Choix d'une méthode de classification pour la partition d'un massif rocheux à partir de TRE

Choice of a clustering method for the partition of a massif based on ERT

Habiba LHARTI[1*], *Colette* SIRIEX[1], *Joëlle* RISS[1], *Cécile* VERDET[1] et *Delphine* LACANETTE[1]

[1]Université de Bordeaux, CNRS, Arts et Métiers Institute of Technology, Bordeaux INP, INRAE, I2M Bordeaux, 351 cours de la Libération F-33405 Talence, France

Abstract. Clustering algorithms, as part of the machine learning family allow us to manage large amounts of datasets. The goal of this study is to compare two clustering methods, the Hierarchical Agglomerative Clustering (HAC) and the K-means method on electrical resistivity tomography (ERT) data characterising the massif surrounding the Lascaux cave. The results of these methods are analysed taking into account the *a priori* knowledge of the geology of the site. In this project, the structuring of the data into 2 to 5 classes aims to subdivide the massif into domains that are as homogeneous as possible in order to assign to each of them different thermal properties in a 3D model. These will be used as input parameters for the thermo-aerodynamic simulations of the cave under different climatic conditions.

Résumé. Les algorithmes de classification permettent de gérer de grandes quantités de données. L'objectif de cette étude est de comparer deux méthodes de classification, la Classification hiérarchique Ascendante (CHA) et la méthode des Centres Mobiles (CM) sur des données de Tomographie de Résistivité Electrique (TRE) caractérisant le massif entourant la grotte de Lascaux. Le rendu de ces méthodes est analysé en tenant compte de la connaissance a priori de la géologie du site. Dans ce projet, la structuration des données en 2 à 5 classes vise à subdiviser le massif en domaines les plus homogènes possible afin de leur affecter des propriétés thermiques différentes dans un modèle 3D. Ces dernières serviront de paramètres d'entrée nécessaires aux simulations thermo-aérauliques de la grotte soumise à différentes conditions météorologiques.

* Habiba LHARTI: habiba.lharti@u-bordeaux.fr

1 Introduction

La grotte de Lascaux, située sur la commune de Montignac en Dordogne (24), constitue l'une des plus célèbres découvertes archéologiques du XXème siècle. Découverte en 1940 et officiellement ouverte au public en 1948, elle accueille un grand nombre de visiteurs jusqu'en 1966. Cette affluence a pour conséquence l'apparition de plusieurs indices d'altération (augmentation de température, moisissures, condensation, etc.). Afin de répondre à ces problématiques, le ministère de la culture crée un comité scientifique international en 2002 pour mener des recherches visant à la conservation du patrimoine de la grotte, en partie sur l'aspect thermique du site.

Le massif de la grotte de Lascaux a été considéré comme homogène du point de vue des caractéristiques thermo-physiques dans le modèle 3D de Lacanette et Malaurent [1], modèle qui est utilisé dans les simulations thermo-aérauliques de la grotte. Les études géophysiques [2] – [4] montrent que le massif est hétérogène. L'objectif de cette étude est de déterminer des ensembles les plus homogènes possible à partir des données de TRE, ensembles pour lesquels les propriétés thermiques pourront être estimées ou mesurées. Pour ce faire, les résultats de deux méthodes de classification (CHA et CM) sont comparés.

2 Site d'étude et méthodes de classification

Les données de tomographie électrique utilisées dans cette étude sont issues des mesures réalisées à la fin de mois de mars 2013 par [3]. Les profils retenus pour cette étude sont les suivants (Fig. 1): 8 d'orientation ouest-est avec 96 électrodes espacées de 1,5 m ; 1 d'orientation sud-ouest/nord-est de 72 électrodes espacées de 1,5 m ; 3 d'orientation nord/sud avec 96 électrodes espacées de 1,5 m ; 6 profils nord-ouest/sud-est dont 2 avec 96 électrodes espacées de 1,5 m et 4 de 72 électrodes espacées respectivement de 0,50 m et 1,0 m. Le nombre total de blocs résultant des inversions avec RES2INV (4.05.38) et auxquels une résistivité est associée est égal à 53800.

