Design, Implementation and Analysis of a Description Model for Complex Archaeological Objects

Aybuke Ozturk

N° d'ordre NNT : 2018LYSE2048

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512 Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 9 juillet 2018, par :

Aybüke ÖZTÜRK

Design, Implementation and Analysis of a Description Model for Complex Archaeological Objects

Devant le jury composé de :

Henning CHRISTIANSEN, Professeur, Roskilde University (Danemark), Président

Nicole VINCENT, Professeure des universités, Université Paris Descartes, Rapporteure

François PINET, Directeur de recherche, Inst. Nat.Rech. en Sces et Tech.environnmt et Agricuult, Rapporteur

Zoï TSIRTSONI, Chargée de recherche, C.N.R.S., Examinatrice

Jérôme DARMONT, Professeur des universités, Université Lumière Lyon 2, Co-Directeur de thèse

Stéphane LALLICH, Professeur des universités, Université Lumière Lyon 2, Co-Directeur de thèse

Sylvie Yona WAKSMAN, Chargée de recherche, C.N.R.S, Co-Directrice de thèse

Contrat de diffusion

Thèse de doctorat de l'Université de Lyon
opéré par
l'Université Lumière Lyon 2

Laboratoire ERIC (EA 3083)
Laboratoire ArAr (UMR 5138)
École Doctorale Informatique et Mathématiques (ED 512)

présentée pour obtenir le grade de
Docteur en Informatique

Élaboration, mise en œuvre et analyse d'un modèle de description d'objets archéologiques complexes

Par: Aybüke ÖZTÜRK SURI

Présentée et soutenue publiquement le 9 Juillet 2018, devant un jury composé de :

Nicole VINCENT, Professeure des universités, Université Paris Descartes	Rapportrice
François PINET, Directeur de Recherche, IRSTEA Clermont-Ferrand	Rapporteur
Henning CHRISTIANSEN, Professeur, Université Roskilde, Danemark	Examinateur
Zoï TSIRTSONI, Chargée de recherche, CNRS Paris	Examinatrice
Jérôme DARMONT, Professeur des universités, Université Lumière Lyon 2	Directeur
Stéphane LALLICH, Professeur des universités émérite, Université Lumière Lyon 2	Directeur
Sylvie Yona WAKSMAN, Chargée de recherche, CNRS Lyon	Directrice

Résumé long de la thèse

L'archéologie est l'étude du passé des hommes à travers les vestiges matériels. Les céramiques sont parmi les artéfacts archéologiques les plus abondants, et fournissent des informations sur de nombreux aspects de l'activité humaine, notamment la chronologie, les échanges commerciaux et la technologie. Ces dernières années, on a pu assister à une forte croissance et une plus grande disponibilité de divers données et réseaux archéologiques. Dans le même temps, les systèmes et outils numériques ont permis une utilisation accrue des données par un grand nombre d'utilisateurs potentiels allant des étudiants aux chercheurs et des conservateurs de musées aux touristes.

En outre, l'évolution des techniques scientifiques et statistiques a également contribué à mieux comprendre les matériaux archéologiques, par exemple les objets céramiques, les coordonnées géographiques et la photographie numérique. Cependant, il n'existe actuellement pas beaucoup de systèmes numériques polyvalents, ni d'outils et de bases de données qui peuvent être facilement utilisés par les archéologues pour étudier des informations variées et les partager. De plus, les céramiques peuvent être utilisées pour déterminer des relations contextuelles, ce qui aide à mettre en évidence les données significatives sur le plan archéologique à partir d'une masse de données individuelles.

En d'autres termes, l'exploitation des données céramologiques permet de découvrir des motifs qui ne sont visibles que sur de larges corpus. En archéologie, les données sont très contextualisées. Ainsi, les céramiques et leurs propriétés peuvent-elles aider à acquérir des connaissances approfondies sur des questions technologiques, culturelles et géographiques, à travers des informations sur l'époque et la provenance de la céramique. En outre, les informations stockées dans les bases de données se concentrent généralement sur une gamme limitée de descripteurs céramologiques et ne sont pas interopérables.

Au cours du processus de documentation d'un site de fouilles, les archéologues tendent à intégrer toutes les données de façon cohérente pour interpréter les matériaux archéologiques afin de mieux comprendre les cultures humaines. Dans ce processus, la construction de ressources réutilisables pour l'étude de la céramique est importante. À partir de là, quelques questions fondamentales sont posées, telles que le lieu et le moment où elles ont été produites, comment elles ont été fabriquées et quelle était leur fonction.

