Italy
Michela.Idiomi@eni.com

Andrea Ortenzi
ENI S.p.A.
San Donato Milanese (MI), Italy
Andrea.Ortenzi@eni.com

Simone Ricci
ENI S.p.A.
San Donato Milanese (MI), Italy
Simone.Ricci@eni.com

ABSTRACT

The task of analyzing the permeability of soil to natural gas and oil has been studied by experts for many years. This analysis is typically achieved by inspecting samples of terrain, called thin sections. An important procedure of this task is the categorization/clusterization of pores, which are microscopical cavities in thin sections. This operation is manually carried out by domain experts through the analysis of high-resolution images retrieved with Scanning Electron Microscopes (SEM). Since the number of pores is very high (more than 10,000 for each thin section), the manual categorization procedure is performed on a small subset of pores and hence the achieved estimations can be imprecise.

To address the pore analysis problem, we propose a custom clustering pipeline that automatically groups pores inside thin sections. The work has been conducted with the help of ENI, a leading company in oil and gas extraction. Specifically, we have designed the Adaptive Multi-level Dendrogram Cut method, a customized version of hierarchical clustering. The proposed method is able to cut the hierarchical dendrogram at multiple levels, being guided by both automatic metrics and domain expert knowledge. In this paper, we show that our methodology produces higher quality results with respect to standard hierarchical clustering algorithms. We also defined a set of techniques to generate interpretable descriptions of the pore clusters. According to ENI's experts, the obtained clusters have a good quality from a geological point of view and help to automatize a very slow process that was carried out manually.

1 INTRODUCTION

Petrography is the branch of petrology focused on classifying and describing rocks from both a microscopical and a megascopical point of view [3]. With the diffusion of Scanning Electron Microscopes (SEM), optical analyses for petrography reached good quality results due to the high resolution of the images. Indeed, SEM images can reveal features that are invisible with optical microscopy [8, 12, 15]. In particular, they allow distinguishing different mineral phases based on variations of gray intensity. Other information such as the size, distribution, shape, orientation and textural relationships of minerals can be retrieved.

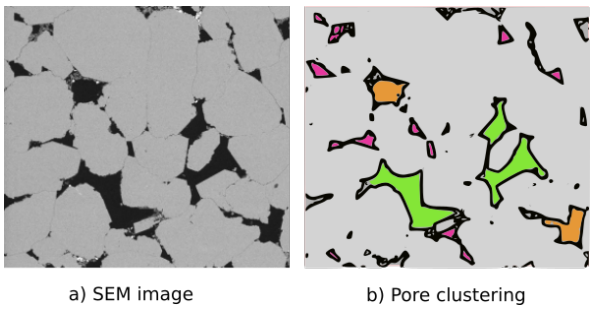


Figure 1: Example of thin section analysis. On the left the image retrieved with SEM, on the right the result after applying our clustering method.

Specifically, SEM images are highly useful for *pore* characterization in rocks [2, 16]. Petrographic Image Analysis (PIA) studies the presence and distribution of pores, which is important for analyzing the permeability of rocks to natural gas and oil [6, 9]. Within this process, SEM images are retrieved from ground samples called *thin sections*. After acquisition, image processing algorithms can extract different characteristics of pores, for example their size and shape. [4, 13]. Finally, domain experts use these results to inspect rock properties such as their permeability.

Since each thin section possibly contains from thousands to millions of pores, their manual categorization is applicable only to a small subset. This entails an approximated result that highly depends on the sampling performed over the initial data. An automatic categorization/clusterization method may help avoiding most of the manual effort that is required to perform the analysis. Both classification and clustering methods could be taken in consideration. However, classification requires having a sufficient amount of training data with manually generated labels. This implies that domain experts should label large quantities of pores to model the characteristics of different categories. Since there are no public datasets with labeled pores and generating new labeled data is very time consuming, we adopted a clustering based approach. Clustering methods are able to learn the data