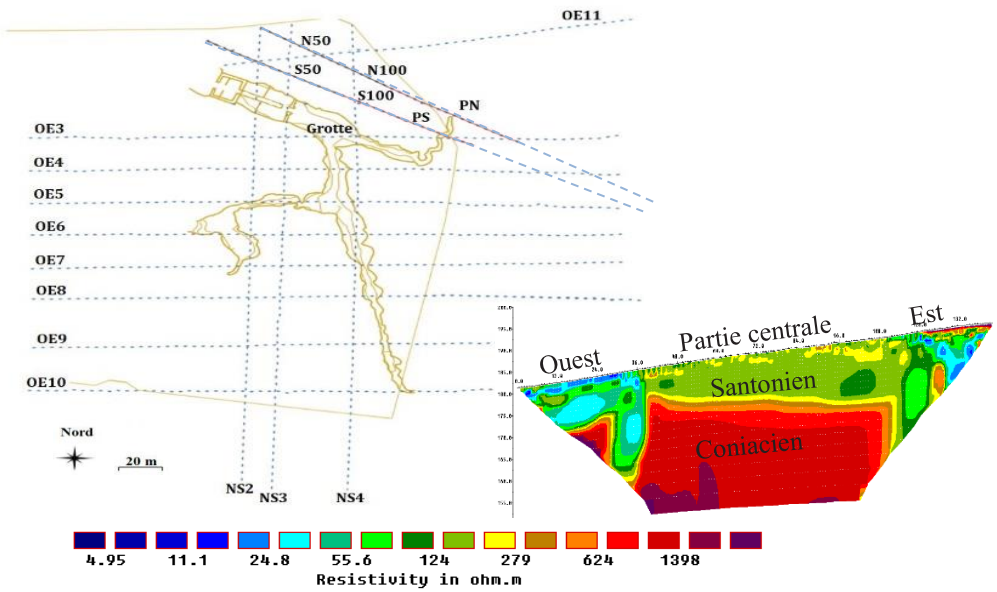

Figure 1. Localisation des profils ouest/est, nord/sud et nord-ouest/sud-est par rapport à la géométrie de la grotte de Lascaux. En bas à droite, un exemple de profils : le profil OE5.

La Classification Hiérarchique Ascendante (CHA) et la méthode des centres mobiles (CM) sont deux méthodes fréquemment utilisées en géophysique. La CHA est une méthode déterministe très utilisée pour la segmentation des données issues de la TRE [2], [5], [6]. Elle permet de construire une partition d'un ensemble de données en fonction de leur matrice de proximité (méthode de Ward) [7]. Le choix du nombre de classes se fait postérieurement à la classification.
La méthode CM, connue sous le nom de K-means, est considérée comme la méthode statistique la mieux adaptée pour l'analyse des données [8] – [10]. L'algorithme consiste à réaliser des itérations afin de partitionner l'ensemble des données en classes dont le nombre est choisi au préalable (le centre des classes peut être choisi aléatoirement ou arbitrairement). Le principe de cette méthode repose sur la minimisation, dans chaque classe, de la distance entre les individus et leur centre de gravité.

3 Résultats

Les résultats déjà obtenus par [3] montrent que la formation géologique la plus profonde (Coniacien), dans laquelle se développe essentiellement la grotte [11], est nettement plus résistive que les parties superficielles (Fig. 1). Par ailleurs, [4] ont montré que, localement et au-dessus de la grotte, la limite supérieure du Coniacien se situait entre 179 et 182 m NGF.

Par ailleurs les résistivités du Coniacien sont supérieures à 209 Ω.m. Cette formation est donc considérée, dans la suite comme un ensemble homogène constituant une classe déterminée a priori, représentée en rouge sur les images ci-dessous. Les données du Coniacien seront donc exclues des classifications. Cela diminue le nombre total de valeurs de résistivité considérées pour les classifications à 37101 comprises entre 4 Ω.m et 7979 Ω.m.

Les classifications ont été effectuées (Minitab® 20.2.) en travaillant avec les valeurs logarithmiques de la résistivité du Santonien (partie supérieure du massif) et des formations détritiques à l'est et à l'ouest qui ont été décrites comme très hétérogènes (Fig. 1). Les méthodes sont utilisées en faisant varier le nombre de classes (de 2 à 5) et en excluant le Coniacien.