C'est ainsi que les données céramologiques brutes et induites peuvent être classées en trois niveaux. Dans le premier niveau, les données sont directement accessibles à partir de l'objet céramique et de son contexte, par exemple la décoration de l'objet et l'emplacement où celui-ci a été trouvé. Ces données sont le plus souvent stockées sans aucune modification ultérieure dans les bases de données. Au second niveau, les données nécessitent un premier degré d'interprétation, notamment sous forme d'hypothèses, comme l'origine supposée d'un objet trouvé sur un site donné, et les analyses scientifiques réalisées pour tester ces hypothèses. Par exemple, le type d'un

objet en céramique est un premier niveau de données et peut être utilisé pour supposer une origine (avant toute analyse), c'est-à-dire une donnée de localisation.

Au troisième niveau, les données sont un résultat, comme l'attribution d'un objet à une origine en fonction d'analyses scientifiques et éventuellement d'autres critères. Par exemple, l'attribution (après analyse) d'une céramique à son origine peut être déduite suite à des analyses pétrographiques ou chimiques.

En raison des besoins de la recherche actuelle, la gestion de données présente certains défis. Trouver des informations utiles dans d'énormes quantités de données très contextualisées est difficile pour les chercheurs et les étudiants. Les données sont globalement très hétérogènes. Les bases de données ont différents formats de fichiers, protocoles d'accès et utilisent différents langages de requête. Il n'y a pas de système de classification commun, ni de terminologie normalisée, qui sont nécessaires pour comprendre les relations à partir des interconnexions. L'interopérabilité est également limitée, avec des bases de données fournissant uniquement une interface web, mais pas d'API (Application Programming Interface).

Ainsi, combiner diverses informations sur des objets archéologiques, tels que des documents textuels, numériques et graphiques, qui permettraient de puissantes analyses informatiques, est au mieux une tâche complexe à ce jour. Le défi de la recherche est d'intégrer différentes dimensions à partir de bases de données distantes qui décrivent les mêmes catégories d'objets de manière complémentaire. Ainsi, nous visons à concevoir des méthodes d'entreposage et d'exploration de données qui aident à mieux analyser et catégoriser les objets complexes. Cette thèse est divisée en deux parties complémentaires. La première partie a trait à la modélisation de données archéologiques complexes, alors que la seconde partie porte sur la classification non supervisée de données archéologiques complexes.

Dans la première partie de la thèse, nous examinons d'abord une sélection de bases de données archéologiques et archéométriques relatives aux céramiques que nous considérons comme représentatives de la diversité des contenus, des formats, des statuts et des caractéristiques. En outre, nous présentons les entrepôts de données archéologiques existants (Chapitre 2).

Par exemple, le projet Levantine Ceramics Project (LCP), dirigé par l'Université de Boston, est une base de données archéologiques centrée sur les céramiques produites au Levant, du Néolithique à l'époque ottomane. Il comprend principalement des données archéologiques (typologiques, chronologiques et géographiques), mais fournit également des données d'analyse pétrographique. Les données LCP sont en format texte et image. Le LCP est une ressource Internet interactive et ouverte[1].

La base de données du MURR Archaeometry Laboratory[2] construite à l'Université du Missouri présente des analyses chimiques d'artefacts en céramique de nombreuses régions, comprenant l'Amérique du Nord, l'Amérique Centrale et l'Amérique du Sud ainsi que la Méditerranée.

[1] `https://www.levantineceramics.org/`

[2] `http://archaeometry.missouri.edu/datasets/datasets.html`

Couvrant un large éventail de périodes et de régions, la base de données Ceramo du Laboratoire d'Archéologie et d'Archéométrie (ArAr) de Lyon[3] était à l'origine principalement une base de données chimiques qui ne contenait que peu d'informations archéologiques. Elle est actuellement développée pour inclure davantage d'informations, notamment sous forme d'images 2D et 3D. Nous présentons la nouvelle base de données Ceramo 3.0 et détaillons sa conception, qui répond aux exigences des spécialistes (Chapitre 3).

Ceramo 3.0 est divisé en trois paquetages principaux, dont les classes et les attributs comprennent plusieurs documents graphiques, des données de localisation, des définitions précises des échantillons céramiques et différents résultats d'analyse. Dans le premier paquetage, nous affichons des informations géographiques telles que PROVENANCE, ORIGINE SUPPOSÉE et ATTRIBUTION. La classe PROVENANCE apporte des informations relatives au lieu où l'objet a été trouvé. La classe ORIGINE SUPPOSÉE fournit une origine supposée avant l'analyse. La classe ATTRIBUTION indique l'origine de l'objet après l'analyse.