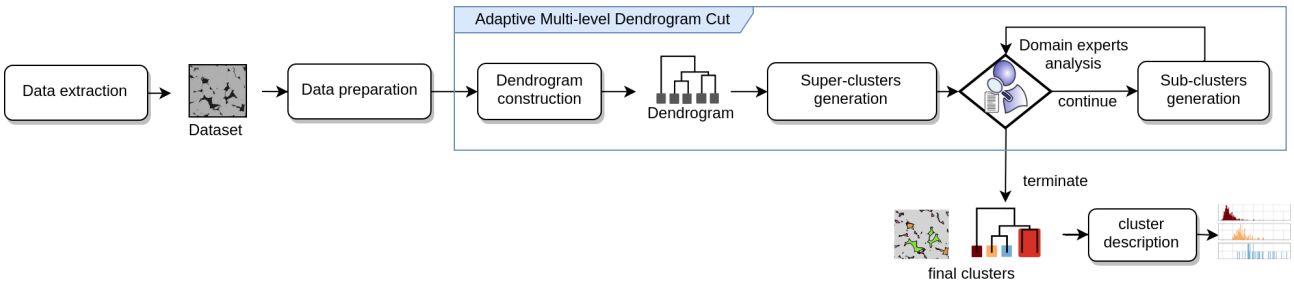


Figure 2: Big Petro pipeline.

distribution and automatically derive a set of groups, which then can be more easily and quickly mapped to geological categories by domain experts. Clustering algorithms do not need training labels and can be applied to thousands of samples. With the proposed methodology, domain experts can concentrate their efforts only on the geological interpretation of each cluster, while leaving to the algorithm the most time consuming task.

In this paper, we propose the Big Petro pipeline, a clustering methodology specifically designed for automatically categorizing pores inside thin sections. The proposed pipeline has been designed and evaluated on real datasets with the support of the domain experts of ENI, a leading company in oil and gas extraction. The contribution of our paper can be summarized as follows:

- Development of a customized *semi-automatic process* for clustering geological pores. This methodology helps domain experts to avoid manually categorizing all pores.
- Definition of a variant of the standard usage of hierarchical clustering. Specifically, we propose to obtain clusters with an *Adaptive Multi-level Dendrogram Cut*, being guided by both automatic metrics and domain experts knowledge.
- Inspection of different methodologies to *describe* the generated clusters in terms of characterizing attributes distribution for each cluster. The description approaches allow domain experts to interpret the different groups and derive interesting insights from the analyzed thin sections.

2 RELATED WORK

Clustering techniques [11] are widely used in many application fields. To the best of our knowledge, this is the first attempt to apply clustering techniques to automatize pore analysis. During our preliminary study of geological pores categorization, we inspected the behavior of the following algorithms: DBSCAN [5], KMeans [7], and agglomerative hierarchical clustering [10].

DBSCAN [5] is a density based algorithm. Its driving idea is joining points in the same cluster when they exceed a minimum neighborhood density. The main issue with this technique is that the distribution of our data shows variable densities. Setting a low minimum density results in a single cluster containing all the pores, while when setting a high density many pores are considered as noise. Multiple runs of DBSCAN with different configurations could potentially allow identifying clusters with different densities. However, also the self-tuning techniques that use multiple runs of DBSCAN [1] did not yield good quality results on our data.

K-means [7] is a well-known centroid based clustering algorithm, which requires specifying the (fixed) desired number of clusters. The results with this technique were not satisfactory, as this method builds globular clusters with a balanced number of points, but the types of pores in our datasets present more complex shapes and are very imbalanced in number. For example the category of small pores is typically an order of magnitude more numerous than the one containing bigger pores.

Hierarchical clustering [10] techniques allow defining clusters by inspecting the cluster hierarchy with a structure called dendrogram. By cutting the dendrogram at a fixed level, k clusters are generated. The generated clusters can be characterized by different densities and imbalanced cardinalities. This technique, already in its standard version, allowed us to obtain the best results, according to the analysis performed by ENI domain experts. For this reason, we selected hierarchical clustering as preferred method to be customized and integrated in our Big Petro pipeline.

