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

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

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

Abstract

Ceramics are one of the most important archaeological materials to help in the reconstruction of past civilizations. Information about complex ceramic objects is composed of textual, numerical and multimedia data, which induce several research challenges addressed in this thesis. From a technical perspective, ceramic databases have different file formats, access protocols and query languages. From a data perspective, ceramic data are heterogeneous and experts have different ways of representing and storing data. There is no standardized content and terminology, especially in terms of description of ceramics. Moreover, data navigation and observation are difficult. Data integration is also difficult due to the presence of various dimensions from distant databases, which describe the same categories of objects in different ways.

Therefore, the research project presented in this thesis aims to provide archaeologists and archaeological scientists with tools for enriching their knowledge by combining different information on ceramics. We divide our work into two complementary parts: (1) Modeling of Complex Archaeological Data and (2) Clustering Analysis of Complex Archaeological Data. The first part of this thesis is dedicated to the design of a complex archaeological database model for the storage of ceramic data. This database is also used to source a data warehouse for doing on-line analytical processing (OLAP). The second part of the thesis is dedicated to an in-depth clustering (categorization) analysis of ceramic objects. To do this, we propose a fuzzy approach, where ceramic objects may belong to more than one cluster (category). Such a fuzzy approach is well suited for collaborating with experts, by opening new discussions based on clustering results.

We contribute to fuzzy clustering in three sub-tasks: (i) a novel fuzzy clustering initialization method that keeps the fuzzy approach linear; (ii) an innovative quality index that allows finding the optimal number of clusters; and (iii) the *Multiple Clustering Analysis* approach that builds smart links between visual, textual and numerical data, which assists in combining all types of ceramic information. Moreover, the methods we propose could also be adapted to other application domains such as economy or medicine.

Keywords: Complex Objects, Archaeology, Archaeometry, Ceramics, Databases, Data Warehouses, OLAP, Clustering, MaxMin Linear Initialization, Cluster Validity, Visual TSFD, Cluster Ensemble, Combined Partition Clustering.

Résumé

La céramique est l'un des matériaux archéologiques les plus importants pour aider à la reconstruction des civilisations passées. Les informations à propos des objets céramiques complexes incluent des données textuelles, numériques et multimédias qui posent plusieurs défis de recherche abordés dans cette thèse. D'un point de vue technique, les bases de données de céramiques présentent différents formats de fichiers, protocoles d'accès et langages d'interrogation. Du point de vue des données, il existe une grande hétérogénéité et les experts ont différentes façons de représenter et de stocker les données. Il n'existe pas de contenu et de terminologie standard, surtout en ce qui concerne la description des céramiques. De plus, la navigation et l'observation des données sont difficiles. L'intégration des données est également complexe en raison de la présence de différentes dimensions provenant de bases de données distantes, qui décrivent les mêmes catégories d'objets de manières différentes.

En conséquence, ce projet de thèse vise à apporter aux archéologues et aux archéomètres des outils qui leur permettent d'enrichir leurs connaissances en combinant différentes informations sur les céramiques. Nous divisons notre travail en deux parties complémentaires : (1) Modélisation de données archéologiques complexes, et (2) Partitionnement de données (*clustering*) archéologiques complexes. La première partie de cette thèse est consacrée à la conception d'un modèle de données archéologiques complexes pour le stockage des données céramiques. Cette base de données alimente également un entrepôt de données permettant des analyses en ligne (OLAP). La deuxième partie de la thèse est consacrée au *clustering* (catégorisation) des objets céramiques. Pour ce faire, nous proposons une approche floue, dans laquelle un objet céramique peut appartenir à plus d'un *cluster* (d'une catégorie). Ce type d'approche convient bien à la collaboration avec des experts, en ouvrant de nouvelles discussions basées sur les résultats du *clustering*.

Nous contribuons au clustering flou (*fuzzy clustering*) au sein de trois sous-tâches : (i) une nouvelle méthode d'initialisation des *clusters* flous qui maintient linéaire la complexité de l'approche ; (ii) un indice de qualité innovant qui permet de trouver le nombre optimal de *clusters* ; et (iii) l'approche *Multiple Clustering Analysis* qui établit des liens intelligents entre les données visuelles, textuelles et numériques, ce qui permet de combiner tous les types d'informations sur les céramiques. Par ailleurs, les méthodes que nous proposons pourraient également être adaptées à d'autres domaines d'application tels que l'économie ou la médecine.

Mots clés: Objets complexes, Archéologie, Archéométrie, Céramique, Bases de données, Entrepôts de données, OLAP, *Clustering*, Initialisation linéaire MaxMin, Validité des *clusters*, *Visual TSFD*, *Clustering* ensembliste, *Clustering* de partitions combinées.

Acknowledgment

First, I would like to thank God for his love and blessings without which this journey would not have been possible. Moreover, this Ph.D. research project became possible because of the support and kindness of many people. Thus, I would like to express my greatest gratitude towards all of them for their help.

I would like to express my sincere gratitude to my advisers Dr. Sylvie Yona Waksman, Professor Jérôme Darmont, and Professor Stéphane Lallich for their support during my thesis and for their patience, motivation, and guidance. I am thankful to them for giving me this opportunity of being a part of such an interesting collaborative project.

I would like to thank all the members of the jury. I thank Professor Nicole Vincent and Dr. François Pinet for accepting to be my thesis reviewers and for their detailed and thoughtful comments. I also thank Dr. Zoï Tsirtsoni and Professor Henning Christiansen for accepting being my thesis examiners.

I would like to thank the Rhône Alpes Region's ARC 5: "Cultures, Sciences, Sociétés et Médiations" for their financial support without which I couldn't have completed this project. I also sincerely thank all the members of the SID team of the "Entrepôts, Représentation et Ingénierie des Connaissances" (ERIC) laboratory, and the Archaeological Ceramics team of the Archaeology and Archaeometry (Archéologie et Archéométrie) laboratory.

Furthermore, I am grateful to my parents especially my dear mother Rabia, my father Kamil, my loving sister Aysena, my parents-in-law: Jagdish and Sweety and my brother-in-law Karan. Your love, support and encouragement helped me to move forward and overcome the hurdles that came up in life. Most importantly, I am thankful to them for keeping their faith and their continuous encouragement throughout my years of study and during this thesis.

I would like to express my gratitude to my loving husband, Kunal. His continues support, encouragement and even criticism helped to pave my path towards success. We faced various difficulties together and in this togetherness, I found the strength to move forward. Thus, I dedicate this thesis to all of you, my loving family. I would like to take this moment to thank all my friends and colleagues in Lyon and others spread across the globe for sharing their love and encouragement during this thesis time.

Thank you again for being part of this journey!

Table of contents

List of Tables

List of Figures

Chapter 1

Introduction

Contents

Archaeology is the study of the human past through material remains. One common archaeological material is pottery, which provides information on many aspects of human activity, including chronology, trade and technology. Potteries and their production change through time. These changes make pottery a tool for chronology, and at the same time, they give clues regarding exchange and trade. Moreover, once a pottery was broken, it could not be recycled, unlike iron or glass, for instance. Therefore, potteries have remained to exist until today. Thence, it is one of the most important archaeological material to help reconstruct past civilizations.

1.1 Context

In recent times, there has been on one hand a high growth rate and availability of various archaeological data and networks. On the other hand, digital systems and tools made possible an increased usage of data by a wide potential number of users ranging from students to researchers, and from museum curators to tourists.

Furthermore, the developments in scientific and statistical techniques have also contributed in gaining deeper insight in the analysis of archaeological materials, e.g., ceramic objects, geographical coordinates and digital photography. However, there are currently not many comprehensive digital systems, tools and databases that can be easily used by archaeologists to study a variety of archaeological information and share their findings easily and consistently.

Moreover, archaeological ceramics[1] can be described in different ways, by archaeologists, museum curators or archaeological scientists, e.g., through criteria related to archaeological contexts, the history of art or the properties of the ceramic material, as may be investigated by chemical, mineralogical and petrographic analyses. In addition, ceramics can be used to determine contextual relationships, which help to highlight archaeologically meaningful data from the mass of individual data. In other words, exploiting ceramic data allows discovering patterns that are only visible in large and distributed ceramic samples. In archaeology, core data are highly contextual. Thence, ceramics and their properties can help obtain comprehensive knowledge about technological, cultural and geographical issues, through information on context period and provenance of ceramics. Furthermore, the information stored in databases typically focus on a limited range of ceramic descriptors and is not interoperable. Thus, research involving archaeological ceramics cannot easily take advantage of combining all these types of information by building smart links between visual, textual and numerical data.

1.2 Nature of Archaeological Data

During the process of documenting and dating an excavation site, archaeologists try to integrate all the data in a coherent way to interpret the archaeological record for a better understanding of human cultures. In this process, the construction of reusable resources for the study of ceramics is important. From this point on, some fundamental questions are asked to better understand the properties that give evidence of the ceramics past, such as where and when they were made, how they were made and what their function was. Therefore, raw and constructed data about ceramics can be categorized into three levels [1].

First level: Data are accessible directly from the ceramic object and its context, e.g., decoration of ceramic object and the location where ceramic samples were found. Such data are mainly stored without any changes later on in databases.

Second level: Data necessitate a first level of interpretation, especially in the form of hypotheses, e.g., the expected origin of an object found at a given site and the scientific analyses carried out to test these hypotheses. For instance, the form of a ceramic object is first level-data and can be used to suppose an origin (before any analysis), that is a location data.

Third level: Data are the result, e.g., the attribution of the object to an origin according to scientific analyses and possibly other criteria. For instance, attributing (after analysis) a location can be obtained as a result of petrographic and chemical analysis.

[1] In this thesis, we use both "pottery" and "ceramic" to designate all the range of categories of these archaeological objects.

1.3 Research Challenges and Motivations

According to the needs of current research, there are some challenges in dealing with highly contextual data. Finding useful information in huge amounts of highly contextual data is difficult for researchers and students. Data are globally very heterogeneous. Databases have different file formats, access protocols and use various query languages. There is no common classification systems, nor standardized terminology, which are required for understanding relationships from interconnections.

Moreover, databases generally have a specific description. For instance, in Lyon, the archaeometric studies carried out on ceramics [2, 3] led to the development of the Ceramo database [4], which began in the late 1970's. In Ceramo, until recently, ceramics were described by their chemical composition together with an archaeological textual summary. Interoperability is also limited, with databases only providing a web interface, but no API (Application Programming Interface). Thus, combining various information about archaeological objects, such as textual, numerical and graphical documents, which would allow powerful computer analyses, is at best an intricate task as of today. The research challenge is to integrate various dimensions from distant databases that describe the same categories of objects in a complementary way.

Figure ***1.1*** presents a schematic representation of our thesis project. Available data lie on the left-hand side of the schema, while our analysis objectives lie on the right-hand side. In between, we mention the methods we aim to use for analyzing complex data, the research challenges and the tasks this process induces.