Dans le deuxième paquetage, nous affichons des informations d'état et de description telles que la classe DESCRIPTION qui présente les descripteurs textuels d'un objet. La classe DATATION stocke les données de datation des objets, à la fois au niveau général et à un niveau précis. Le troisième paquetage contient les résultats de différents types d'analyse en laboratoire. Par exemple, la classe CHIMIE rassemble les résultats d'analyse chimique d'un objet. Cette nouvelle base de données stocke ainsi des données complexes. En outre, des applications ont été conçues pour fonctionner avec Ceramo permettant la mise à jour, l'interrogation des données et l'utilisation de traitements statistiques. Cependant, une nouvelle tendance en archéologie est de construire des entrepôts de données [1], qui sont des bases de données analytiques.

Les entrepôts de données comportent un modèle multidimensionnel spécifique qui permet l'analyse en ligne (OnLine Analytical Processing ou OLAP). Par exemple, pour analyser l'énorme quantité de données liées à la civilisation chinoise ancienne, l'Université de Chine du Nord travaille à la construction d'un entrepôt de données distribué, qui aide à gérer, partager et analyser les informations relatives à l'antiquité [2]. Des chercheurs du Département d'histoire et du Centre de Recherche en Géomatique de l'Université Laval (Québec) ont travaillé à résoudre le problème de l'enregistrement et de l'analyse des données de fouilles archéologiques en utilisant un système basé sur les systèmes d'information géographiques (SIG). En général, les SIG aident à enregistrer, analyser et visualiser les données spatiales. Ici, le SIG a contribué à la construction d'un système intégré d'exploration archéologique (ISAE) qui prend en charge les analyses multicritères [3].

Nous avons également travaillé sur la façon dont les données céramologiques peuvent être entreposées pour permettre l'OLAP (Chapitre 4). De telles analyses aident à naviguer et à observer les données selon différentes perspectives, fournissant ainsi

[3] `http://www.arar.mom.fr/qui-sommes-nous/laboratoire-de-ceramologie/les-bases-de-donnees`

aux chercheurs un meilleur aperçu de leurs données. Le principal avantage de cette approche est d'identifier les motifs cachés. Dans un entrepôt de données, les données observées sont appelées faits, par exemple les ventes dans un contexte commercial. Ils sont caractérisés par des mesures qui sont généralement numériques, par exemple les quantités vendues et les montants correspondants. Les faits sont observés suivant différents axes d'analyse appelés dimensions, par exemple les produits vendus, l'emplacement du magasin et la date de vente. Ainsi, les schémas d'entrepôt de données sont appelés schémas multidimensionnels. Pour permettre la navigation OLAP dans les données de Ceramo, nous devons sélectionner les faits à observer, les axes d'analyse (dimensions) et importer les données dans l'entrepôt de données.

Le résultat est appelé un cube (un hypercube lorsque le nombre de dimensions est supérieur à 3), où les valeurs des dimensions sont des coordonnées qui définissent une cellule de fait. Dans un scénario type, nous choisissons d'observer les groupes de céramiques résultant d'analyses chimiques par rapport à la provenance, la datation, la description. Dans l'analyse, nous utilisons des fonctions d'agrégation pour analyser en profondeur les données relatives aux céramiques. Les résultats montrent comment OLAP peut contribuer à la compréhension des relations économiques et culturelles à une période spécifique, grâce à sa capacité à analyser l'information selon différents points de vue. De plus, les modèles que nous proposons peuvent facilement être adaptés à d'autres domaines d'application, par exemple l'économie ou la médecine, qui partagent des problèmes similaires de modélisation et d'analyse de données. Ces contributions ont été publiées dans les actes de la 9e conférence internationale interdisciplinaire Modeling and Using Context (CONTEXT 2015) [4].

Dans la deuxième partie de la thèse, nous nous concentrons sur le clustering (classification non supervisée). Le clustering est un domaine de recherche qui appartient aux domaines de la fouille de données (data mining) et de l'apprentissage automatique (machine learning) (Chapitre 5). Le clustering permet de regrouper un ensemble de points de données (occurrences) non étiquetés décrits par des attributs (variables), de sorte que les points d'un même cluster (groupe) ont des caractéristiques similaires, tandis que les points de différents clusters ont des caractéristiques différentes. Il existe différentes catégories de clustering. L'un des critères de classification pour le clustering est la gestion du chevauchement des clusters. En clustering dur, un point appartient à un groupe et un seul, alors que dans un clustering flou [5], un point peut appartenir avec plus ou moins d'intensité à plusieurs clusters. Le clustering flou est très utile dans de nombreuses applications, notamment dans la catégorisation textuelle de diverses informations en différents groupes. Par exemple, si l'on considère trois groupes ayant trait respectivement à l'économie, l'énergie et la politique, le mot clé "pétrole" est susceptible de renvoyer à chacun des trois groupes. En outre, il est également possible d'ouvrir des discussions avec les experts du domaine lorsque l'on analyse les résultats d'un clustering flou. Il existe plusieurs méthodes de clustering flou, comme les C-Means flous (FCM) ou les Fuzzy K-Medoids (FKM).