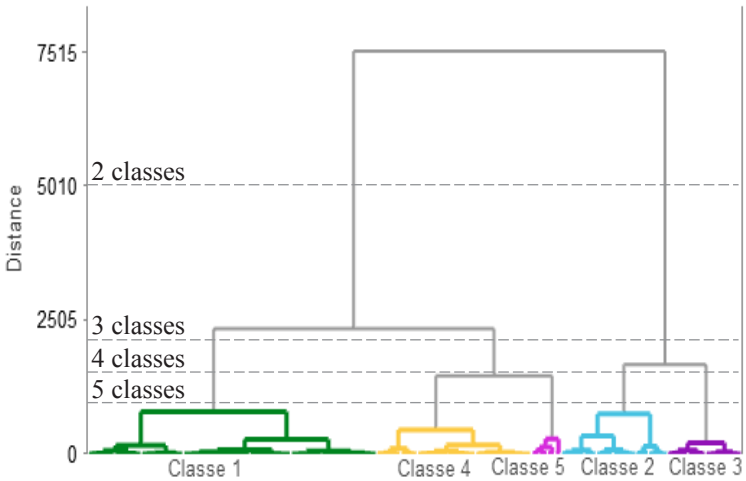

Figure 2. Dendrogramme issu de la CHA avec les coupures à 2, 3, 4 et 5 classes.

L'analyse de l'ensemble des données (notée CHA globale) est représentée sous forme de dendrogramme (Fig.2). Les classes qui représentent des intervalles de résistivité sont portées sur l'axe des abscisses et les distances auxquelles les classes se regroupent sont portées sur l'axe des ordonnées. Le dendrogramme permet l'identification de deux grandes classes bien distinctes compte tenu de la distance à laquelle elles sont finalement réunies ; la première classe (cf. classe verte figure 3.a) rassemble deux sous-classes (1 et 4) avec un CV de 139% et une médiane de 151 Ω·m, et la deuxième classe (cf. classe bleue figure 3.a) regroupe deux sous-classes (2 et 3) avec un CV de 46% et une résistivité médiane de 47 Ω·m.

La méthode CM permet également de distinguer deux classes, la première (cf. classe verte figure 3.b) présent un CV de 140% avec une résistivité médiane égale à 148 Ω·m, ce qui n'est pas significativement différent de celle obtenue avec la CHA, la deuxième classe a un CV de 44% et une médiane de 42 Ω·m.

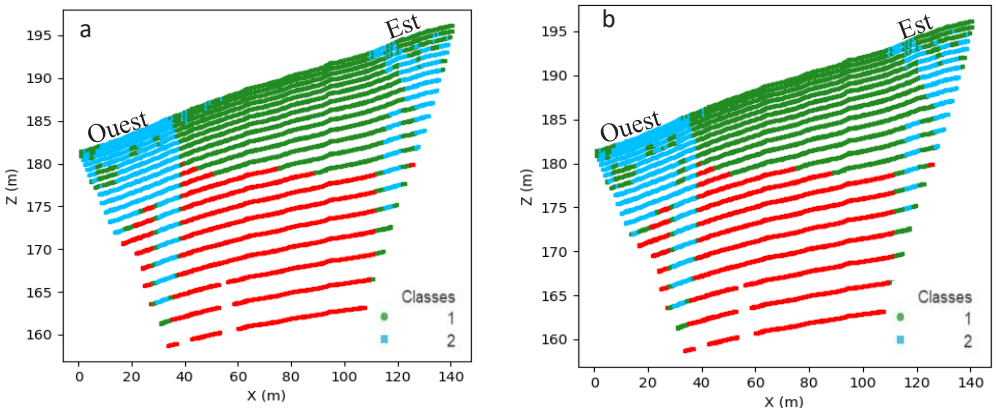

Figure 3. Profil OE5 (classification à 2 classes, le Coniacien est exclu) **a**. méthode CHA **b**. méthode des centres

Les deux méthodes CHA et CM (Fig.3) permettent de retrouver les deux formations connues de la géologie du site : le Santonien qui est représenté par la première classe pour chacune des deux méthodes et les formations détritiques qui le sont par la classe 2. La méthode CM montre une distribution spatiale légèrement différente de celle de la CHA en ce qui concerne les formations détritiques (classe 2) en relation avec une limite de la résistivité entre les deux classes égale à 90 Ω·m avec la CHA et à 79 Ω·m avec les CM.