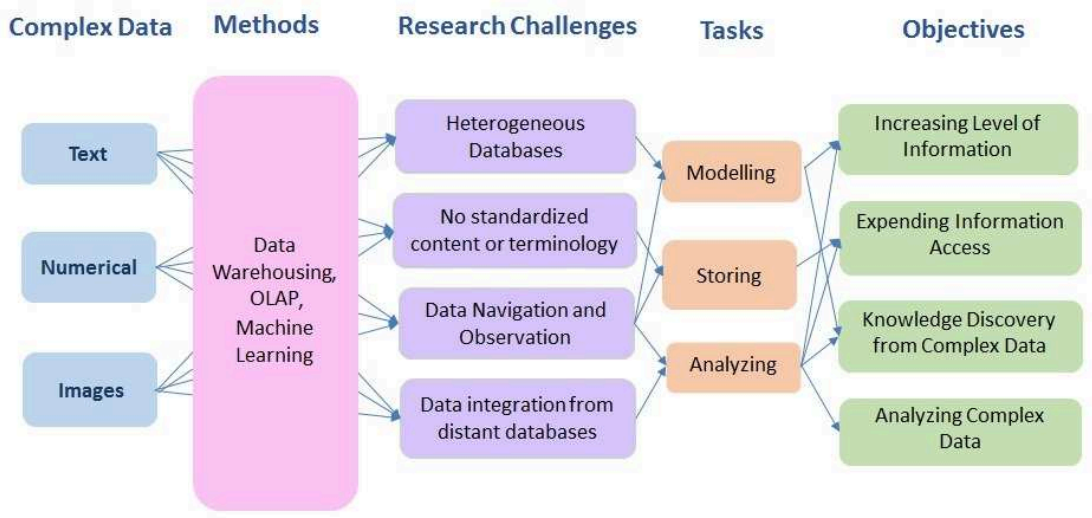

Figure 1.1: Overview of the thesis project

The global objective of this thesis is expanding information access by using all the

information provided by complex data, in order to take benefit from each and every piece of information. Applied on the Ceramo database, we aim at designing data warehousing and data mining methods that help better analyzing and categorizing complex objects.

More precisely, our motivation is to address the following research issues:

- model and store complex data;
- find analysis methods having linear time-complexity;
- deal with problems related to clustering, such as finding an efficient clustering initialization technique to build homogeneous clusters, evaluating clustering quality to obtain accurate analysis results for helping experts;
- find analysis methods to deal with mixed types of data.

1.4 Contributions

This thesis is divided into two complementary parts. Our contributions in each part are summarized below.

1. Modeling of Complex Archaeological Data
2. Clustering Analysis of Complex Archaeological Data

1.4.1 Part 1: Modeling of Complex Archaeological Data

Database model for Archaeology and Archaeometry: We introduce a new database model called Ceramo 3.0. It models previously little-exploited textual descriptions of ceramic samples, which include rich locational data, graphical descriptions (technical drawings, photos) and results of different kind of laboratory analyses (such as petrographical and chemical ones). The aim of this new database model is to store information about each and every piece of ceramic sample in a more systematic manner.

Multidimensional model for OLAP analysis: Online analytical processing (OLAP) helps interactively navigate a data warehouse to discover outliers or hidden patterns. We use Ceramo 3.0 to source a data warehouse, which is original in its storage of data that are not only numerical. Thanks to OLAP, we integrate various points of view on ceramic objects, e.g., text descriptors, chemical analysis results, to be able to learn deep contextual information and to make an observation from these different viewpoints. This new multidimensional model can be a base model for other archaeological and archaeometric data warehouses.

1.4.2 Part 2: Clustering Analysis of Complex Archaeological Data

Data mining techniques and tools are essential to analyze the information available in big amounts of data. The main challenge is to find a method that is well suited for performing clustering on archaeological data. For this purpose, we aim at improving the clustering results by using textual, numerical and graphical data possibly coming from different databases. To achieve this goal, we propose a fuzzy approach that is well suited to collaborate with experts by allowing to open new discussions during clustering analysis. More precisely, our contributions to fuzzy clustering follow.

Initialization method for fuzzy clustering: We propose a new, fast and easy to implement linear initialization method for fuzzy clustering called MaxMin Linear, which outperforms existing methods on a variety of numerical real-world and artificial datasets. One noteworthy characteristic of MaxMin Linear is that it can be applied on both numerical and textual data.

Quality index for fuzzy clustering: The performance of a clustering algorithm critically depends on the number of clusters. The optimal number of clusters can be estimated using quality indices. However, the existing quality indices that are well-suited to fuzzy clustering are limited when different kinds of datasets come into play. Thus, we propose the *Visual Transformed Standardized Fuzzy Difference (TSFD)*, an innovative, visual quality index for the well-known Fuzzy C-Means (FCM) method. Visual $TSFD$ outperforms state-of-the-art quality indices on several numerical real-world and artificial datasets. Furthermore, Visual $TSFD$ can be also applied on both numerical and textual data.

Multiple clustering analysis approach: This approach consists of two alternative solutions to deal with heterogeneous data. Our first solution introduces a new ensemble clustering scheme that uses both the Fuzzy C-Means and Fuzzy K-Medoids methods. This includes starting clustering initialization with MaxMin Linear and applying the Visual $TSFD$ quality index to determine the optimal number of clusters. Then, the main fuzzy centers obtained during ensemble clustering allow understanding the structure of data. Our second solution introduces a combined partition clustering method that hardens clustering results before they are cross-tabulated. This method helps differentiate the groups obtained by clustering.

1.5 Thesis Outline

The remainder of this thesis is organized as follows.

Chapter 2 discusses the state-of-the-art for the first part of the thesis. It notably presents a selection of ceramic database projects, as well as of archaeological data warehouses.

Chapter 3 presents the design and implementation of the Ceramo 3.0 database

model.

Chapter 4 details the multidimensional remodeling of the Ceramo 3.0 database. It also includes a sample OLAP scenario that exploits the obtained archaeological and archaeometric data warehouse.

Chapter 5 notably surveys various clustering methods. This chapter presents several data mining methods applied in archaeology and archaeometry. It also presents characteristics of iterative fuzzy clustering methods, clustering ensemble methods and how to deal with mixed types of data.

Chapter 6 presents the approaches we apply for detecting image features from ceramic data, using techniques such as segmentation and color detection.

Chapter 7 surveys various clustering initialization methods. It details the MaxMin Linear initialization method, along with its implementation and experimental validation achieved on both real-life and artificial datasets.

Chapter 8 presents quality indices. It also presents the Visual $TSFD$ validation method, along with its implementation and experimental validation achieved on both real-life and artificial datasets.

Chapter 9 presents the *expert-defined groups*, which are groups of ceramic samples having similar characteristics. These groups are defined by the experts from the Laboratory "Archaeology and Archaeometry" (ArAr) in Lyon, which serve as ground truth for our work. This chapter also presents the disjoint cluster analysis steps for each kind of data, i.e., chemical data, description data and image data. Lastly, the disjoint cluster analysis results are interpreted with the support of the expert-defined groups.

Chapter 10 presents our multiple cluster analysis for both numerical and categorical data, using two different techniques. First, fuzzy clustering ensemble deals with mixed types of data. Second, combined partition clustering allows visualizing multiple clustering results in a synthetic way.

Finally, **Chapter 11** concludes this thesis and provides insights for future research.

Part I

MODELING COMPLEX ARCHAEOLOGICAL DATA

Chapter 2

Ceramic Databases and Archaeological Data Warehouses

Contents

In recent years, several databases were created to highlight different perspectives in pottery research. These databases have different types of contents, depending on the aspects of ceramics studies they focus on. Moreover, specific formats may be implied based on different contents, e.g., numbers for chemical analyses or text and/or images for petrographic analyses.

Ceramic databases usually have a main type of content and may focus on specific categories of ceramics, time periods or regions. These databases can be either publicly available on-line or not. Additionally, some interface features may also be available, such as interactive maps, interactive 2D or 3D views or statistical tools, the latter being of particular interest in the context of our research.

In addition, most archaeological applications feature operational databases, i.e., they allow updating and querying the data. However, a new trend in archaeology is to build data warehouses [5], which are analytical databases. Data warehouses bear a specific, multidimensional model that allows On-Line Analytical Processing (OLAP). OLAP helps navigate and observe data in order to find hidden patterns.

Thus, in this chapter, we first review a selection of archaeological ceramic databases that we consider representative of the diversity of contents, formats, statuses and features. Further, we present existing archaeological data warehouses.

2.1 Ceramic Database Projects

The Levantine Ceramics Project (LCP), directed by Boston University, is an archaeological database focusing on ceramic wares produced in the Levant, from the Neolithic to the Ottoman periods. It mainly includes archaeological (typological, chronological

and geographical) data, but also provides fabric and petrographic data. The format of LCP data is in text and image. The LCP is an open, interactive internet resource[1].

Roman Amphorae: a digital resource[2], proposed by the University of Southampton, provides an online introductory resource for the study of Roman amphorae, based on a rich corpus of archaeological information together with petrographic and fabric data.

POTSHERD[3] is a collection of pottery from the Roman period (1st cent. BC – 5th cent. AD) in Britain and Western Europe, including distribution maps and links to complementary resources.

The Worcestershire Online Ceramic Database[4] is designed to make available pottery fabric and form type series for Worcestershire, from the Neolithic to the early post-medieval period.

The *Information sur la Céramique Médiévale et Moderne* (ICERAMM)[5] proposes a database that focuses on medieval and modern ceramics in western and northern France, Belgium and Switzerland.

The *Prototype d'Encyclopédie Céramologique en Ligne* (PECL)[6] is a project of encyclopedia for ceramics of the Mediterranean and sub-Saharian region of all periods, including detailed archaeological contexts information.

ASCSA Digital Collections (ASCSA)[7] presents archaeological objects and contexts from the excavations of the American School of Classical Studies at Athens, in the Athenian Agora and in Corinth.

Other types of databases particularly focus on petrographic and fabric data. The National Roman Fabric Reference Collection (NRFRC)[8] is the online version of a reference book providing detailed and standardized fabric descriptions of Roman wares found in Britain [6].

Fabrics of the Central Mediterranean (FACEM), directed by the University of Vienna, focuses on fabric data of Greek, Punic and Roman pottery in the Southern Central Mediterranean area. FACEM includes interactive maps and allows downloading detailed information[9].

Petrodatabase is a petrographic relational database featuring interactive maps [7]. There are also several image databases that are designed for a larger audience, using digital representations of ceramics. One of them, *Sgraffito in 3D*[10], proposes 3D reconstructions of late medieval pottery collection from the Museum Boijmans Van

[1] `https://www.levantineceramics.org/`
[2] `http://archaeologydataservice.ac.uk/archives/view/amphora_ahrb_2005/`
[3] `http://potsherd.net/atlas/potsherd`
[4] `https://www.worcestershireceramics.org/`
[5] `http://iceramm.univ-tours.fr/bdceramm.php`
[6] `http://pecl.fr/`
[7] `http://ascsa.net/research?v=default`
[8] `http://romanpotterystudy.org/nrfrc/base/index.php`
[9] `http://facem.at/`
[10] `http://www.sgraffito-in-3d.com/`

Beuningen in Rotterdam.

Yet other databases focus on chemical data. The CeraDAT project[11], developed by the Demokritos National Centre for Scientific Research in Athens, is a prototype relational database including interactive maps and focusing on the Aegean and the wider Eastern Mediterranean Region [8].

The MURR Archaeometry Laboratory Database[12] built at the University of Missouri features chemical analyses of ceramic artifacts from many regions, including Northern, Central and Southern America and the Mediterranean. It also gives access to "historical" chemical databases, such as the Berkeley laboratory's. Archaeological information is actually presented as a bibliography [9].

The Archaeometry Research group at the University of Fribourg has also established a bibliography of several reference groups of ancient ceramics from Switzerland, Italy, France and Germany, with their chemical composition.

Covering a large range of periods and regions, the Ceramo database from the Laboratory of ArAr in Lyon initially used to be mainly a chemical database [3, 10], including only limited archaeological information. As presented in Chapter 3, the new Ceramo 3.0 database has been re-centered on ceramic objects and enriched with archaeological and multimedia contents.

Table ***2.1*** summarizes the features of the ceramic databases mentioned above. In Table ***2.1***, the primary contents of databases are indicated by X, secondary contents by x and occasional contents by (x).

2.2 Archaeological Data Warehouses

In the literature, research works related to data warehousing and OLAP on archaeological data may be divided into two main groups: (1) OLAP on the features of archaeological materials and (2) OLAP on top of Geographical Information Systems (GISs), i.e., Spatial OLAP (SOLAP).