De nombreux archéologues utilisent des méthodes issues de l'informatique et de la

statistique pour étudier les données générées au cours des différentes phases de leur recherche, avant, pendant et après les fouilles archéologiques. Par exemple, l'analyse discriminante (AD) [6] est une technique d'apprentissage supervisé utilisée lorsque deux groupes ou plus sont connus a priori et qu'une ou plusieurs nouvelles observations doivent être attribuées à l'un des groupes connus en fonction des caractéristiques mesurées. Une autre technique est l'analyse des correspondance (AFC) [7], qui est utilisée pour comprendre le lien entre variables catégorielles (plutôt que continues). Dans le laboratoire ArAr de Lyon, les archéomètres définissent des groupes d'objets céramiques en se basant en premier lieu sur leur composition chimique. Pour déterminer l'origine des objets, ils s'appuient sur des classifications hiérarchiques (méthode ascendante) et sur des analyses discriminantes appliquées aux données chimiques [8] [9].

Pour effectuer un bon clustering, plusieurs critères doivent être pris en compte, parmi lesquels le choix de la méthode de clustering, la procédure d'initialisation, le choix du nombre de clusters et la recherche d'outils efficaces pour évaluer la qualité des résultats obtenus. De plus, pour obtenir des clusters stables, on doit souvent gérer des données de types différents (hétérogènes). Cette hétérogénéité est communément rencontrée dans les applications de fouille de données en sciences humaines et sociales, notamment en archéologie et en archéométrie.

Pour ces raisons, nous présentons d'abord des images de matériaux céramiques (fabrics), puis les méthodes utilisées dans la littérature pour la détection des caractéristiques des images (Chapitre 6). Les "fabrics" correspondent aux caractéristiques des matériaux céramiques telles qu'elles peuvent être observées à l'œil nu ou à l'aide d'une loupe binoculaire. Elles comportent deux composantes principales : la matrice et les inclusions. Pour exploiter les images correspondantes lors d'un clustering, nous avons choisi d'utiliser la couleur des inclusions comme caractéristique. Ceci peut être obtenu plus précisément en utilisant des méthodes de détection de couleur plutôt qu'à l'œil nu. Cette caractéristique peut aider à définir la similarité entre les céramiques, en construisant des groupes d'objets cohérents. La plupart de ces images ont une couleur de fond. Pour cette raison, nous appliquons d'abord la méthode de segmentation d'image de MathWorks Image Processing Toolbox[4] pour détecter un objet entier. Cependant, seuls quelques objets sont correctement détectés, car l'objet et les couleurs d'arrière-plan sont trop similaires. Ainsi, nous ajoutons un masque créé manuellement avec la fonction Roipoly à partir de MathWorks Image Processing Toolbox. Cette fonction permet de sélectionner la région d'intérêt manuellement. Ensuite, pour détecter la couleur, nous appliquons une méthode de segmentation fondée sur les couleurs, initialement conçue pour les images médicales [10], qui repose sur un clustering obtenu à l'aide des K-Means.

L'approche est subdivisée en trois étapes. La première étape commence par la lecture des images au format JPEG. Ensuite, les images sont converties de l'espace colorimétrique RGB vers l'espace colorimétrique L*a*b* pour adoucir les variations

[4] `https://www.mathworks.com/products/image.html`

de luminosité et facilement distinguer visuellement une couleur d'une autre. L'espace L*a*b* est constitué d'une couche de luminosité (L*) contenant la valeur de luminosité de chaque couleur[5], d'une couche de chromaticité (a*) indiquant la couleur de l'axe rouge-vert et d'une autre couche de chromaticité (b*) indiquant où se situe la couleur le long de l'axe bleu-jaune.

La deuxième étape vise à classifier les couleurs de l'espace a*b* en ayant recours à un clustering par les K-Means. En utilisant la distance euclidienne, nous regroupons les pixels en quatre clusters (le nombre de clusters est déterminé de manière empirique). K-Means renvoie pour chaque pixel d'entrée un index correspondant à un cluster. On peut alors étiqueter chaque pixel de l'image par son index de cluster.

Dans la troisième étape, pour chaque résultat de regroupement, la couche L* permet d'extraire la couleur la plus claire et la plus sombre de chaque cluster. De là, 8 images différentes (résultats du clustering) sont obtenues à partir de chaque image. Parmi ces résultats, nous sélectionnons manuellement certains d'entre eux qui sont les plus représentatifs des inclusions.