Compte tenu de notre objectif (prise en compte des hétérogénéités, en particulier du Santonien) une classification à deux classes quel que soit la méthode est insuffisante. Ainsi, afin de mieux circonscrire la distribution spatiale des domaines, les classifications ci-dessous sont réalisées en faisant varier le nombre de classes entre 3 et 5 (Tableau 1).

Le tableau 1 présente les médianes, les valeurs minimales et maximales des résistivités et les coefficients de variation par classe obtenus avec l'ensemble des profils (37101 données) ; ces classifications sont qualifiées de « globales ». Les résultats sont ensuite analysés en prenant comme exemple l'un des profils : le profil OE 5.

Tableau 1 Valeurs de la médiane des résistivités, des limites inférieure et supérieure et du coefficient-de variation (CV : Rapport de l'écart type à la moyenne) pour chaque classe issue de la CHA et des CM.

	Statistiques par classes issues de la CHA et CM globale (37101 valeurs de résistivité en Ω·m)						
Méthode	Nbre de classes		Classe1	Classe2	Classe3	Classe4	Classe5
CHA	3	Médiane	134	47	197	-	-
		Min	90	4	162		
		Max	162	90	7979		
		CV	14%	46%	136%		
	4	Médiane	134	33	73	197	-
		Min	90	4	55	162	
		Max	162	55	90	7979	
		CV	14%	37%	14%	136%	
	5	Médiane	134	33	73	190	603
		Min	90	4	55	162	345
		Max	162	55	90	345	7979
		CV	14%	37%	14%	19%	94%
CM	3	Médiane	143	510	39	-	-
		Min	71	288	4		
		Max	288	7979	71		
		CV	28%	102%	42%		
	4	Médiane	78	580	29	152	-
		Min	45	324	4	109	
		Max	109	7979	45	324	
		CV	24%	97%	33%	24 %	
	5	Médiane	134	667	64	191	27
		Min	90	405	39	162	4
		Max	162	7979	90	404	39
		CV	14%	88%	23%	24%	31%

3.1 Classification à 3 classes

La figure 4 montre des différences dans la distribution spatiale des classes. Pour la partie superficielle, la classe 1 (en vert) pour la CHA présente un CV de 14% et une résistivité médiane de 134 Ω·m. Pour la méthode des CM, la classe 1 (en vert) présente un CV plus élevé (28%) et une médiane de 143 Ω·m donc une plus grande dispersion que la CHA. Avec 3 classes, la méthode CM ne permet pas de distinguer les hétérogénéités mises en valeur par la CHA et présentes sur la TRE dans le Santonien au-dessus de la grotte.

Les différences à l'ouest sont dans le nombre de blocs légèrement plus important avec la CHA (classe 2, en bleu) qui regroupe environ 520 blocs et une résistivité médiane de 47 Ω·m tandis que la méthode CM (classe 3) regroupe 452 blocs avec une résistivité 39 Ω·m.
Pour la partie est du profil, les blocs jaunes en surface correspondent à des formations détritiques sableuses connues avec une médiane de 510 Ω·m pour les CM, mais elles sont, avec la CHA à 3 classes, confondues avec les hétérogénéités de la partie centrale.