In the first group of approaches, to analyze the huge amount of antiquity-related data from the ancient Chinese civilization, the North China University works on building a distributed data warehouse, which helps manage, share and analyze antiquity information [11]. This data warehouse composes of (1) an architecture of local data warehouses that process data and (2) a global data warehouse that integrates all data and supports OLAP.

The *Soprintendenza Speciale Archeologia Belle Arti e Paesaggio di Roma* (SSBAR), an Italian public institute, does research to better understand the stone resources present in the Roman area, especially tuff from the quarries of Lazio [12]. SSBAR's data warehouse and OLAP analyses helped obtain detailed information and plan a new view of natural archaeological parks in Italy.

[11] `http://www.ims.demokritos.gr/ceradat/`

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

***Table 2.1*:** Ceramic database features

Ceramic databases	Database type				Data type			Features		
	Archaeological	Chemical	Petrographic	Fabric	Textual	Numerical	Multimedia	Online	Structured	Stat tools
LCP	X		x	x	x		x	x	x	
Roman Amphorae	X		X	X	x		x	x	x	
POTSHERD	X			x	x		x	x		
Worcestershire Ceramics	X		(x)	X	x		x	x	x	
NRFRC	(x)		X	X	x		x	(x)		
FACEM	x			X	x		x	x	x	
Petrodatabase	(x)		X	x	x		x	(x)	x	
ICERAMM	X			x	x		x	x	x	
PECL	X		(x)	(x)	x		x	x	x	
ASCSA	X				x		x	x	x	
Sgraffito in 3D	x				x		x	x		
CeraDAT	(x)	X			x	x		x	x	x
MURR	(x)	X			x	x		x		downloadable
Fribourg	(x)	X			x	x		x		
Ceramo	x	X	(x)		x	x			x	x

To create accurate models for historical and archaeological sites, various tools are used, such as three-dimensional (3-D) scanner technology. However, the scanning process creates a large number of virtual collections, which take a very long time to index and are very costly. Thus, the Canadian National Research Council (NRC) and the University of Ottawa worked on a project involving data warehouses to preserve old and current 3D visual information of historical sites. Using a data warehouse helped enforce scalability [13].

Data cubes are structures that precompute data aggregations to speed up OLAP. However, computing and storing of data cubes can consume lots of computing power and disk space in the first place. Thus, researchers from Aix-Marseilles University works on a partitioned cube that aims to provide storage reduction and perform experiments using archaeological excavation data [14].

In the second group of approaches, researchers from the Department of History and the *Centre de Recherche en Géomatique* at Université Laval (Québec) worked towards solving the problem involved in data recording and analysis of archaeological excavation results by using a GIS-based system. In general, GISs help record, analyze and visualize spatial data. Here, the GIS helped build an Integrated System for Archaeological Excavation (ISAE) that supports multicriteria analyses [15].

Finally, *Moving megaliths in the Neolithic* (MEGAGEO) [16] is a multi-disciplinary project conducted by the universities of Évora, Aveiro and Lisbon in Portugal. This project aims at developing a large spatial data warehouse that stores geochemical and

petrographic information. Using SOLAP allows to find the source of slabs used in the construction of dolmens from archaeological sites.

All the above-stated data warehouse project are designed specifically based on concrete project requirements such as to deal with the features of the archaeological materials or to work with the geometrical information. However, none of them can be considered as a comprehensive data warehouse project that can be used for doing OLAP analysis on the complex ceramic data. This creates a room to introduce a new model for experts to do OLAP analysis on the ceramic data.

2.3 Summary

Cultural Heritage Information Systems (CHISs) are increasingly being used by several potential users from different backgrounds and domains. Such systems are complex to use, as they focus on specific contents, have different formats and are reachable via different access protocols. This makes it hard to have a comprehensive tool that would collect information from these systems. Moreover, in addition to the aforementioned complexity, it is difficult to observe hidden patterns from huge amounts of data.

Thus, in this chapter, we survey the tools that allow analyzing data from CHISs. We first present a selection of ceramic databases and compare their features. Then, we introduce some archaeological and archaeometric data warehouses, which allow OLAP and SOLAP analyses based on different features of the data, such as the characteristics of materials, geolocation and 3D features.

Chapter 3

Design of the Ceramo 3.0 Database Model

Contents

The Ceramo database was initially designed to meet the needs of archaeological scientists working in Lyon laboratory. It was mainly used for recording chemical data, to which multivariate statistical treatments where applied.

3.1 Data Modeling

The design of the new Ceramo 3.0 database model involves all types of analytical data. All these data are combined with extended definitions of the analyzed ceramic samples, along with rich location data. In addition to these data, various graphical documents such as drawings and images of ceramic samples are added in the new model. This information is necessary to complement ceramic information. This new database thus stores complex data. Data may indeed be qualified as complex if they are [17]:

- multiformat, i.e., represented in various formats (databases, texts, images, sounds, videos...);
- and/or multistructure, i.e., diversely structured (relational databases, XML documents repository...);
- and/or multisource, i.e., originating from several different sources (distributed databases, the Web...);

- and/or multimodal, i.e., described through several channels or points of view (radiographies and audio diagnosis of a physician, data expressed in different scales or languages...);
- and/or multiversion, i.e., changing in terms of definition or value (temporal databases, periodical surveys...).

Ceramo 3.0 is a complex database, now centered on pottery samples that are defined by various descriptions, geographical features and results of different types of analyses (Figure ***3.1***). Each package (Geography, Status and Description, Analysis) from Figure ***3.1*** is further detailed in the following subsections. All conceptual models are depicted as UML class diagrams [1].

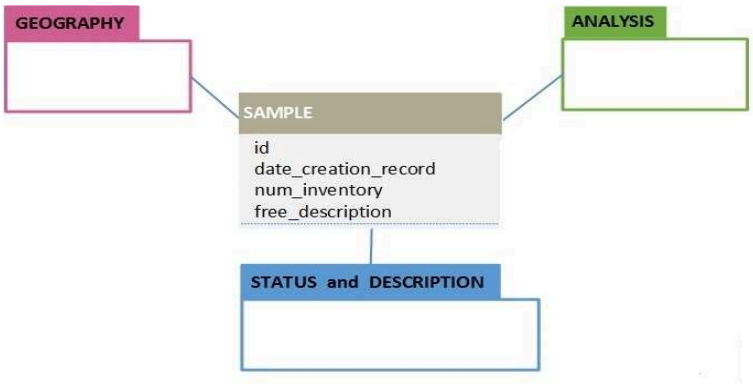

***Figure 3.1*:** Ceramo 3.0 conceptual schema: global simplified view

3.1.1 Geography Package

Figure ***3.2*** displays the Geography package of the data model. The LOCATION class connects geolocation data to PROVENANCE, SUPPOSED ORIGIN, ATTRIBUTION and STORAGE OUTSIDE LABORATORY classes.

The PROVENANCE class bears information regarding the location data where the object was found. The SUPPOSED ORIGIN class provides a supposed origin before analysis. The ATTRIBUTION class indicates where the object was demonstrated to come from after analysis. The SITE, TOWN/MUNICIPALITY, REGION and COUNTRY classes represent a hierarchy of information within LOCATIONS. These three classes regard objects as historical objects, while the STORAGE OUTSIDE LABORATORY class represents where objects are physically stored. Even though they do not encompass the same level of information, all these classes are connected by the structure of geographical information.

[1] `http://www.uml.org`

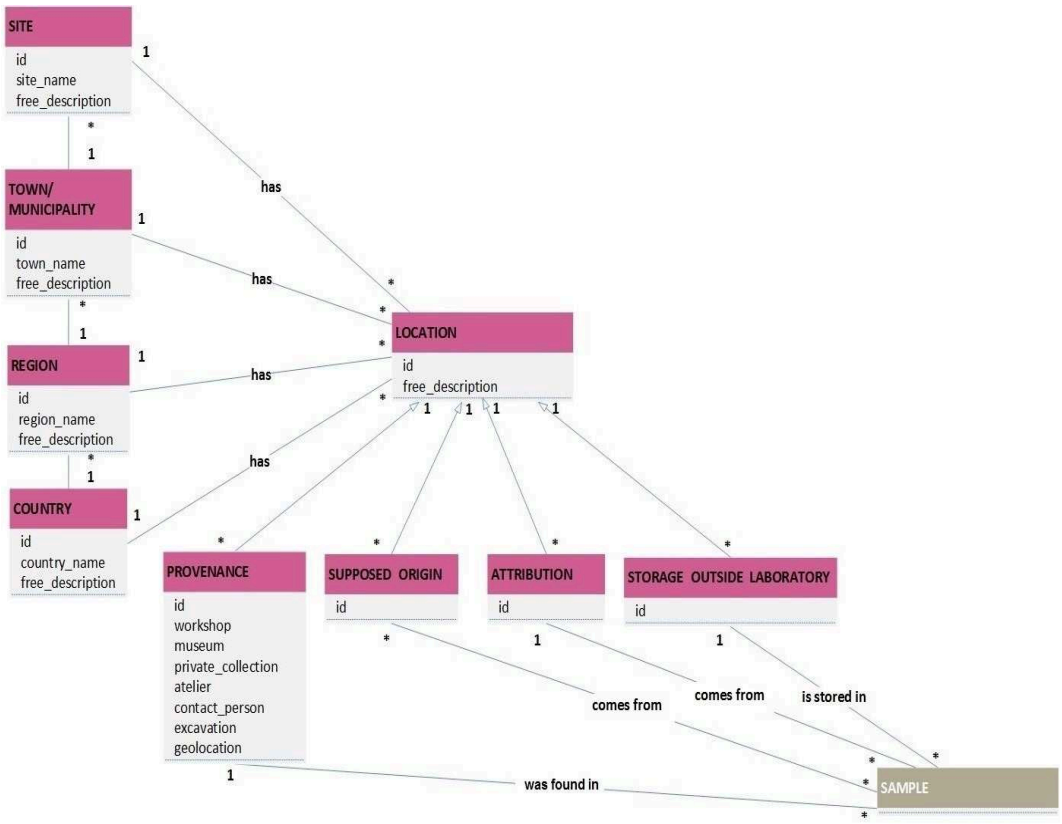

***Figure 3.2*:** Geography package

3.1.2 Status and Description Package

The Status and Description package of the data model is depicted in Figure ***3.3***. The DESCRIPTION class bears textual descriptors of an object. It includes most of the information that helps identify the object archaeologically. The CATEGORY class helps categorize ceramics, e.g., "COMM." (common ware) or "CARREAU" (tile). The DATING class stores all the periods of samples both at a general level, e.g., "Médiéval" (Medieval) and at a precise level, e.g., "13e siècle" (13^{th} century). It also includes the methods used for dating. The BIBLIOGRAPHY and LINKS classes contain the bibliographical data related to each sample, with the links referring to public resources. The LOCAL REFERENCE class specifies whether an object is a local reference or not. The FIRING MODE class contains data about the firing mode, which is coded as mode A, mode B, mode C, etc. [18, 19]. The STORAGE IN LABORATORY class bears location data if the object is stored in the laboratory. The LEGAL STATUS OF OBJECT class contains data about ownership of the object.