Ensuite, deux autres caractéristiques sont ajoutées manuellement : la taille des inclusions (petite, moyenne et grande) et l'abondance des inclusions (absente, rare, fréquente, commune et abondante). Il y a plusieurs limitations à ce travail. Par exemple, les images ont été obtenues à partir de diverses sources et dans différentes conditions d'éclairage, de fond et de réglages de caméra. De là, une comparaison précise des images, même avec un œil humain, est difficile. En outre, la sélection manuelle des résultats de clustering représentatifs est subjective, bien qu'elle aide à distinguer visuellement les différentes inclusions et les couleurs de la matrice.

Une exigence dans notre projet de thèse est d'éviter l'utilisation de méthodes de clustering trop complexes. À cet effet, une solution consiste à utiliser des méthodes itératives. Pour le cas du clustering dur, nous retenons les K-Means pour traiter les données continues et les K-Medoids pour traiter les données catégorielles ou booléennes. S'agissant du clustering flou, nous utilisons les C-Means floues (FCM) dans le cas de données continues et les K-Medoids flous (FKM) dans le cas de données catégorielles ou booléennes. Pour appliquer ces méthodes itératives, une question primordiale est la manière de choisir K points de données (où K est le nombre de clusters) comme centroïdes initiaux (ou graines) pour enclencher la méthode itérative retenue. Une méthode d'initialisation efficace doit être linéaire, de sorte que l'algorithme itératif qui l'utilise reste également linéaire.

Nous avons d'abord procédé à une revue de la littérature consacrée aux méthodes d'initialisation (Chapitre 7). La plupart des méthodes d'initialisation y sont présentées dans le cadre des K-Means et des K-Medoids, mais ces méthodes peuvent aussi être utilisées pour les versions floues de ces algorithmes.

La méthode la plus simple est celle proposée par MacQueen [11], qui propose d'utiliser les K premiers points de données comme centroïdes. Mais une telle procédure

[5] `https://fr.mathworks.com/help/images/examples/color-based-segmentation-using-k-means-clustering.html`

est sensible à l'ordre des données. MacQueen propose aussi de choisir les K graines de départ totalement au hasard parmi les points de données (méthode que nous appelons MacQueen2).

Faber propose d'effectuer de multiples relances de la méthode MacQueen2. Son inconvénient est que des valeurs aberrantes peuvent être choisies. D'un autre côté, plusieurs relances garantissent que la qualité de l'échantillon choisi s'améliore. Parmi les différentes méthodes proposées, la méthode MaxMin (aussi appelée Maximin) [12] est particulièrement intéressante. MaxMin calcule d'abord toutes les distances entre les points pris deux à deux. Ensuite, à chaque étape, on ajoute comme nouvelle graine le point qui est le plus éloigné de la graine dont il est le plus proche parmi les graines déjà choisie, ce qui a le grand intérêt d'améliorer l'homogénéité des clusters en construction. Cependant, le choix des deux premiers centres rend MaxMin quadratique.

Deux versions linéaires de MaxMin ont été proposées dans la littérature. Gonzalez suggère de choisir aléatoirement le premier centre et de choisir comme second centre l'objet le plus éloigné du premier centre [13]. Malheureusement, cette version dépend entièrement du choix aléatoire du premier centre. Son inconvénient est que des valeurs aberrantes peuvent être choisies. En revanche, Katsavounidis et al. proposent de considérer la moyenne globale des données comme premier centre [14]. Ainsi, seule la distance de chaque point à la moyenne globale doit être calculée pour déterminer le second centre, ce qui rend la méthode linéaire. Malheureusement, le recours à la moyenne globale n'est pas approprié aux données booléennes. Pour remédier à ce problème nous proposons MaxMin Linear, une variante de MaxMin qui applique son principe tout en restant de complexité linéaire et en étant adaptée aussi bien aux données booléennes qu'aux données continues. La moyenne générale de tous les points est d'abord calculée. Ensuite, nous choisissons comme premier centroïde le point le plus proche de la moyenne globale. Le deuxième centroïde est le point qui a la plus grande distance au premier centroïde. Ainsi, la complexité de la variante proposée reste linéaire par rapport au nombre de points de données. Ensuite, le choix des centroïdes suivants reste le même que dans MaxMin. Ainsi, MaxMin Linear peut servir dans un ensemble de clustering flou sur des données hétérogènes. Cela fait de MaxMin Linear une contribution simple mais très efficace. Nous comparons expérimentalement MaxMin Linear à plusieurs méthodes d'initialisation de la littérature. Notre méthode surpasse les méthodes existantes sur 22 ensembles de données synthétiques et réels. En outre, MaxMin Linear peut être utilisé avec des algorithmes autres que FCM, tels que Fuzzy K-Modes et FKM, qui s'appliquent aux données catégorielles et booléennes. Cette contribution a été publiée dans les actes de la 14e conférence internationale Machine Learning and Data Mining (MLDM 2018) [15].