3 BIG PETRO PIPELINE

The objective of our work is to automatically discover different categories of geological pores in SEM images of rock samples by means of the proposed Big Petro pipeline. The Big Petro pipeline is composed of different blocks, depicted in Figure 2. In the following paragraphs we describe the characteristics of each block.

Data extraction. Each dataset to be analyzed consists of a single thin section acquisition, represented as a grayscale image. Figure 1a shows an example of a small portion of thin section, where the black irregular spots are the pores to be categorized. In this first step, a tool developed by ENI processes the grayscale image that automatically selects the darker regions representing pores. Afterwards, the tool analyzes each pore and extracts a set of geometrical features describing the pore shape and size. The result of this operation is a structured dataset, with a record for each pore, characterized by 32 numerical attributes. Section 5 describes more in detail the characteristics of the considered datasets, extracted from different thin section samples.

Data preparation. The second step of the pipeline involves the preprocessing of the datasets obtained in the data extraction phase. Each dataset is typically characterized by the presence of many small pores represented by only few pixels. The low pixel resolution of these pores does not allow computing most of the geometrical characteristics necessary for the analysis. Hence, we filter out the smaller pores by applying a domain provided threshold on their size. This operation also yielded smaller, more manageable datasets (from millions of pores to thousands), which allowed a reduction of the clustering algorithm processing time. In Section 5, Table 1 shows further details of the dataset cardinality before and after this filtering procedure.

Next, attribute values are normalized by applying the z-score normalization, after which each attribute domain range is characterized by mean equal to zero and unitary variance. This step

is necessary to guide the distance metric used for clustering to treat all the attributes in the same way, independently of their value range.

Adaptive Multi-level Dendrogram Cut. Pore clustering takes place in this phase. It is composed of four steps: (i) dendrogram construction, (ii) *super-clusters* generation, (iii) domain experts analysis, and (iv) *sub-clusters* generation. The first task is addressed by running an agglomerative hierarchical clustering algorithm, which builds the dendrogram describing the hierarchy of the pore aggregations. The second step consists of generating a first set of clusters by cutting the dendrogram at a specific height. These clusters correspond to macro-groups of pores, called *super-clusters* and provide an initial coarse-grain aggregation of pores. Their number is chosen by analyzing the silhouette score [14], as described in Section 5.

Next, each super-cluster is manually analyzed by domain experts to decide if further partitioning is needed. The standard approach consisting of one single cut of the dendrogram may generate some clusters already containing homogeneous pores and some other clusters grouping heterogeneous pores. Simply cutting the dendrogram at a different height will split also the already high-quality macro-clusters. For this reason, only for the subset of super-clusters that are deemed by the domain experts to need further split in sub-clusters, the *sub-cluster* generation step is executed. Specifically, given a cluster to be further divided, we analyze its corresponding dendrogram sub-tree and cut the hierarchy at a new level of depth. The process continues with a loop involving the domain expert analysis and the sub-clusters generation. At each iteration, new levels of sub-clusters are added until the results are satisfactory. Further details on this process are provided in Section 4.

Cluster description. In this step, the obtained clusters are analyzed with different descriptive techniques. In Section 5, we show how, by inspecting the attribute distribution in the obtained clusters and the aggregation hierarchy in the dendrogram, interesting insights are obtained. We also consider PCA plots to show the distribution of the cluster labels along the directions where data have maximum variance.

4 ADAPTIVE MULTI-LEVEL DENDROGRAM CUT

The dendrogram is generated by an agglomerative hierarchical clustering algorithm. Since our datasets consist of continuous attributes, the euclidean metric is used for computing the distance between pore pairs. After inspecting the results obtained with different linkage techniques, we selected Ward's method [17]. In particular the single linkage criterion did not fit our purposes, as it is sensible to noise and the points belonging to lower density clusters are not merged properly in the dendrogram. More specifically, single linkage connects two clusters when they present a pair of points which are very close together. For this reason higher density clusters are merged together in the first stages of the dendogram and points in lower density regions appear in different isolated clusters. Instead, complete linkage tends to break bigger clusters to have equal-sized groups. The algorithm behaviour with this metric is approximately similar to K-means and does not produce good quality results. Average linkage produces a trade-off result between the single and complete methods. However, by inspecting the first levels of the dendrogram together with ENI domain experts, we observed that Ward's method yields a better separation of the main pore groups.