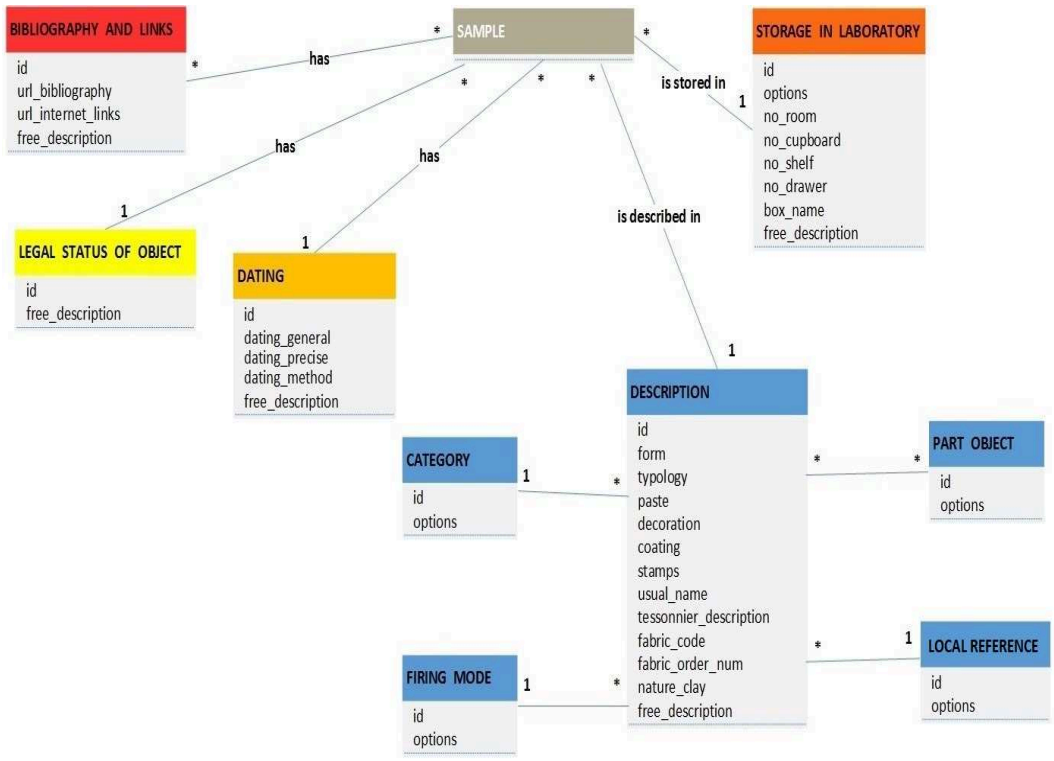

***Figure 3.3*:** Status and Description package

3.1.3 Analysis Package

Figure ***3.4*** depicts the Analysis package of the data model. Analysis results of a sample are represented in class ANALYSIS RESULT. Each analysis corresponds to a series of separate records, each of which contains an individual measure. This is due to the fact that samples may be analyzed using several methods, such as chemistry or petrography. A given sample may also be analyzed several times using the same technique, but with different parameters. For example, as a growing number of chemical elements have been analyzed since the 1970's, when a given sample is analyzed again, some chemical elements (e.g., aluminum or calcium) are assigned several concentration values.

The DIFFRACTION, DILATO, PETRO, CHEMISTRY, BINO, SEM and OTHER ANALYSIS classes bear data regarding diffraction, dilatometry, petrography, chemistry, binocular microscopy, scanning electron microscopy and additional, miscellaneous analyses, respectively. These data are of different nature: numerical (e.g., chemical data), text (e.g., descriptions of fabrics), images (e.g., photos under the binocular or under the petrographic microscope, diffractograms). The results of analyses are used by the experts of the ArAr laboratory to cluster samples into groups

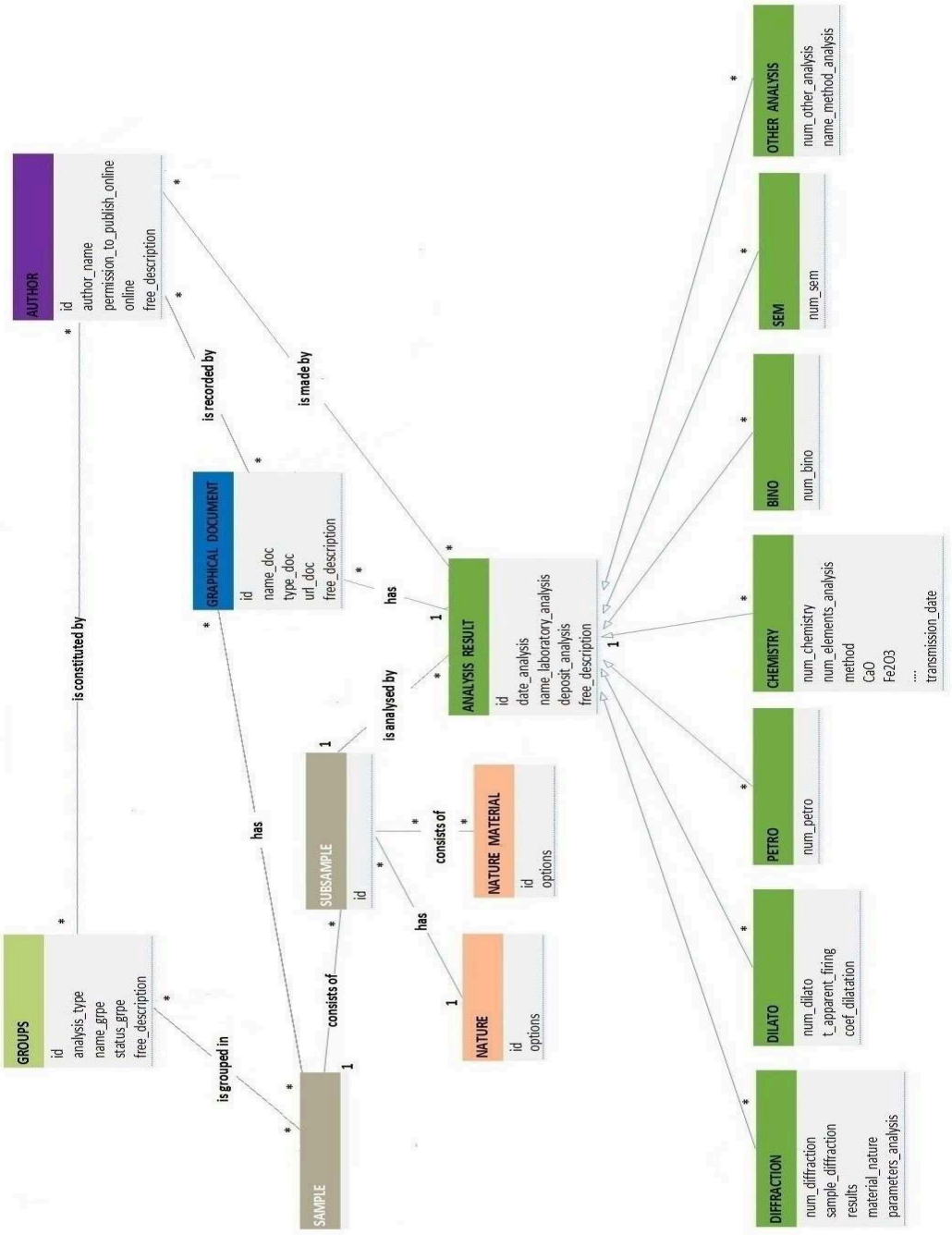

***Figure 3.4*:** Analysis package

sharing the same characteristics, according to analyses results. These correspond to the class GROUPS, they will be refered to as the "expert-defined groups" in the following parts of this research.

As shown in the Table ***3.1***, Ceramo 3.0 database is a comprehensive database model compared to other archaeological ceramic databases.

***Table 3.1*:** Ceramic database features compared to Ceramo 3.0

Ceramic databases	Database type				Data type			Features		
	Archaeological	Chemical	Petrographic	Fabric	Textual	Numerical	Multimedia	Online	Structured	Stat tools
LCP	X		x	x	x		x	x	x	
Roman Amphorae	X		X	X	x		x	x	x	
POTSHERD	X			x	x		x	x		
Worcestershire Ceramics	X		(x)	X	x		x	x	x	
NRFRC	(x)		X	X	x		x	(x)		
FACEM	x			X	x		x	x	x	
Petrodatabase	(x)		X	x	x		x	(x)	x	
ICERAMM	X			x	x		x	x	x	
PECL	X		(x)	(x)	x		x	x	x	
ASCSA	X				x		x	x	x	
Sgraffito in 3D	x				x		x	x		
CeraDAT	(x)	X			x	x		x	x	x
MURR	(x)	X			x	x		x		downloadable
Fribourg	(x)	X			x	x		x		
Ceramo	x	X	(x)		x	x			x	x
Ceramo 3.0	**X**	**X**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**

3.2 Summary

The ongoing developments in scientific research and analysis techniques have led to an increase in the quantity of archaeological and archaeometric data. At this point, designing comprehensive databases or tools that can store all complex data is needed to combine various point of views and parameters together.

In this chapter, we focus on the needs of ceramic specialists to store and analyze complex archaeological and archaeometric data. We introduce the new Ceramo 3.0 database and detail its design, which satisfies the requirements of ceramic specialists. Ceramo 3.0 is divided into three main packages, whose classes and attributes include several graphical documents, rich locational data, extensive definitions of ceramic samples and different analysis results.

Chapter 4

Warehousing Complex Archaeological Data

Contents

The Ceramo 3.0 database is designed to provide solutions to existing issues related to the analysis of complex archaeological and archaeometric data. It is also motivated toward improving the usability of these complex data. In this chapter, we show how ceramic data can source a data warehouse to perform OLAP. Such analyses help navigate and observe data from different perspectives, thus providing researchers with better insight into their data. The main advantage of this approach is to identify hidden patterns of ceramic samples. Moreover, this chapter introduces a sample OLAP scenario.

4.1 Exploring Data

Data warehouses are actually databases with a specific model tailored for efficient OLAP analyses. In a data warehouse, the observed data are called facts, e.g., sales in a business context. They are characterized by measures that are usually numerical, e.g., quantities sold and amounts of money. Facts are observed with respect to different analysis axes called dimensions, e.g., sold products, store location and sale date. Thus, data warehouse schemas are called multidimensional schemas or more casually star schemas, for facts are usually represented in the center of the model, with dimensions gravitating around. Star schemas help answer queries such as "total sales revenue of each product in Lyon in 2014", to go on with our business example.

Moreover, dimensions may be organized in hierarchies, e.g., a time dimension could be subdivided into day, month, quarter and year. Such a structure helps observe facts

at different granularity levels, e.g., "dezooming" from one quarter of a year to said year to have a more global (aggregate) view of sales or "zooming" from one month to one day in this month to have a more detailed view of sale events. These operations actually correspond to OLAP's rollup and drill-down operators, respectively.

Thus, to allow OLAP navigation in Ceramo's data, we must select facts to observe, axes of analysis (dimensions) and import data into the data warehouse. The result is called a cube (hypercube when the number of dimensions is greater than 3), where dimension values are coordinates that define a fact cell.

4.2 Multidimensional Model

In this sample scenario, we choose to observe results of chemical analyses with respect to ceramic sample provenance, dating, description and groups. Our data warehouse's star schema is provided in Figure ***4.1***, again as a UML class diagram. Facts are modeled as a quaternary association-class connected to dimension classes. To make use of numerical values for analyses, the SAMPLE class from Figure ***3.1*** is combined with the Analysis package (Figure ***3.4***) into the SAMPLE ANALYSES class in Figure ***4.1***, which models our analysis facts. In our case, aggregates (summaries) are the number of samples, the average number of sample and the number of analyses, etc. Dimension classes are PROVENANCE, GROUPS, DESCRIPTION and DATING, which are the same as in the Ceramo database (Figures ***3.2***, ***3.3*** and ***3.4***). Moreover, the LOCATION class individually connects to all classes in the SITE, TOWN/MUNICIPALITY, REGION and COUNTRY hierarchy to still allow a connection in case of missing value at one hierarchy level (Section 4.4).