Pour étudier l'impact du choix des paramètres sur la qualité d'un clustering, nous avons besoin d'un critère de qualité (Chapitre 8). Par exemple, lorsque l'ensemble de données est bien séparé et n'a que deux variables, un diagramme de dispersion peut aider à déterminer le nombre de clusters. Cependant, lorsque le jeu de données

comporte plus de deux variables, un bon index de qualité est nécessaire pour comparer différentes configurations de clusters et choisir le nombre approprié de clusters (K). En clustering, il n'y a pas de norme de référence liée aux données, permettant de statuer sur le nombre de clusters et la qualité du clustering obtenu, car en non supervisé les notions d'erreur et de taux d'erreur n'ont pas de sens, contrairement au cas de l'apprentissage supervisé. En outre, différents experts peuvent avoir des points de vue différents sur les mêmes données et exprimer des contraintes différentes sur le nombre, la taille et la forme des clusters. Ceci implique la nécessité de disposer d'indices de qualité.

Grâce à une approche visuelle (par exemple, le graphique qui considère les variations de l'indice de qualité en fonction du nombre de clusters), différentes solutions peuvent être présentées par rapport aux données. Ainsi, les experts peuvent-ils faire un compromis entre leur opinion et les meilleures solutions locales proposées par l'indice visuel. Selon Wang et al., Il existe deux types d'indices de qualité [16]. Les premiers sont fondés uniquement sur les valeurs d'appartenance aux centroïdes, alors que les seconds associent les valeurs d'appartenance aux centroïdes et les données.

Les indices fondés sur la décomposition de l'inertie (I) en inertie intra (W) et inertie inter (B), avec I = W + B, sont bien adaptés au clustering dur, car dans ce cas I garde sa valeur initiale tout au long du processus itératif. Ce n'est pas le cas en clustering flou, car l'inertie floue FI = FB + FW (où FW est l'inertie floue intra, alors que FB est l'inertie floue inter) dépend des coefficients d'appartenance aux clusters de chaque objet, ce qui fait que FI change de valeur au fil des itérations.

Lorsque le nombre de clusters augmente, la valeur des indices de qualité augmente mécaniquement aussi. Il faut donc arbitrer entre la complexité du modèle de clustering et sa qualité, en se demandant à chaque étape du processus itératif si l'ajout d'un nouveau cluster est utile. Pour répondre à cette question, les solutions les plus courantes sont la pénalisation et la règle du coude (Elbow rule). Parmi tous les indices de qualité, il n'en existe pas qui donne le meilleur résultat pour n'importe quel ensemble de données. Ainsi est-il intéressant de proposer un nouvel index de qualité spécialement conçu pour le clustering flou qui puisse aider l'utilisateur à choisir la valeur de K. Nous proposons donc un nouvel indice de qualité pour FCM appelé Visual TSFD, qui permet de déterminer visuellement le nombre de clusters. Nous comparons expérimentalement les résultats de Visual TSFD à ceux des indices de qualité issus de l'état de l'art et nous montrons que Visual TSFD les surclasse sur divers ensembles de données. De plus, Visual TSFD peut également être utilisé dans le cas de données catégorielles avec les Fuzzy K-Medoids [17]. Visual TSFD permet donc de traiter des ensembles de données hétérogènes, ce qui est particulièrement intéressant dans notre contexte applicatif. Cette contribution a été publiée dans les actes de la 14e conférence internationale Artificial Intelligence Applications and Innovations (AIAI 2018) [18].

Nous avons appliqué ces nouvelles méthodes aux données de la base Ceramo (Chapitre 9). Nous avons effectué deux types d'expériences, d'abord en opérant

séparément un clustering flou des objets céramiques à partir de différents types de données de Ceramo, puis en construisant un comité de classifieurs flous (ensemble clustering) issu des clusterings séparés. Nous comparons les résultats obtenus aux groupes définis par les experts en archéométrie du laboratoire ArAr. Ceux-ci reposent sur l'interprétation raisonnée de classifications ascendantes hiérarchiques portant sur les données chimiques relatives à ces objets. Dans nos expériences, nous considérerons ces groupes comme les groupes de référence (vérité terrain).