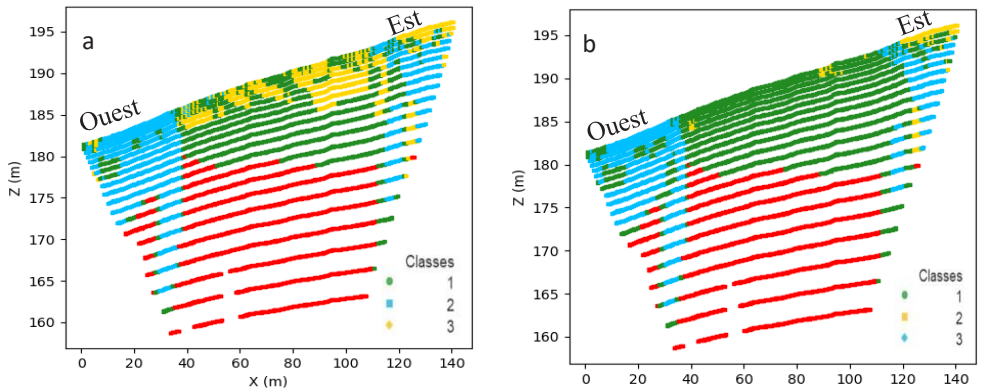

Figure 4. Profil OE5 (classification à 3 classes) **a**. Méthode CHA **b**. Méthode des centres mobiles

Les intervalles de répartition diffèrent d'une méthode à l'autre, mais dans le cadre de notre projet, la caractérisation du massif n'est pas encore satisfaisante avec la méthode des CM qui ne permet pas de distinguer les hétérogénéités centrales du Santonien nécessaire à la modélisation thermique par contre la CHA ne permet pas de distinguer les sables de la partie est, ce que fait la méthode des CM.

3.2 Classification à 4 classes

La figure 5 montre les résultats des classifications à 4 classes qui, dans le cadre de la CHA, apporte des détails dans les zones est et ouest et maintient les limites des classes pour le Santonien. La répartition spatiale pour la partie centrale (zones vertes et jaunes) en CHA n'évolue pas significativement par rapport à la classification précédente (Fig.4) car, entre autres, le coefficient de variation (CV) reste égal à 14%. Pour les CM, les hétérogénéités ne sont toujours pas visibles mais les limites de répartition ont varié. La limite inférieure de la classe 4 (en vert) a augmenté de 38 Ω·m et la limite supérieure a augmenté de 36 Ω·m par rapport au cas précédent tandis que le CV a légèrement diminué (de 28% à 24%) ce qui traduit une diminution de la dispersion des données.

Pour les parties est et ouest, les zones bleues présentent la même résistivité minimale mais la valeur maximale a diminué de 29 Ω·m pour les CM et de 35 Ω·m pour la CHA. Les CV ont diminué dans les deux cas de 46% à 37% pour la CHA et de 42% à 33% pour les CM. L'augmentation du nombre des classes implique en effet une diminution de la dispersion des données.

Pour la partie ouest, une nouvelle classe représentée en violet apparait avec 150 blocs de résistivité médiane de 73 Ω·m pour la CHA et 244 et 78 Ω·m pour les CM. La distribution spatiale de la zone bleue est très similaire avec un nombre de blocs égale à 365 pour la CHA et à304 pour les CM. A l'est du profil, la segmentation des classes est très semblable pour l'une et l'autre des deux méthodes hormis les zones jaunes en surface qui regroupent toujours plus de blocs pour la CHA que pour les CM.

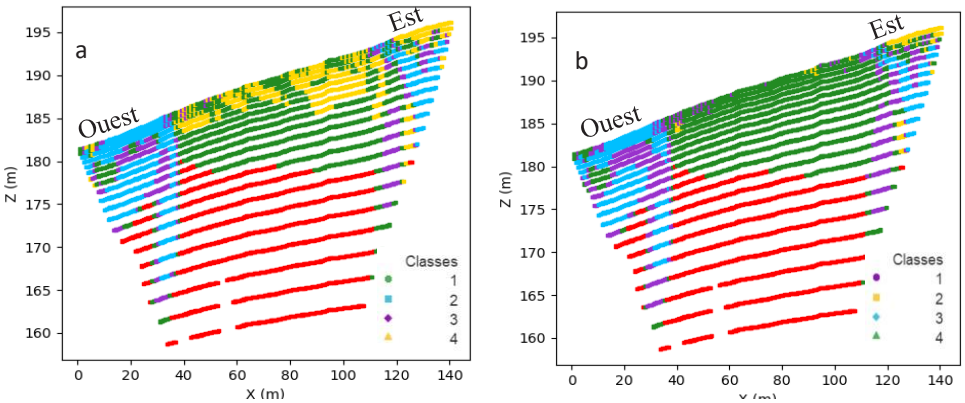

Figure 5. Profil OE5 (classification 4 classes) **a**. Méthode CHA **b**. Méthode des centres mobiles

La classification à 4 classes apporte une subdivision dans les formations détritiques tant pour la CHA que les CM. Par contre, avec les CM et pour la partie centrale les zones hétérogènes du Santonien ne sont toujours pas visibles mais, le CV a diminué par rapport à la classification à 3 classes.