4.3 Example of OLAPing Archaeological Ceramic Data

Once part of Ceramo's data is multidimensionally remodeled, OLAP analyses can be performed. OLAP actually helps interactively navigate the data warehouse, e.g., to discover outliers or hidden patterns. We use Pentaho Business Analytics [1], a suite of open source business intelligence, as our OLAP engine. Pentaho features a user console that is a web-based design environment. The console helps visualize and navigate hypercubes, which are created from the data warehouse with the help of the schema-workbench tool.

As an example of OLAP analysis, let us examine the content of a specific chemical group coming from the GROUP class and compare it to the initial typological classification from the DESCRIPTION class. Samples within a given chemical group belong to the same pottery production, i.e., they share the same origin. They usually come from several excavations (PROVENANCE class) and their circulation and corresponding fluxes provide insight into past contacts between populations and trade

[1] `http://www.pentaho.com`

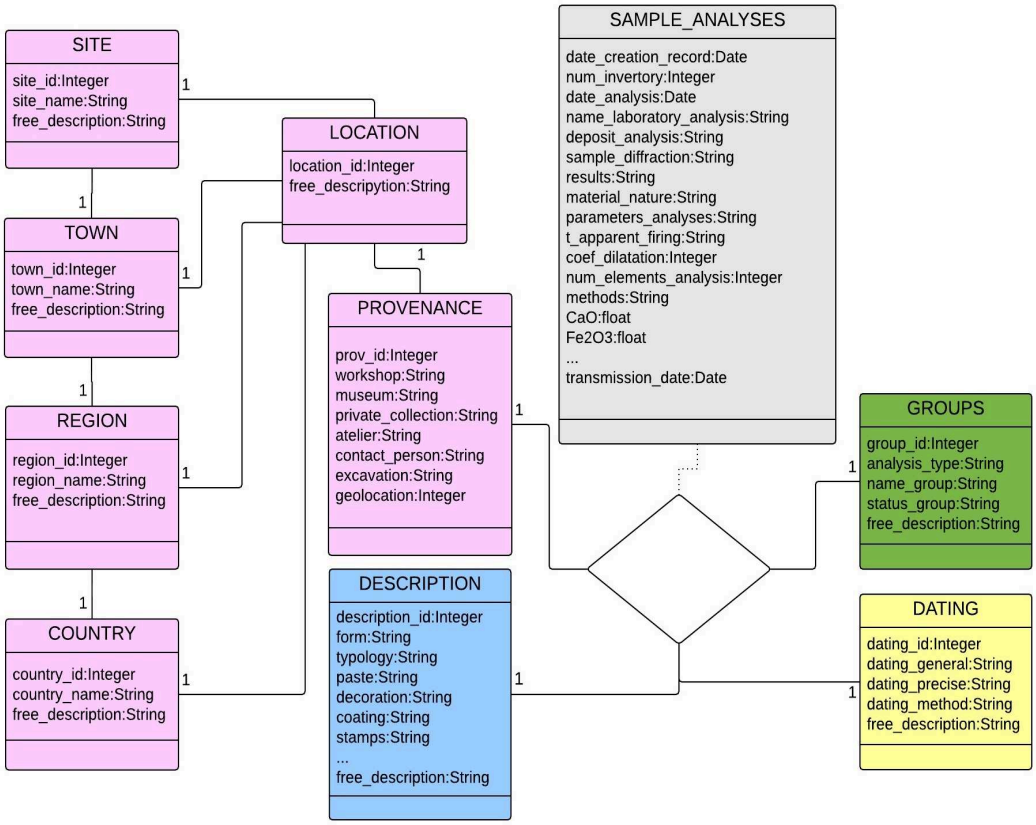

***Figure 4.1*:** Chemical data warehouse multidimensional schema

networks. When different workshops manufacture similar wares, classified under the same typology, chemical analysis "sorts out" the different productions and enables archaeologists and historians to better understand economic trends and cultural influences.

In a first analysis, successive rollups help aggregate PROVENANCE data at the country level to achieve a coarser view of data. We take interest in ceramics of the Byzantine period (the Medieval period in the DATING class) called "Zeuxippus Ware"; we select the DATING, DESCRIPTION (typology) and PROVENANCE (country) dimensions; and eventually count "Zeuxippus Ware" occurrences (OLAP slice and dice operators, respectively). "Zeuxippus Ware" corresponds to a typological class that has 163 occurrences in the database. "Zeuxippus Ware" was found all over the Mediterranean and beyond, but was also largely imitated.

In a second analysis, we slice on PROVENANCE, GROUPS and DESCRIPTION and dice on "Zeuxippus Ware stricto sensu". A research program enabled to define several distinct chemical groups, including one corresponding to the "Zeuxippus Ware

stricto sensu" (87 samples), which is the "prototype" of this ware. We identify the features of each production, including information on its geographic distribution, related to trade networks [20].

Figure ***4.2*** compares data with the number of samples whose description includes the term "Zeuxippus", i.e., including both "Zeuxippus Ware stricto sensu" and wares imitating it or related to it typologically. In all countries, examples of both prototype and imitations were found. It is nonetheless noticeable that a larger proportion of imitations comes from Greece, a new insight that may be significant. In the histogram, "chemical classification"(in blue) is based on the actual diffusion of ceramic products; it is related to economic factors. It confirms the large distribution of this ware in countries of the Mediterranean and Black sea areas. Although the bias introduced by the initial sampling needs to be taken into account, the number of samples from each country gives an idea of the abundance of this ware, e.g., only very few examples were found in France. "Typological classification" (in red) refers to the diffusion of models and fashions and is thus more related to cultural factors. This example somehow simulates the comparison of data obtained on the same categories of objects from two databases, focusing each on another aspect of these wares. It shows the discrepancies, but also the added value that may be obtained when connecting information. It also shows how OLAP can contribute to the understanding of economic and cultural relationships at the Byzantine period, thanks to its ability to look at information from a different viewpoint.

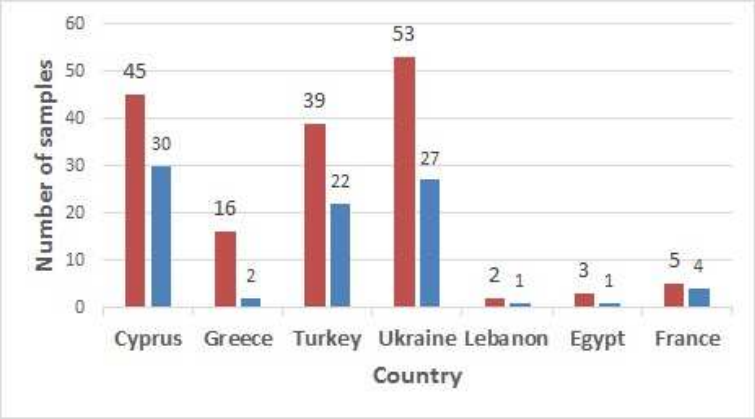

***Figure 4.2*:** Distribution of Samples by Country, for descriptions including the term "Zeuxippus Ware" (in red) and in chemical group "Zeuxippus Ware stricto sensu" (in blue)

4.4 Issues in Archaeological Ceramic Data Analysis

We have been confronted with a couple of major challenges before OLAP analyses and when performing OLAP onto Ceramo's data. First, we encountered a classical problem in databases, i.e., missing values. In our example scenario, there is location information for provenance studies, but in practice, some information, i.e., site, town, region or country might be missing in the database. For example, some samples relate to Sudak (Ukraine) and Acre (Israel), with no archaeological site reference. This is why we complement the geographical hierarchy with direct associations from provenance to site, town, region and country. Some of this missing information (town to region to country relationships) shall be found in external sources, though.

Moreover, data about archaeological ceramics mostly consists of textual and numerical data. On one hand, textual data include information about the characteristics of ceramics, such as form, chronological information, etc. On the other hand, numerical data include information produced by various analyses performed on the ceramic material. Both can be warehoused. However, while classical OLAP provides a good tool for analyzing numerical data (through aggregation functions such as sum, average, minimum, maximum, etc.), it is not very convenient for textual data, whose individual values can only be counted. Thus, in order to gain in-depth knowledge about ceramics, there is a crucial need to better take textual data into account in OLAP.

4.5 Summary

In this chapter, we introduce how ceramic data can source a data warehouse. This data warehouse can be used to perform OLAP analyses. First, we implement a multidimensional model. Then, OLAP is performed to navigate and observe ceramic data from different perspectives. The issues during this analysis are presented to open new discussions.

Part II

CLUSTERING COMPLEX ARCHAEOLOGICAL DATA

Chapter 5

Data Clustering Methods

Contents

Clustering is a field of research that belongs to both the domains of data mining and machine learning. Clustering is a useful technique for grouping a set of unlabeled data points (instances) described by attributes (variables) such that points in the same cluster (group) have similar characteristics, while points in different clusters have dissimilar characteristics. During the last 60 years, several clustering methods have been created to solve various problems involving continuous and/or categorical data. To perform a cluster analysis, Halkidi et al. [21] propose to consider five necessary steps to follow.

The first step is feature selection. Each feature should be non-redundant and relevant regarding user interest. The second step is that each variable should be preprocessed. The third step is to select a clustering method. Two fundamental parts play a role in this step: (1) a proximity measure that evaluates pairwise dissimilarities or similarities between the objects being considered; and (2) a clustering criterion that consists of rules and a function to be optimized. The forth step is to validate

clustering result. Evaluation is needed for achieving the final partition of data, e.g., by using quality indices. The last step is an interpretation of clustering results that is performed by help of application area experts.

5.1 Clustering Methods

One of the classification criteria for cluster analysis is to handle cluster overlapping. In crisp (or hard) clustering, a data point belongs to one and only one cluster, while in fuzzy clustering [22], a data point belongs to several clusters. Fuzzy clustering is very useful in many applications, e.g., the text categorization of various news into different clusters: an economy, an energy and a politics cluster; where an article containing the keyword "petrol" could belong to all three clusters. Furthermore, it is also possible to open discussions with domain experts when using fuzzy clustering.

5.1.1 Crisp Clustering Methods

According to Halkidi et al. [21], there are four different categories in which clustering algorithms can be summarized: hierarchical clustering, density-based clustering, grid-based clustering and partitional clustering.

Hierarchical clustering [23] proceeds by either merging small clusters into larger clusters (ascendant hierarchical method) or by splitting larger clusters to small clusters (descendant hierarchical method). In the first case, the ascendant hierarchical method produces a sequence of clustering schemes of decreasing number of clusters at each step according to a chosen criterion. At each step of the algorithm, the pair of clusters with the shortest distance are combined into a single cluster. The algorithm stops when all samples are merged into a single cluster. The result of the algorithm is a tree diagram called dendrogram, which shows how clusters are related. On the contrary, the descendant hierarchical method produces a sequence of clustering schemes increasing the number of clusters at each step. That is, at each step of the algorithm, one cluster is chosen, and this cluster is partitioned into a pair of clusters having the maximum distance between each other. The algorithm stops when each cluster contains only one sample.

Density-based clustering [24] groups neighboring objects of a dataset into clusters based on density conditions. Density-based clustering is suitable to handle arbitrarily-shaped collections of points and cluster of different sizes. It is also an effective method to handle the separation of outliers (extreme values). An example of density-based clustering algorithm is DBSCAN, where it is required to specify the neighborhood radius of a point and the minimum number of points in the neighborhood. Thence, DBSCAN is sensitive to the parameters that are required to start clustering. Such parameter needs to be known in advance or needs to be tested with a range of parameters to find appropriate settings, which is a time consuming process.

Grid-based clustering [25] is mainly aimed at spatial data mining. Instead of data points, the neighborhood surrounding data points, which are represented by cells, are clustered in a grid data structure. The number of cells can be significantly smaller than the number of data points. In such cases, grid-based clustering can obtain better clustering performance than other clustering methods. STING [26] is a much-cited grid-based algorithm. It divides the spatial area into rectangular cells to store the statistical parameters of object numerical features. The grid structure facilitates parallel processing and incremental updating. However, STING results depend on the granularity of the lowest level of the grid because a parent cell is constructed without consideration of the spatial relationship between the children and their neighboring cells. In this approach, the cluster boundaries are horizontal or vertical for a cell and no diagonal boundary is detected. This result decreases quality and accuracy of the clusters.