Dans le cas des clusterings séparés, nous appliquons successivement le clustering flou sur les données chimiques, les données de description et les données d'images, pour examiner la cohérence de nos résultats par rapport aux groupes définis par des experts. Les résultats obtenus avec les données chimiques et avec les données de description montrent tout à la fois la faisabilité de notre méthode et la bonne cohérence de ses résultats avec les opinions des experts.

Les résultats issus du clustering sur les données d'images ne sont pas corrélés avec les groupes définis par les experts, car les échantillons appartenant à différents groupes définis par des experts peuvent avoir des caractéristiques similaires, telles que la couleur et la taille des inclusions. Cela crée des difficultés pour les séparer à partir des méthodes de clustering. Dans les résultats des données chimiques, certains des groupes ont une très petite taille, alors que ce n'est pas le cas avec les résultats des données de description. Finalement, un point important est que nous comparons notre méthode, qui est automatique et dont la complexité est en O(n), à la méthode de regroupement des experts, qui est partiellement manuelle et dont la complexité est O(n2 log n).

Ensuite, nous cherchons à combiner les données chimiques et descriptives pour obtenir une meilleure correspondance avec les groupes définis par les experts (Chapitre 10). Un comité de regroupeurs (classifieurs non supervisés) simule en quelque sorte la collaboration entre chercheurs ou laboratoires utilisant des critères différents pour regrouper des objets. On citera comme exemple un laboratoire d'archéométrie opérant par clustering des données chimiques et un laboratoire d'archéologie opérant par regroupement de données descriptives. Nous discutons d'abord de plusieurs stratégies possibles pour effectuer un clustering ensembliste, puis nous détaillons la construction des solutions les plus pertinentes, c'est-à-dire un comité de regroupeurs et le comité de partitions combinées. Enfin, nous comparons les résultats des méthodes par comités aux groupes définis par les experts du domaine.

Lors de la conception de notre méthode de comité de regroupeurs, nous proposons différentes façons d'évaluer la dissimilarité globale entre deux objets en fonction des dissimilarités issues des clusterings flous opérés sur chaque type de données, à l'aide d'une moyenne généralisée (minimum, harmonique, géométrique, arithmétique, quadratique ou maximum), ce qui permet d'accorder plus ou moins de poids aux valeurs moyennées. Les meilleurs résultats sont obtenus avec la dissimilarité minimale qui attache plus d'importance à la ressemblance qu'à la dissemblance. Même si les résultats de prédiction du comité ne sont pas meilleurs que la classification portant

uniquement sur les données chimiques, cela donne un regroupement plus satisfaisant en termes de taille des clusters.

Nous proposons ensuite une nouvelle méthode pour combiner les clusterings issus de chaque type de données : la méthode des partitions combinées, qui consiste à durcir les partitions floues obtenues pour chaque type de données, pour en opérer ensuite la combinaison par tableau croisé (ou hyper-tableau croisé s'il y a plus de deux regroupeurs), dont on ne conserve que les croisements non vides. Cette méthode présente trois avantages : (1) sa complexité est linéaire ; (2) elle donne dans notre cas des résultats totalement cohérents avec les groupes définis par les experts, permettant de prédire sans erreur le groupe d'appartenance d'un objet ; et (3) les résultats peuvent être présentés de manière synthétique en utilisant un tableau croisé qui rend compte de l'homogénéité de chaque groupe en termes de centres de classification associés. Dans notre étude de cas, chaque cellule du tableau croisé correspond à un seul groupe défini par un expert, mais un groupe peut correspondre à plusieurs cellules, ce qui est un complément d'information intéressant. À partir de là, il est possible de déterminer sans erreur (au moins dans notre échantillon) le groupe défini par l'expert auquel appartient un objet céramique, en fonction de son cluster résultant des données chimiques et de son cluster résultant des données de description.

Bien qu'il existe déjà une longue tradition dans les domaines de l'archéologie et de l'archéométrie de développement des outils informatiques et statistiques, cette thèse a été stimulante de par son caractère interdisciplinaire. Dans ce travail, nous simulons en quelque sorte des processus impliqués dans la recherche interdisciplinaire, en croisant des points de vue sur les mêmes objets ou catégories d'objets caractérisés et définis selon différents critères. Nous avons également traité des données ayant un caractère hétérogène : numériques, textes, images. Des améliorations pourraient certainement être obtenues dans la façon dont les deux dernières catégories ont été traitées, en particulier pour les données d'images. Par ailleurs, le fait est que la méthodologie que nous avons développée dans cette thèse pourrait potentiellement être appliquée à une grande variété de données hétérogènes. Cette perspective est importante dans un contexte de disponibilité croissante de différents types de données, notamment via Internet.