3.3 Classification à 5 classes

Avec 5 classes (Fig.6), la classe 1 en vert présente les mêmes limites inférieure et supérieure quelle que soit la classification et la même valeur du coefficient de variation (14%). Pour identifier les zones hétérogènes du Santonien, il a fallu une classification à 5 classes avec les CM alors que 3 suffisent avec la CHA.

Pour les zones bleues, la limite inférieure des résistivités reste naturellement identique pour les deux méthodes, mais la limite supérieure présente une diminution de 6 Ω·m pour les CM par rapport à la classification à 4 classes et un CV légèrement plus faible, alors que, avec la CHA, elle reste inchangée. Les zones violettes se caractérisent par un coefficient de variation égal à 14% avec la CHA plus faible que celui (23%) obtenu avec les CM ce qui est dû à une limite inférieure plus faible pour les CM que pour la CHA.

Les résistivités des zones en magenta ont une limite inferieure qui diffère de 60 Ω·m (345 Ω·m avec la CHA et 405 Ω·m avec les CM), les limites supérieures étant naturellement les

mêmes, par suite le coefficient de variation est légèrement inférieur pour les CM (88% pour les CM et 94% pour la CHA).

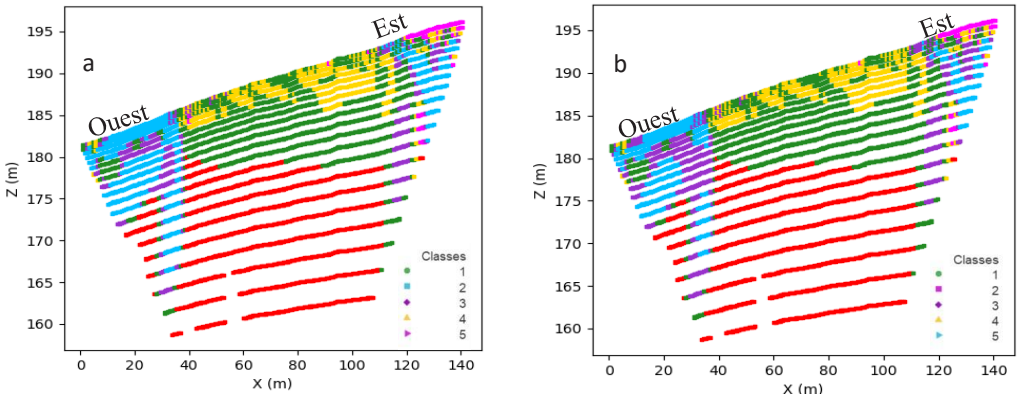

Figure 6. Profil OE5 (classification 5 classes) **a**. Méthode CHA **b**. Méthode des centres mobiles

Les deux méthodes convergent vers une partition de l'espace très ressemblante mais avec cependant des différences dans les limites des intervalles de résistivité de 4 des 5 classes. Par contre, les limites des résistivités de la partie centrale (classe représentée en vert) sont identiques d'une classification à l'autre (Fig.6). Cependant, une partition à 5 classes a été nécessaire pour mettre en évidence les hétérogénéités dans le Santonien avec les CM, hétérogénéités observées précédemment dès la CHA à 3 classes. On note, toutefois, pour les hétérogénéités du Santonien, un CV légèrement plus élevé avec la CHA (24%) qu'avec les CM (19%) traduisant une dispersion des valeurs légèrement plus importante.

4 Conclusion

Les algorithmes de classification permettent de partitionner un grand nombre de données rapidement. Dans le cadre de cette étude, deux méthodes ont été comparées (CHA et CM) afin d'obtenir la solution qui concorde le mieux avec la géologie du site.