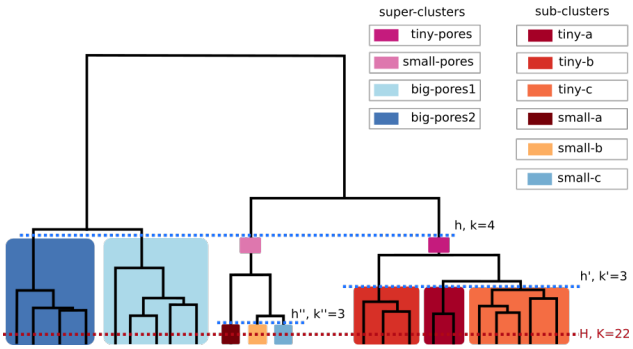


Figure 3: Dendrogram for Dset1.

As mentioned before, agglomerative hierarchical clustering yields a binary tree structure, called dendrogram. The leaves of the tree represent single pores to be clustered, while nodes describe the merging points of clusters at different hierarchical levels. The height of each merging point is defined by the linkage distance between the two nodes to be merged.

Let C_h be the set of k clusters obtained by cutting the dendrogram at a specific height h:

$$C_h = \{c_0^h, ..., c_i^h, ..., c_{k-1}^h\}$$

The dendrogram cut at height h is exploited to generate the first level of groups, denoted as *super-clusters*. For example in Figure 3 the obtained super-clusters are *tiny-pores*, *small-pores*, *big-pores1* and *big-pores2* (these names have been conventionally assigned by domain experts after the analysis of the macro-clusters). The value of h is selected by inspecting the silhouette value of the obtained groups. In Section 5 we further discuss how this value is chosen for the different analyzed datasets.

When the domain experts deem the granularity of some macro-clusters as insufficient, the iterative generation of *sub-clusters* is activated. Sub-clusters are generated by inspecting the descendants of their corresponding super-cluster. The set of sub-clusters of a cluster c_i^h obtained by cutting the dendrogram at height h' is defined as:

$$subclusters_{h'}(c_i^h) = \{c_j^{h'} : c_j^{h'} \in descendants(c_i^h)\}$$

where $descendants(c_i^h)$ is the set of descendants of c_i^h in the tree hierarchy and $c_j^{h'}$ are the nodes obtained by cutting the dendogram at the new height h'. For example, in Figure 3 the set $subclusters_{h'}(tiny\text{-}pores)$ is composed of sub-clusters *tiny-a*, *tiny-b*, *tiny-c*. The definition of $subclusters_{h'}(c_i^h)$ can be applied recursively to obtain multiple sets of subclusters, at different levels of the hierarchy for different cluster subsets, until the results are satisfactory (see Figure 2).

5 EXPERIMENTAL RESULTS

ENI provided us three different datasets (*Dset1*, *Dset2* and *Dset3*), whose samples were extracted from three thin sections mostly composed by carbonates. As described in Section 3, the smallest pores in each dataset are filtered out using a threshold specified by ENI domain experts. Consider the left part of Table 1. The biggest dataset, *Dset3*, presents $1.3M$ pores, while after filtering its size is $4K$ rows. The smallest dataset, *Dset2*, changes from $300K$ to $5.9K$ pores. Note that, even if the original size of the datasets can be quite different, the number of pores after filtering takes similar values, ranging between $4K$ and $6K$.

Datasets size			Super-clusters size				
Dataset	# pores	# filtered pores	# tiny-pores	# small-pores	# big-pores1	# big-pores2	# big-pores3
Dset1	330346	6771	6194	489	74	14	
Dset2	301828	5880	5025	634	213	7	1
Dset3	1349554	4164	4034	103	2	25	

Table 1: Datasets and super-clusters size.