Eventually, Partitional clustering [27] works with pre-defined (supposed known) number of clusters and it determines the partitions by optimizing some criterion function, which is an iterative procedure. The most significant examples of partitional clustering are K-Means [28], K-Medoids [29] and K-Mode [30]. K-Mode is based on the concepts of K-Means.

5.1.2 Fuzzy Clustering Methods

The aim of the K-Means algorithm is to minimize the distance within each clusters. It starts by choosing K data points as initial centroids (seeds) of the clusters. Then, each data point of the dataset is assigned to the cluster of the closest centroid. Cluster centroids are updated as the average of all points in each cluster until a termination criterion is reached. In K-Means, this criterion can be a fixed number of iterations t, e.g., $t = 100$. Alternatively, another termination criterion is that until the modification of centers are negligible.

Fuzzy inertia is a core measure in fuzzy clustering. Fuzzy inertia FI (Equation 5.1) is composed of fuzzy within-inertia FW (Equation 5.2) and fuzzy between-inertia FB (Equation 5.3). Membership coefficients u_{ik} of data point i to cluster k are usually stored in a membership matrix U that is used to calculate FW, FB, and FI. Note that $FI = FW + FB$. Moreover, contrary to case of crisp case, FI is not constant because it depends on u_{ik}. When FW changes, the values of FI and FB also change.

$$FI = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^{m} d^2(x_i, \overline{x}) \tag{5.1}$$

$$FW = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^{m} d^2(x_i, c_k) \tag{5.2}$$

$$FB = \sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{m} d^2(c_k, \overline{x}) \tag{5.3}$$

where n is the number of instances, K is the number of clusters, m is the fuzziness coefficient (by default, $m = 2$), c_k is the center of the k^{th} cluster $\forall\ 1 \leq k \leq K$, $\overline{x}$ is the grand mean (the arithmetic mean of all data – Equation 5.4), and function $d^2()$ computes the squared Euclidean distance.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.4}$$

In our work, we use two fuzzy iterative methods, Fuzzy C-Means (FCM) [31] and Fuzzy K-Medoids (FKM) [32]. These methods keep the clustering process linear, are easy to implement and are easily understood by the experts.

5.1.2.1 Fuzzy C-Means

FCM is a common method for fuzzy clustering that adapts the principle of the K-Means algorithm. FCM, proposed by Dunn [33] and extended by Bezdek et al. [31], applies on numerical data. Since numerical data are the most common case, we choose to experiment our proposals with FCM.

The aim of the FCM algorithm is to minimize FW. It starts by choosing K data points as initial centroids of the clusters. Then, membership matrix values u_{ik} (Equation 5.5) are assigned to each data point in the dataset. Centroids of clusters c_k are updated based on Equation 5.6 until a termination criterion is reached. In FCM, this criterion can be a fixed number of iterations t, e.g., $t = 100$. Alternatively, a threshold ϵ can be used, e.g., $\epsilon = 0.0001$. Then, the algorithm stops when $FW_{K+1} \div |FW_{K+1} - FW_K| < \epsilon$.

$$u_{ik} = \frac{1}{\sum_{j=1}^{K}(\frac{\|x_i - c_k\|^2}{\|x_i - c_j\|^2})^{\frac{1}{m-1}}} \tag{5.5}$$

$$c_k = \frac{\sum_{i=1}^{n} u_{ik}^{m} x_i}{\sum_{i=1}^{n} u_{ik}^{m}} \tag{5.6}$$

5.1.2.2 Fuzzy K-Medoids

The Fuzzy K-Medoids (FKM) algorithm was proposed by Krishnapuram et al. [32]. In contrast to FCM, FKM chooses data points as centers (prototypes). The principle of FKM is also based on minimizing FW (Equation 5.2). The method starts by choosing K random data points as initial medoids (centroids of FKM). Then, membership matrix values u_{ik} (Equation 5.5) are assigned to each data point in the dataset (as in

FCM). Medoids of clusters c_k are updated based on Equation 5.7 until a termination criterion (the same as in FCM) is reached.

$$c_k = x_q$$

$$q = \underbrace{argmin}_{1 \leq i \leq n} \sum_{i=1}^{n} u_{ik}^m \|x_i - c_k\|^2 \tag{5.7}$$

5.2 Data Mining Methods Applied in Archaeology and Archaeometry

Many archaeologists use methods from computer science and statistics, to study data that are generated during various research phases, before, during and after the excavations (see for instance the activities of the international association "Computer Applications and Quantitative Methods in Archaeology"). These data may range from individual objects to archaeological regions or analysis results. Different methods can be applied to these data, depending on the nature of the problem, e.g., the description, comparison or classification of datasets [34].

Hug et al. [35] and Orton [36] have reviewed the main statistical methods in the literature that are used in archeology. Multivariate statistical methods, which are a subdivision of statistics for studying more than one variable, have notably been used.

Discriminant Analysis (DA) [37] is a supervised learning technique. DA is used when two or more groups are known *a priori* and one or more new observations are to be attributed to one of the known groups based on measured characteristics.

By contrast, Principal Component Analysis (PCA) [38] is an unsupervised learning technique. It is mainly used for the data reduction of continuous variables, when data are collected on a large number of variables from a single group. The variables can be standardised by centering them and dividing by the standard deviation (Equations 5.8 and 5.9). This helps to avoid any particular distribution of the data since there are features on different ranges. During the transformation, the first principal component (first new variable) has the largest possible variance. The second principal component has the second largest and so on. The result is presented as a scatter plot.

$$x_{New} = \frac{x - \overline{x}}{s} \tag{5.8}$$

where x is the value of the i^{th} object's j^{th} variable ($1 < i < n, 1 < j < K$), n being the number of objects and K the number of clusters. The normalized x is x_{New}, $\overline{x}$ being the mean of the values in variable j (Equation 5.4) and s the standard deviation of the values in variable j (Equation 5.9).

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{5.9}$$

Another technique is Correspondence Analysis (CA) [39], which is used for understanding the link between two categorical variables (rather than continuous variables). In this analysis, Euclidean distance is replaced with a Chi-square distance ($\tilde{\chi}^2$) [40].

To deal with a large set of categorical variables, Benzécri et al. [39] propose Multiple Correspondence Analysis (MCA), an extension of CA algorithm. The CA algorithm is applied to the complete disjunctive table constructed from the categorical variables. For instance, the laboratory *Trajectoires* at the Université de Paris 1 Panthéon-Sorbonne, France, works on an interesting case study referring to Neolithic sites. In their work, the χ^2 test and correspondence analysis are applied to analyze the data matrix of artifacts made out of different material, e.g., flint and sandstone. The results highlight a strong relationship between the types of flint artifacts and their features [41].

In the ArAr laboratory (Lyon), archaeological scientists usually define groups of ceramic objects based on their chemical composition. They perform hierarchical clustering (ascendant method) and discriminant analysis onto chemical data as tools in provenance studies (see Section 9.1) [10].

5.3 Characteristics of Iterative Fuzzy Clustering Methods

For both iterative fuzzy methods FCM and FKM, primordial work is to understand characteristics of methods such as selection of membership coefficient (m) as fuzziness value and a number of clusters (K) as initial centroids (for FKM, it indicates initial medoids). These selections may effect clustering algorithm. For example, the algorithm stays fuzzy if $m > 1$. If a K value is chosen less than required number, clusters are merged while they could be more separated.

5.3.1 Fuzzy Clustering Initialization

Initialization is a very important step in a clustering analysis. Initialization methods for the continuous and Boolean data should be selected carefully to keep the selected algorithm linear (see Chapter 7).

5.3.2 Fuzzy Clustering Quality Indices

Parameter selection affects the quality of clustering. A quality criterion is needed to evaluate the choice of parameters. Since we desire to use FCM and FKM as our basis algorithms without any changes, we focus on dealing with the determination of an appropriate number of clusters (see Chapter 8) to obtain accurate clustering results.

5.4 Dealing with Mixed Types of Data

In clustering, variable (or feature) selection plays a key role. The goal is to select variables that together help to obtain the maximum possible information. There are two types of data (Figure ***5.1***):

- qualitative data that do not allow to measure the information about modality, e.g., ceramics' color such as red and gray;
- quantitative data that allow to measure the information about modality with numbers, e.g., the number of ceramics in an archaeological context.

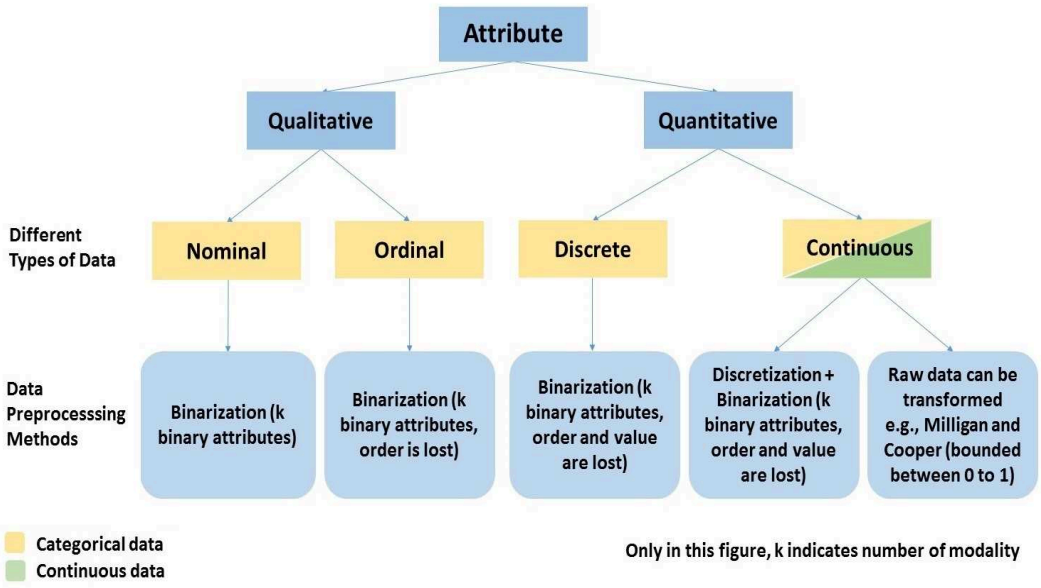

***Figure 5.1*:** Data types and corresponding preprocessing steps before cluster analysis

Within the class of qualitative variables, variables are categorized as nominal or ordinal. Nominal variables are unordered qualitative variables such as ceramic types, e.g., Aegean ware and Miletus ware. Ordinal variables are ordered qualitative variables such as the hardness of pottery, e.g., soft, hard and very hard. However, the range difference between categories is inconsistent. For instance, the difference between soft and hard is probably much harder than the difference between hard and very hard.

Within the class of quantitative variables, variables are usually categorized as discrete or continuous (interval). On the one hand, if we can count a set of items, then it is a discrete variable, e.g., in the ArAr laboratory, the total number of chemical elements analyzed in ceramics is 24. That is, any number of chemical elements (between

1 and 24) can be used in analyses. Thus, the number of chemical elements can be considered a discrete variable. On the other hand, if a variable's values lie an interval of $\mathbb{R}$, then it is a continuous variable, e.g., the concentration of chromium (Cr) in a ceramic piece.

A specific case of categorical data is binary data (Booleans), which have only two modalities, i.e., presence or absence. Therefore, they can be considered as numerical data with two values: 0 and 1. In this case, the mean is the proportion of 1s.