Les deux méthodes à deux classes permettent de retrouver les deux grandes formations géologiques de l'environnement de la grotte de Lascaux (formations détritiques et calcaire Santonien) avec de légères différences dans la délimitation des formations détritiques. Cependant, les images de TRE présentent des hétérogénéités dans le Santonien qui ne sont pas identifiées avec cette partition. Pour les faire apparaître, un plus grand nombre de classes s'est avéré nécessaire: trois avec la CHA si l'intérêt se porte essentiellement sur la partie centrale, puis 4 (voire 5) si l'intérêt se porte aussi sur les hétérogénéités des formations détritiques. Par contre, avec la méthode des CM, il faut une partition à cinq classes pour que l'ensemble des hétérogénéités apparaissent.

Notre projet porte sur les zones localisées au-dessus de la grotte, donc aux hétérogénéités présentées dans la TRE du Santonien. La CHA à partir de 3 classes correspond à nos attentes compte tenu de l'hétérogénéité connue de cette formation et de la concordance avec les données géologiques du site.

L'utilisation de ces méthodes de classification est amenée à se développer en géophysique afin de permettre une interprétation plus rapide et assurer aussi la production d'images plus

synthétiques de sites d'étude. De plus, des codes sont désormais en libre accès sur de nombreux logiciels gratuits favorisant leur utilisation. Le choix de la méthode de classification doit donc être étudié de façon plus systématique afin d'utiliser la méthode la plus opportune en fonction des objectifs fixés mais aussi de la connaissance préalable du site étudié.

Références

1. D. Lacanette et P. Malaurent, « Prévision climatique 3D dans la grotte de Lascaux »,
2. *karstologia*, vol. 63, p. 49-57, 2014.
3. S. Xu, C. Sirieix, J. Riss, et P. Malaurent, « A clustering approach applied to time-lapse ERT interpretation — Case study of Lascaux cave », *Journal of Applied Geophysics*, vol. 144, p. 115-124, sept. 2017, doi: 10.1016/j.jappgeo.2017.07.006.
4. S. Xu, C. Sirieix, A. Marache, J. Riss, et P. Malaurent, « 3D geostatistical modeling of Lascaux hill from ERT data », *Engineering Geology*, vol. 213, p. 169-178, nov. 2016, doi: 10.1016/j.enggeo.2016.09.009.
5. C. Verdet, C. Sirieix, A. Marache, J. Riss, et J.-C. Portais, « Detection of undercover karst features by geophysics (ERT) Lascaux cave hill », *Geomorphology*, vol. 360, p. 107177, juill. 2020, doi: 10.1016/j.geomorph.2020.107177.
6. F. Genelle, C. Sirieix, J. Riss, et V. Naudet, « Monitoring landfill cover by electrical resistivity tomography on an experimental site », *Engineering Geology*, vol. 145-146,p. 18-29, sept. 2012, doi: 10.1016/j.enggeo.2012.06.002.
7. D. Delforge, « Time-series clustering approaches for subsurface zonation and hydrofacies detection using a real time-lapse electrical resistivity dataset », *Journal of Applied Geophysics*, p. 15, 2021.
8. J. H. Ward, « Hierarchical Grouping to Optimize an Objective Function », *Journal of the American Statistical Association*, vol. 58, n° 301, p. 236-244, mars 1963, doi: 10.1080/01621459.1963.10500845.
9. Y. Wang, M. Wang, Y. Yan, et S. Gong, « A method to recognize the contaminated area using K-means in ERT contaminated site surveys », in *2018 IEEE International Conference on Information and Automation (ICIA)*, Wuyishan, China, août 2018, p. 1587-1591. doi: 10.1109/ICInfA.2018.8812583.
10. Kutbay Ugurhan et al., « Underground electrical profile clustering using K-MEANS algorithm », juin 2015.
11. D. Delforge, A. Watlet, O. Kaufmann, M. Van Camp, et M. Vanclooster, « Time-series clustering approaches for subsurface zonation and hydrofacies detection using a real time-lapse electrical resistivity dataset », *Journal of Applied Geophysics*, vol. 184, p. 104203, janv. 2021, doi: 10.1016/j.jappgeo.2020.104203.
12. J.-P. Platel, «Le Crétacé supérieur de la plate-forme septentrionale du Bassin d'Aquitaine. Stratigraphie et évolution géodynamique », Document B.R.G.M, vol. 164, p. 581, 27 nov, 1987.