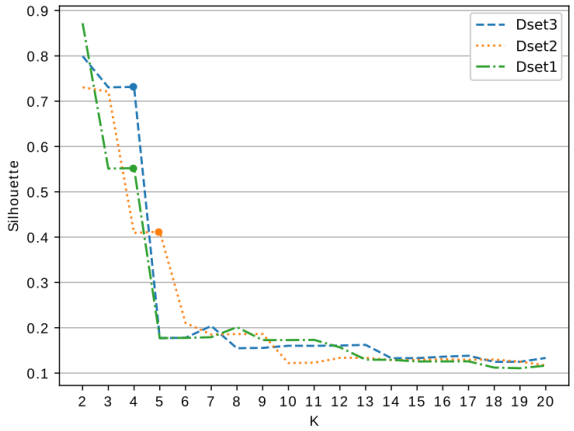


Figure 4: Super clusters generation. Silhouette varying the number of clusters k.

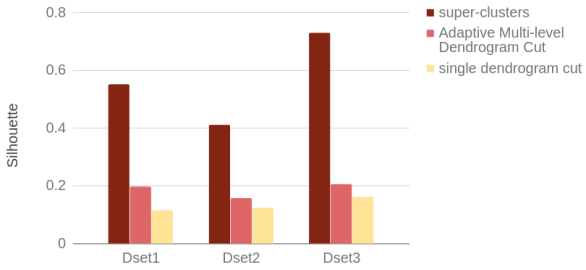


Figure 5: Silhouette for different methods.

Super-clusters generation. After building the dendrogram on the three datasets, we selected the proper values of h to cut the hierarchy tree and obtain the super-clusters. In particular, Figure 4 plots the silhouette values against the number k of super-clusters in each dataset. For all datasets the silhouette score decreases when k grows. However, even if the silhouette is higher with lower values of k, a very low number of super-clusters does not provide meaningful groups from a geological point of view. Datasets *Dset1* and *Dset3* show stable silhouette values in the range $k = 3$ to $k = 4$. Hence, we selected $k = 4$, the maximum value of k before the silhouette rapidly decreases, for these two datasets. Instead, dataset *Dset2* shows stable silhouette values for $k = 2, 3$ and $k = 4, 5$. For the same motivation, we set $k = 5$, i.e., the highest number of clusters in the region of stable silhouette values.

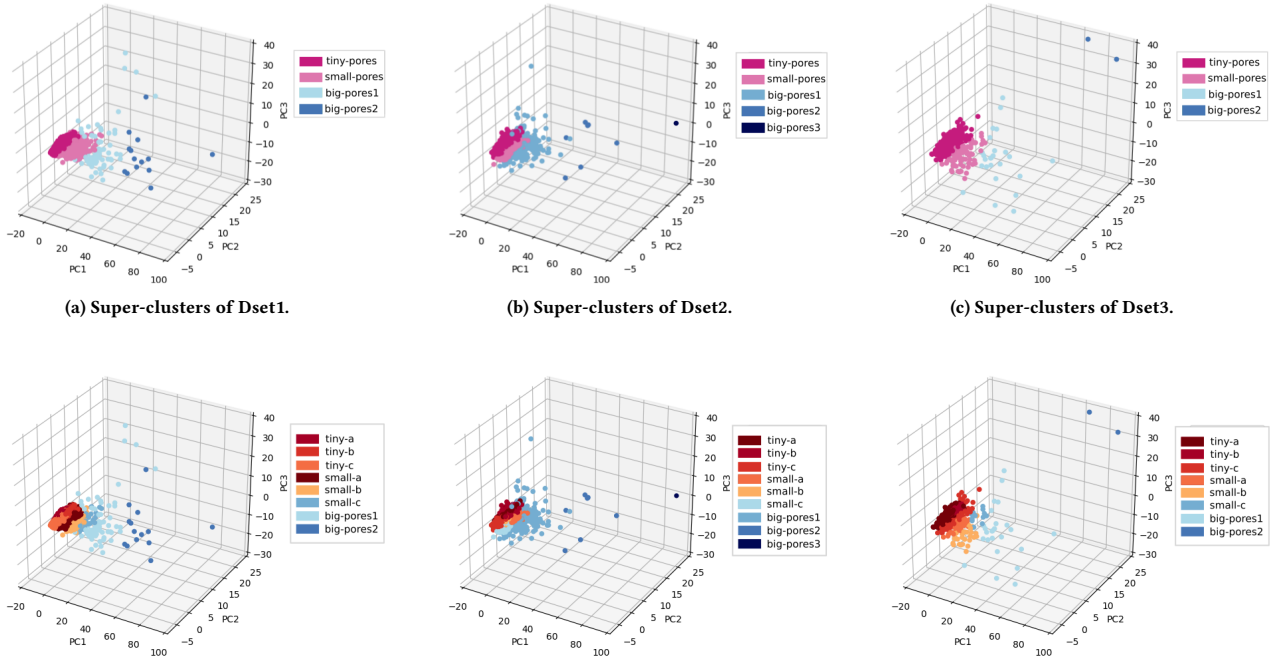
Table 1 shows the size of the obtained super-clusters. All datasets are characterized by two big clusters (*tiny-pores*, *small-pores*) and two smaller ones (*big-pores1, big-pores2*). Moreover, *Dset2* shows a cluster (*big-pores3*) with a single pore, which can be considered an outlier. This confirms that the choice $k = 5$ for dataset *Dset2* produces similar results to the ones obtained with the other two datasets.