Statistical pieces of software usually distinguish categorical data from continuous data. From a practical point of view, one can consider that all data that are represented by a frequency distribution can be treated as categorical data, e.g., nominal or ordered qualitative data, discrete quantitative data or regrouped continuous quantitative data. However, non-regrouped continuous quantitative data must be treated as continuous data [42].

In data mining, the two following strategies can be used to cluster mixed types of data.

- Converting all variables into the poorest data type, e.g., Boolean. For each type of data, a preprocessing step can be applied, such as binarization [43] and discretization [44, 45] (Figure ***5.1***). After the conversion, a single algorithm is applied to the whole dataset.

- Retaining the original data and applying adequate algorithms with respect to data types. For instance, He et al. propose that the original mixed dataset is divided into two sub-datasets [46]: a pure categorical dataset and a pure numeric dataset. Then, adequate mining algorithms are applied to the datasets separately.

5.4.1 Binarization

Binarization is a procedure for converting categorical attributes into a binary form. For instance, k different binary attributes can be created if a categorical attribute has k distinct values. Each binary attribute fits in one logical value of the categorical attribute. Thus, in binarization, only one k attribute takes the value 1 and remaining attributes take the value 0 [47].

5.4.2 Discretization

Discretization is an essential data preprocessing task that can help handle missing values, asymmetry and outliers; obtain more compact and harmonious datasets; and simplify data so as to reduce computation complexity and memory space. More details about discretization may be found in [48].

Discretization transforms continuous data into categorical data. This process starts by partitioning the range of continuous data into k segments. For instance, the dating of ceramics can be transformed into discrete values, each representing a slice

of 100 years, e.g., 0 to 100 as 1, 100 to 200 as 2, 200 to 300 as 3, etc. Then, any value (e.g., 362) can be represented as both a Boolean vector (e.g., 00010) and a symbolic value (e.g., 4) representing an interval. However, intra-slice variations are lost.

5.4.3 Variable Transformation

Milligan et al. propose a method that starts with transforming continuous values to standard score form [49]. Then, they determine the minimum and maximum values of each variables of the dataset and apply Equation 5.10.

$$x^*_{i,j} = \frac{x_{i,j} - Min(x_j)}{(Max(x_j) - Min(x_j))} \tag{5.10}$$

where $x_{i,j}$ is the value of the i^{th} object's j^{th} variable, $x^*_{i,j}$ is the transformed value of $x_{i,j}$, $Min(x_j)$ being the minimum of the j^{th} variable and $Max(x_j)$ the maximum of the j^{th} variable.

To deal with variable asymmetry, one solution is to do logarithmic or square root transformation of the considered variables. Thus, in practice, it is essential to select proper methods depending on the datasets and the learning context.

5.5 Clustering Ensemble Methods

In clustering analyses, we need to manage different types of data for achieving stable clusters, since mixed types of datasets are common in real-life data mining and social science applications, including archaeology and archaeometry. However, existing, well-established clustering algorithms address different types of attributes separately. There is currently no appropriate clustering method to deal with mixed type of attributes. Thence, clustering ensemble is used to combine by aggregating several clustering results achieved by different methods.

Clustering ensemble methods aim to find better clustering solutions by fusing information from several data clusterings. All clustering ensemble method follow two steps: partition generation and partition integration (using a consensus function process). In the partition generation process, one can apply either the same clustering algorithm with different parameters or different clustering algorithms, e.g., K-Means and hierarchical methods. For instance, data resampling is a partition generation method [50]. In the consensus function process, multiple partitions generated by different clustering algorithms are combined into a single clustering solution. This process is quite challenging. For instance, let C_1 and C_2 be the results of a first clustering, and C^*_1 and C^*_2 the results of a second clustering. It is a complicated task to combine them accurately, since clusterings do not have any direct associations.

Yet, there are a number of advantages in using clustering ensemble instead of single clustering [51]:

- Robustness and stability: clustering is less sensitive to noise and outliers.
- Novelty: a committee solution ensures a better exploration the space of solutions.
- Parallelization and scalability: data subsets can be clustered in parallel with a subsequent combination of results which allows to reduce the size of the data that we apply with the algorithm. By the same way, ensemble methods can integrate solutions from multiple data sources.

Both supervised and unsupervised ensemble methods exist. In supervised learning, training data include both exogenous variables and labels. Supervised methods are usually fast and accurate. They give correct results with good accuracy when new unlabeled data are given as input. A good supervised ensemble uses a diversity of classifiers having each one a good accuracy.

In unsupervised learning, clustering is one of the technique that is used to cluster input data in groups by their statistical properties only. There is no label. Therefore, there is no ground truth nor golden standard. As in supervised ensembles, classifiers (clusteres) must be diverse. This can be achieved by setting a variety of initial conditions, doing data resampling, using different algorithms, distance and grouping criteria.

Finally, in many situations, fuzzy clustering is more natural than crisp clustering. Since our main aim is to give suggestions and open new discussions with archaeologists, it motivates us to use fuzzy clustering, which can capture the uncertainty of real-life archaeological ceramic data analysis.

5.6 Summary

Clustering is a field of research that belongs to both data mining and machine learning. It groups together a set of data points described by attributes. To perform a good clustering analysis, there are several criteria to be taken into consideration, which include the selection of the clustering method, appropriate initialization and effective quality analysis.

In this chapter, we review several clustering methods, as well as data mining methods that have been applied in archaeological and archaeometric studies. Then, we discuss the importance of clustering initialization and quality indices that are used in fuzzy clustering. Moreover, we survey several methods that deal with mixed types of attributes, i.e., clustering ensemble.

Chapter 6

Dealing with Fabric Images

Contents

Graphical documents stored in the Ceramo 3.0 database include, e.g., pictures of objects, technical drawings and pictures taken under a binocular microscope, which are called fabric images. Such graphical documents, especially ceramic fragment images and technical drawings, have been used to determine ceramic features [52–54]. Thus graphical documents constitute an interesting material for improving clustering. Due to time constraints in this thesis project and for the sake of simplicity, we only consider the features of fabric images as a first step.

In this chapter, we present ceramic fabric images, and then methods used in literature for image feature detection. Later, we consider what features are important in fabric images, how to design and implement a feature detection process, and what the limitations and issues faced during image feature analysis are.

6.1 Fabric Images

Fabrics correspond to the characteristics of ceramic materials as may be observed with the naked eye or using a binocular microscope (magnification is usually 10 to 20 times) [55].

A fabric image consists of two main basic components: the matrix and inclusions. The matrix is composed of clay minerals, which cannot be distinguished at this scale. It contains different kinds of inclusions, e.g., rock fragments and minerals (Figure ***6.1***). The nature of the matrix and inclusions is mainly governed by the choice of raw materials used to manufacture ceramics, and by the manufacturing process. Both the matrix and inclusions are modified to some extent by firing conditions [1].

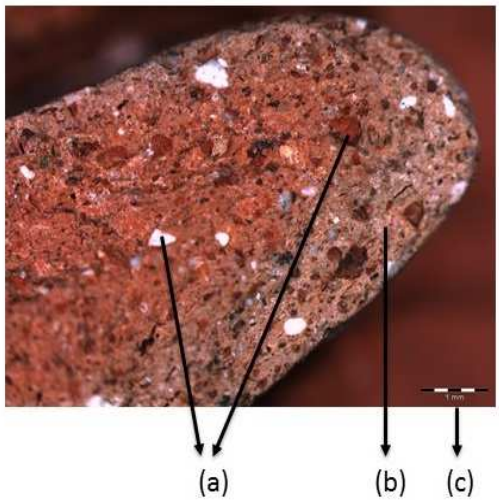

***Figure 6.1*:** Sample fabric image with inclusions in different colors (a), matrix (b) and image scale (c)

6.2 Methods for Image Features Detection

One of the main approaches in image processing is segmentation. It helps divide an image into several regions and distinguish (separate) objects. Surveys from the literature, such as Pal and Pal's [56], Hedberg's [57], Khan's [58] and Zaitoun and Aqel's [59] identify the main techniques, i.e., threshold, edge, region and clustering-based segmentation.

Threshold-based segmentation is a simple and ancient technique, typically used for images having light objects on a dark background. There are two types of threshold-based segmentation methods: local and global. In the global method, a single threshold value is set to the background to distinguish from the remaining region. In contrast, the local method uses multiple segmentation thresholds to divide an image into several regions. The advantage of threshold-based segmentation is that its computation cost is much lower than other segmentation techniques. On the other hand, its disadvantage is that it is quite sensitive to noise.

Edge-based segmentation aims at finding a discontinuity in an image that is called an edge or a boundary. Thanks to features of the image, e.g., color and texture changes, and different gray values, it is possible to find discontinuities in images. There are two types of edge-based segmentation methods: search and zero-crossing. In the search-based method, a measure of edge strength, which is a first-order derivative expression, is computed to detect edges by using various operators such as Sobel operator is used to find the approximate gradient magnitude at each point in an input

grayscale image [60]. In the zero-crossing based method, a second-order derivative expression is computed to detect edges. The edge method is advantageous as it is similar to how humans segment images. Yet, in mainly works well with images having a good contrast between an object and the background.

Region-based segmentation is based on grouping pixels with similar characteristics by comparing neighboring pixels. There are three region-based segmentation approaches: region growing, splitting and merging. The region growing method starts with selecting a seed pixel. Then, it merges similar pixels around the seed pixel into a region until there is no pixel left in the neighborhood. Region splitting starts with the entire image as a seed. Then, it splits the image into a predetermined number of sub-regions and uses each sub-region as a seed until all sub-regions are homogeneous. Finally, the region merging method merges any similar-enough neighborhoods based on a chosen threshold value. The disadvantage of region-based segmentation is that results generally depend on the initialization and the order in which regions are processed.

Eventually, clustering-based segmentation exploits various attributes in an image, e.g, size, color and texture. One of the most commonly used clustering methods is K-Means, since it is simple, fast and scalable. However, the disadvantage of this method is that it needs knowing in advance the number of clusters. It is suitable for convex clusters since it is a distance based partitioning method.

These days, various image analysis tools are available to both academia and the industry. These tools support image segmentation, which can be done automatically or manually. For instance, the MathWorks Image processing Toolbox[1] includes the Crop function [2], which supports both automatic and manual segmentation methods. In the automatic version, it crops only a small region from the middle of the image. On the other hand, using the manual method helps choose an area of interest. Further, a mask can be created using the Roipoly function[3] to select the region of interest (ROI) manually. As an alternative, Adobe Photoshop[4] can also be used to freely choose the ROI and avoid some image acquisition problems such as blurry areas.

Furthermore, images can be segmented based on color. Color is perceived by humans as a combination of three primary colors: Red (R), Green (G) and Blue (B). From the RGB representation, other kinds of color representations can be derived by using either linear or nonlinear transformations. To do color detection, color image segmentation techniques can be applied using different color representations. There are several color models (also called color spaces) such as RGB, hue saturation intensity (HSI), International Commission on Illumination (ICI), L*u*v* and L*a*b*. When doing color image segmentation, none of the color space gives a better result than the other spaces, for any kind of images [61]. However, the L*a*b* color model

[1] `https://www.mathworks.com/products/image.html`
[2] `https://fr.mathworks.com/help/images/ref/imcrop.html`
[3] `https://fr.mathworks.com/help/images/ref/roipoly.html`
[4] `https://www.adobe.com/fr/products/photoshop.html`

represents some colors that can be seen by the human eye easily [62]. The L*a*b* space consists of a luminosity layer (L*) that contains the brightness value of each color, a chromaticity layer (a*) indicating where the color falls along the red-green axis, and another chromaticity layer (b*) indicating where the color falls along the blue-yellow axis.