Notre travail souligne tout d'abord l'importance de travailler sur des corpus relativement équilibrés, afin d'obtenir des clusters de taille plus grande et plus régulière. Nous aurions aussi besoin d'approfondir l'analyse de la performance de notre méthode de partitionnement combiné en distinguant apprentissage et généralisation. Pour ce faire, comme l'analyse est supervisée par les groupes définis par les experts, nous pourrions organiser une validation croisée. Par exemple, pour une validation croisée de type 2-fold, nous devons d'abord diviser de façon aléatoire l'ensemble de données en deux ensembles de taille égale. Le premier ensemble est utilisé comme un ensemble d'apprentissage et le second pour évaluer la qualité du modèle issu de l'apprentissage. Ensuite, le jeu de données est utilisé en s'entraînant sur le second ensemble et en évaluant sur le premier. Enfin, le taux d'erreur en généralisation est calculé en faisant

la moyenne des deux taux d'erreur obtenus.

Pour appliquer la validation croisée dans le cas où le modèle est la partition combinée, nous devons être en mesure d'insérer un nouvel objet dans cette partition. Dans le cas du clustering dur, la procédure de base consiste à insérer un nouvel objet dans le cluster dont le centre est le plus proche de l'objet considéré. Il serait intéressant de proposer une nouvelle procédure d'insertion, qui serait bien adaptée au cas du clustering flou. De plus, nous devons améliorer la méthode de choix des centres principaux associés à un objet donné, en tenant compte non seulement de l'ordre, mais aussi de la valeur de chaque coefficient flou. Enfin, lorsque nous naviguons avec OLAP, nous analysons les données avec des fonctions d'agrégation classiques, telles que somme, moyenne, maximum, etc. Il serait également intéressant de prendre en compte aussi les données textuelles, car il existe des défis pour agréger efficacement les données textuelles.

Bibliography

[1] Manuella Kadar. Data modeling and relational database design in archaeology. *Acta Universitatis Apulensis*, 3:73–80, 2002.

[2] Quanhong Sun, Qi Xu, and Qiaoqiao Li. Multidimensional analysis of distributed data warehouse of antiquity information. *The Open Cybernetics & Systemics Journal*, 9(1), 2015.

[3] Michel Fortin and Bernard Lachance. Conception of an integrated system for archaeological excavations. *JT Clark et EM Hagemeister (éds), Digital Discovery. Exploring New Frontier In Human Heritage, CAA*, pages 459–466, 2006.

[4] Aybüke Öztürk, Louis Eyango, Sylvie Yona Waksman, Stéphane Lallich, and Jérôme Darmont. Warehousing complex archaeological objects. In *International and Interdisciplinary Conference on Modeling and Using Context*, pages 226–239. Springer, 2015.

[5] Enrique H. Ruspini. Numerical methods for fuzzy clustering. *Information Sciences*, 2(3):319–350, 1970.

[6] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

[7] Jean-Paul Benzécri et al. *L'analyse des données*, volume 2. Dunod Paris, 1973.

[8] Maurice Picon. Le traitement des données d'analyse. PACT 10(379–499), 1984.

[9] Sylvie Yona Waksman. Etudes de provenance de céramiques, in Dillmann. P. and Bellot Gurlet. L, (dir.). *Circulation et Provenance des Matériaux dans les Sociétés Anciennes*, (195–216), 2014.

[10] P. J. Baldevbhai and R. S. Anand. Color image segmentation for medical images using l* a* b* color space. *IOSR Journal of Electronics and Communication Engineering*, 1(2):24–45, 2012.

[11] James B. Macqueen. Some methods for classification and analysis of multivariate observations. In *5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[12] Boris Mirkin. *Clustering for data mining: a data recovery approach*. Chapman and Hall/CRC, 2005.

[13] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[14] Ioannis Katsavounidis, C-C Jay Kuo, and Zhen Zhang. A new initialization technique for generalized Lloyd iteration. *IEEE Signal Processing Letters*, 1(10):144–146, 1994.

[15] Aybüke Öztürk, Stéphane Lallich, Jérôme Darmont, and Sylvie Yona Waksman. MaxMin linear initialization for fuzzy c-means. To appear, Springer: Machine Learning and Data Mining in Pattern Recognition, 2018.

[16] Weina Wang and Yunjie Zhang. On fuzzy cluster validity indices. *Fuzzy sets and systems*, 158(19):2095–2117, 2007.

[17] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2):3336–3341, 2009.

[18] Aybüke Öztürk, Stéphane Lallich, and Jérôme Darmont. A visual quality index for fuzzy c-means. volume 519, pages 546–555. IFIP Advances in Information and Communication Technology Springer, Cham, 2018.