Figures 6a, 6b, and 6c show the results of a *Principal Component Anaysis* (PCA) on the pores of the three datasets by visualizing the projection of the points on the first three components. The different colors represent points belonging to the generated super-clusters. For all three datasets, the two biggest clusters (*tiny-pores*, *small-pores*) are characterized by the highest density, while the others are sparser. Furthermore, the cluster structure is similar for all datasets. The biggest clusters are characterized by lower values of component *pc1*, while the smaller ones by higher values of *pc1*. Finally, the outlier in *Dset2* is clearly visible as an isolated point in Figure 6b (*big-pores3*).

Sub-clusters generation. According to the analysis performed by ENI domain experts, clusters *big-pores1* and *big-pores2* are pure and well characterized for all datasets. Instead, clusters *tiny-pores* and *small-pores* need to be further separated. Being guided by ENI domain experts we generated $k' = k'' = 3$ sub-clusters from *tiny-pores* and *small-pores* respectively. Figure 3 depicts the resulting dendrogram for dataset *Dset1*. The names of the sub-clusters are obtained by appending a suffix (*-a,-b,-c*) to the name of their corresponding super-cluster (e.g., *tiny-a, tiny-b*). Note that the cut point for the sub-clusters of *small-pores* is deeper in the dendogram than the one for the three sub-clusters of *tiny-pores*. This result cannot be obtained with a single dendrogram cut and motivates the need for the proposed Adaptive Multi-level Dendrogram Cut. To reach the granularity of the *small-pores* sub-clusters with a single dendrogram cut we need to choose a height H corresponding to $K = 22$ clusters. However, this K value would break the clusters *big-pores1*, *big-pores2*, *tiny-a*, *tiny-b*, *tiny-c* and would produce a lower quality clustering.

Figure 5 shows the silhouette value for the final clusters obtained with Adaptive Multi-level Dendrogram Cut and those generated with a single dendrogram cut. The number of clusters for the single dendrogram cut is $K = 22, 19, 13$ for *Dset1, Dset2* and *Dset3* respectively. The silhouette of the clusters obtained with our method is clearly higher than the one based on one single cut. This entails that our method not only produces a better partition from a geological point of view, but also a better domain-agnostic quality. A further discussion on the results shown in Figure 5 is provided in Section 6.

Cluster description. We provide a description of the obtained clusters by inspecting the attributes that allow characterizing them. The PCA representation in Figure 6 shows the distribution of the different super and sub-clusters. We inspected the attributes which are most representative of each of the three principal components, by considering the ones with highest eigenvalues. For example, the first component, *pc1*, is characterized by size attributes. In particular, higher values of *pc1* entail higher size of the pores. In Figures 6a, 6b, 6c, the super-clusters *tiny-pores* and *small-pores* are characterized by lower values of *pc1*, while *big-pores1* and *big-pores2* present higher values. Figure 7a shows the distribution of one of the *size* attributes in *Dset1* for the different super-clusters. It confirms that the distinction between super clusters is partially driven by the size of the pores.



(a) Super-clusters of Dset1.

(b) Super-clusters of Dset2.

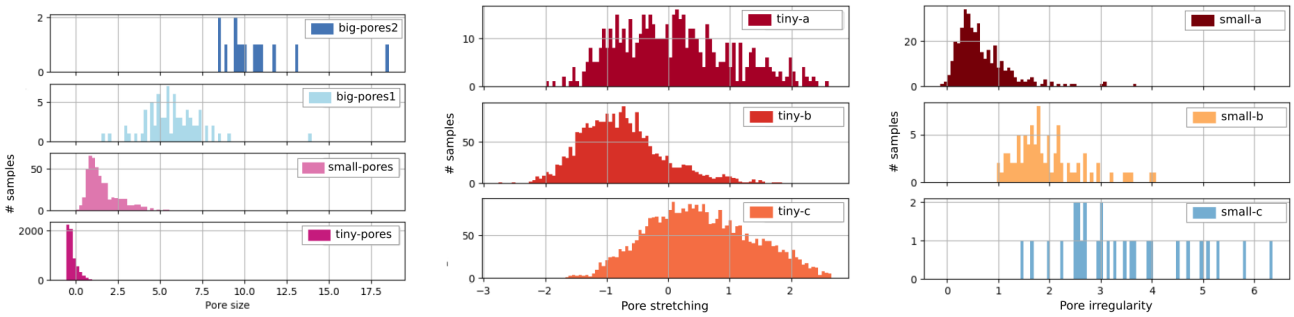
(c) Super-clusters of Dset3.

(d) Final clustering of Dset1.

(e) Final clustering of Dset2.

(f) Final clustering of Dset3.

Figure 6: Clusters obtained with hierarchical clustering for all datasets. PCA plots.



(a) Super-clusters of Dset1.

(b) Sub-clusters of Dset1.

(c) Sub-clusters of Dset1.

Figure 7: Cluster distributions.

Figures 6d, 6e, 6f show the final clustering, including both the sub-clusters and the super-clusters that have not been further divided. It can be noticed that sub-clusters present a finer separation that also involves components *pc2* and *pc3*. For example in *Dset1* clusters *tiny-b* and *tiny-c* are well separated by *pc2*. The most relevant attributes for *pc2* describe how much pores are stretched, obtaining a non circular shape. Looking at Figure 7b it is possible to see that pores in *tiny-c* show a higher stretching than those in *tiny-b* that are more circular. The third principal component, *pc3*, is related to pore irregularity, instead. From Figure 6d pores in cluster *small-a* present a lower shape irregularity than those in *small-c*. This is also confirmed by the histograms in Figure 7c. Following this methodology, ENI domain experts are able to derive interesting insights, describing the geological characteristics of the analyzed thin section.

6 LESSONS LEARNED

The analysis conducted in the previous sections highlights the following results. First, the silhouette score, a domain-agnostic quality index for evaluating clusters, is not always a good metric from a domain driven point of view. Consider Figure 5. The silhouette score computed on the super-clusters is higher than the one obtained with the Adaptive Multi-level Dendrogram Cut. For example, in *Dset3* super-clusters reach a silhouette score of 0.73, while Adaptive Multi-level Dendrogram Cut clusters score 0.2. However, the quality of the clusters from the domain experts point of view is higher for the proposed method, as the super-clusters show an excessively coarse subdivision. In fact, super-clusters have a strong correlation with the pore size, but cannot capture more complex characteristics of the pore shape.

Second, the cluster description confirmed that sub-clusters represent a finer subdivision of the initial groups according to

more complex characteristics, such as the pore stretching and shape irregularity. According to the domain experts, this grouping has a good quality from a geological point of view, because it generates groups that are distinguished by both geometrical and genetic characteristics. These characteristics are fundamental for geologists to inspect the structure of the analyzed thin section and its permeability. Hence, the role of domain knowledge is fundamental in driving the clustering process to obtain higher quality results.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed the Adaptive Multi-level Dendrogram Cut method for clustering geological pores in thin sections and we evaluated its results on three different real datasets provided by ENI. We demonstrated that the Adaptive Multi-level Dendrogram Cut allows obtaining sub-clusters that capture more complex characteristics of the pore shape and have higher silhouette values than the sub-clusters that would be obtained with the corresponding single level dendrogram cut. Currently, the Big Petro pipeline is under further evaluation by ENI domain experts, which are planning to deploy it shortly.

As future work, we plan to integrate this pipeline with domain driven self-tuning techniques that will allow fully-automatic pore clustering. Furthermore, we will work on improving the cluster description step to help domain experts in interpreting clustering results.

Acknowledgements

The research leading to these results has been supported by the SmartData@PoliTO center for Big Data and Machine Learning technologies.

REFERENCES

[1] Daniele Apiletti, Elena Baralis, Tania Cerquitelli, Paolo Garza, Danilo Giordano, Marco Mellia, and Luca Venturini. 2016. SeLINA: a self-learning insightful network analyzer. *IEEE Transactions on Network and Service Management* 13, 3 (2016), 696–710.

[2] Philip W Choquette and Lloyd C Pray. 1970. Geologic nomenclature and classification of porosity in sedimentary carbonates. *AAPG bulletin* 54, 2 (1970), 207–250.

[3] M.G. Edwards. 2008. *Introduction to Optical Mineralogy and Petrography - The Practical Methods of Identifying Minerals in Thin Section with the Microscope and the Principles Involved in the Classification of Rocks.* Read Books.

[4] Robert Ehrlich, Stephen K Kennedy, Sterling J Crabtree, and Robert L Cannon. 1984. Petrographic image analysis; I, Analysis of reservoir pore complexes. *Journal of Sedimentary Research* 54, 4 (1984), 1365–1378.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.

[6] Edward L Etris, David S Brumfield, Robert Ehrlich, and Sterling J Crabtree. 1988. Relations between pores, throats and permeability: a petrographic/physical analysis of some carbonate grainstones and packstones. *Carbonates and Evaporites* 3, 1 (1988), 17.

[7] B-H Juang and Lawrence R Rabiner. 1990. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Transactions on acoustics, speech, and signal Processing* 38, 9 (1990), 1639–1641.

[8] Kitty L Milliken, Lucy T Ko, Maxwell Pommer, and Kathleen M Marsaglia. 2014. SEM petrography of eastern Mediterranean sapropels: Analogue data for assessing organic matter in oil and gas shales. *Journal of Sedimentary Research* 84, 11 (2014), 961–974.

[9] Theodore T Mowers and David A Budd. 1996. Quantification of porosity and permeability reduction due to calcite cementation using computer-assisted petrographic image analysis techniques. *AAPG bulletin* 80, 3 (1996), 309–321.

[10] Fionn Murtagh. 1983. A survey of recent advances in hierarchical clustering algorithms. *Comput. J.* 26, 4 (1983), 354–359.

[11] T. Pang-Ning, Steinbach M., and Kumar V. 2006. *Introduction to data mining.* Addison-Wesley.

[12] Kenneth Pye and David H Krinsley. 1984. Petrographic examination of sedimentary rocks in the SEM using backscattered electron detectors. *Journal of Sedimentary Research* 54, 3 (1984), 877–888.

[13] Chandra L Reedy. 2006. Review of digital image analysis of petrographic thin sections in conservation research. *Journal of the American Institute for Conservation* 45, 2 (2006), 127–146.

[14] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[15] Ashok K Singh, Mamta Sharma, and Mahendra P Singh. 2013. SEM and reflected light petrography: A case study on natural cokes from seam XIV, Jharia coalfield, India. *Fuel* 112 (2013), 502–512.

[16] Sandra Neils Tonietto, Margaret Z Smoot, and Michael Pope. 2014. PS Pore Type Characterization and Classification in Carbonate Reservoirs. (2014).

[17] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.