6.3 Feature Selection from Fabric Images

Fabric images may be described in terms of matrix color and inclusions shape, size, frequency and color, as illustrated by several fabric image samples in Figure ***6.2***. Yet, observations are subjective and also time-consuming. Moreover, they require some expertise. To exploit fabric images in clustering analysis, color information could be used as a feature. It can be obtained more precisely by using color detection methods than by the naked eye. For this sake, we chose the determination of inclusion colors. This feature can help to define similarity between ceramics, by constructing coherent groups of objects.

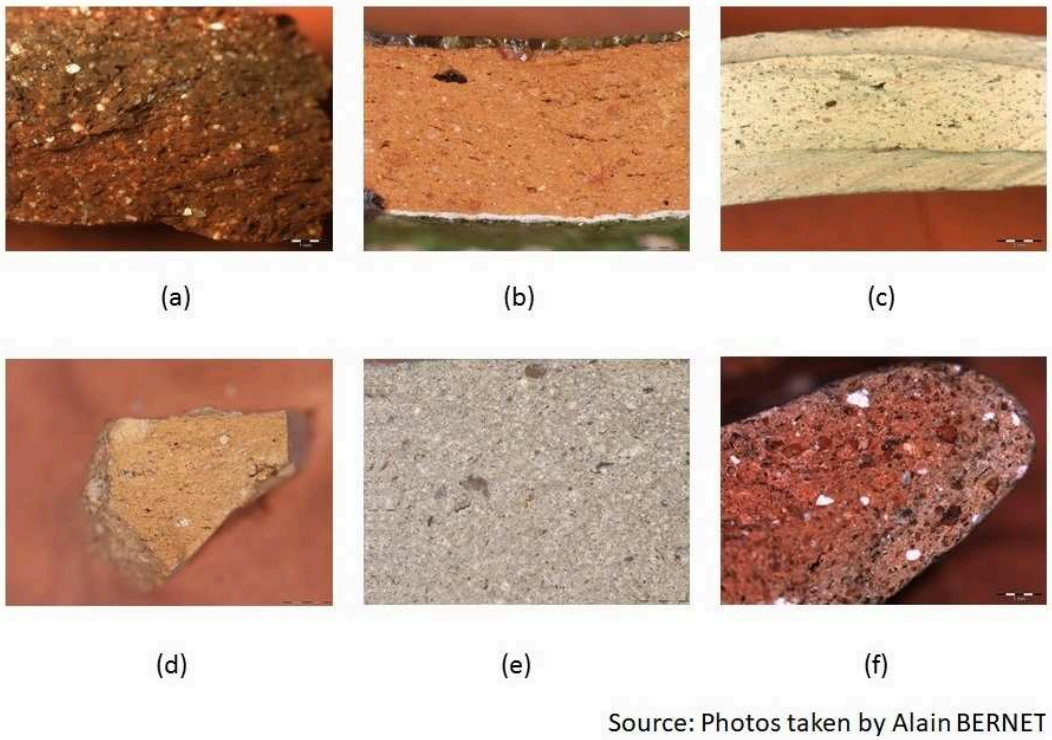

Figure 6.2: Sample fabric images

6.4 Image Segmentation of Fabric Images

Fabric image samples from Figure ***6.2*** are quite diverse. The region of interest usually does not cover the whole field, and there are some sample objects with a background (objects a, b, c, d and f). Moreover, the surfaces of ceramic samples are not flat, since these are observations on fresh cuts. It means that some parts of the fabric images may be blurry. Therefore, in most cases, images have to be preprocessed to detect a region of interest (ROI) before any other analysis can take place.

First, we apply the image segmentation method from the MathWorks Image Processing Toolbox[5] to detect a whole object. The method starts with reading fabric images in JPEG format. Then, the Sobel operator (Section 6.2), which calculates the gradient of an image, helps detect the background and the object's boundary and outputs a binary mask containing the segmented cells. The binary gradient mask shows lines that have some interior gaps on segmented cells (the object). The gaps are filled by using the Imfull function[6]. Next, the segmented object's border is removed with the Imclearborder function[7], and the Strel function[8] makes it look natural.

However, only few objects are detected correctly, because the object and the background colors are too similar. Thus, to achieve a more reliable result, we add another mask that is manually created with the Roipoly function from the MATLAB Image Processing Toolbox. The function helps select the ROI manually (Figure ***6.3***).

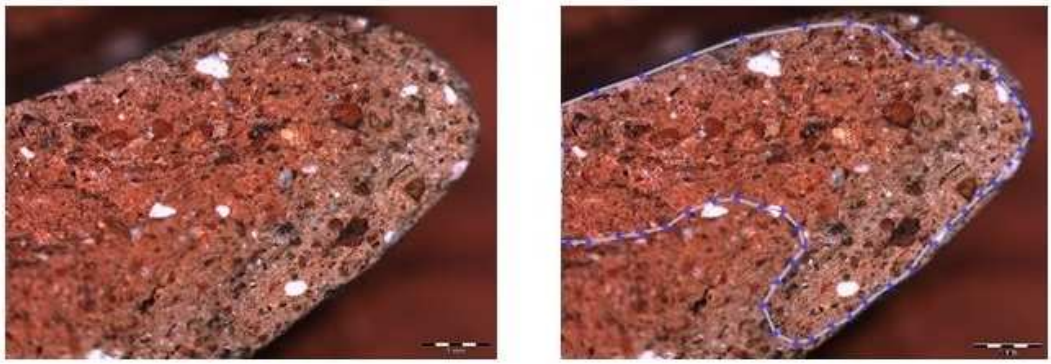

Figure 6.3: Sample fabric image with background and blurry areas (a) and segmented image (b)

[5] `https://fr.mathworks.com/help/images/examples/detecting-a-cell-using-image-segmentation.html`
[6] `https://fr.mathworks.com/help/images/ref/imfill.html`
[7] `https://fr.mathworks.com/help/images/ref/imclearborder.html`
[8] `https://fr.mathworks.com/help/images/ref/strel.html`

6.5 Color Detection of Fabric Images

To detect color, we apply the Color-Based Segmentation Using K-Means Clustering method[9] initially designed for medical images [62] onto segmented fabric images. The approach is subdivided into three steps.

The first step starts with reading fabric images in JPEG format. Then, images are converted from the RGB color space to the L*a*b* color space to soften variations in brightness and easily visually distinguish one color from another.

The second step aims at classifing colors from the a*b* space using K-Means clustering. Let us recall that K-means clustering treats each object as having a location in space, and requires a number of desired clusters and a distance metric to quantify how close two objects are to one another. Since color information exists in the a*b* space, objects are pixels with a* and b* coordinates. Using the Euclidean distance, we cluster pixels into four clusters (the number of clusters is determined empirically through experiments). For every input pixel, K-Means returns an index corresponding to a cluster. Every pixel in the image is labeled with its cluster index.

In the third step, for each clustering result, the L* layer helps extract the brightest and darkest color of each cluster. Thence, 8 different images (clustering results) are obtained from every fabric image. Figure ***6.4*** shows an example of original image and representative clustering results that are manually selected to distinguish inclusion colors. In this example, the five colors that are the most representative of inclusions are labeled manually: white, light gray, dark brown, light red and dark red. These clustering results give an opportunity to notice how easily detected colors can be visually distinguished from one another, whereas this task was much more difficult from the original image.

We apply this methodology to classify the inclusions from each sample with respect to their color. Two other features are added manually: the size of inclusions, i.e., small, medium and big; and the abundance of inclusions, i.e., no, rare, frequent, common and abundant. These features come from the description sheet of fabrics used in the ArAr laboratory (Appendix A.1). For example, in Figure ***6.4***(b), there are rare medium, rare big and no small white inclusions.

6.6 Limitations and Issues in Feature Detection

Several fabric images may be used to show different details in the same object. For instance, in the Ceramo database, some samples are associated with more than one fabric image. One of the limitations of our current work is that we use only one of them.

Moreover, image acquisition is far from ideal, because we had no control on the image acquisition process. Images were indeed obtained from various sources and in

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

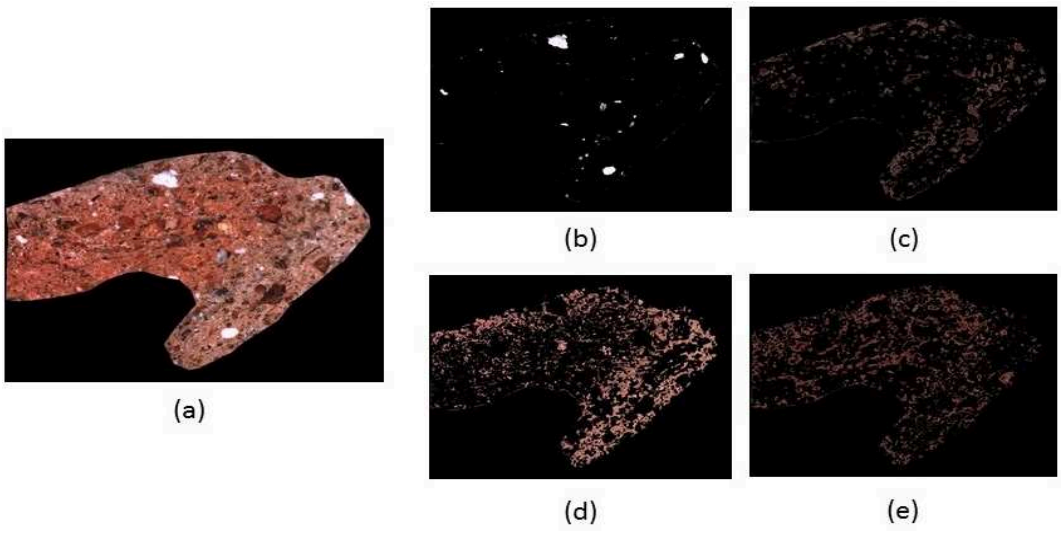

***Figure 6.4*:** Sample segmented fabric image (a); white and light gray (b), dark brown (c), light red (d) and dark red (e) color clusters

different conditions of lightning, background and camera settings. Thence, accurate comparison of such images even with a human eye is difficult.

Another limitation is that the manual selection of representative clustering results is subjective, though it definitely helps visually distinguish different inclusions and matrix colors.

6.7 Summary

Graphical representations of ceramics are useful material to extract new features that can be used in clustering analysis. In this chapter, to detect color, size and frequency of inclusions from fabric images, we use image segmentation and color detection methods. These features are used in Chapter 9.

Chapter 7

MaxMin Linear Initialization for Fuzzy C-Means

Contents

One requirement in our thesis project is to avoid the use of highly complex clustering methods. Thus, one solution is to use iterative fuzzy methods such as Fuzzy C-Means (FCM) in case of continuous data and Fuzzy K-Medoids (FKM) in case of Boolean data. To apply these two iterative methods, a primordial issue is the way of choosing K data points as initial centroids (seeds). An efficient initialization method should be linear, so that the two considered algorithm stay linear, too.

In the following, the case of continuous data (Section 7.1) is first presented and then the case of Boolean data (Section 7.2). The Section 7.3 is devoted to the new initialization method we propose. In Section 7.4, validity indices that are well suited to the fuzzy case are used to compare our proposal with several other initialization methods.

7.1 Initialization Methods for Continuous Data

In the case of continuous data, most initialization methods are studied in terms of K-Means clustering concepts [28], but these methods can also be used with Fuzzy C-Means. We have reviewed various works from the literature, including much-cited papers by Steinley and Michael [63], Maitra et al. [64] and Celebi et al. [65]. In our study, we make use of commonly mentioned linear methods from these three papers.

The first one by MacQueen (who introduced K-Means) [28] uses the first K data points as centroids. It is used by default in SPSS [66], but is sensitive to the order of data. Thus, a second method by MacQueen, which we label MacQueen2, takes K random data points as centroids.

Moreover, Faber proposes to perform multiple relaunches of MacQueen2 [67]. Among the different relaunches, the one that optimizes FW (Equation 5.2) is considered as the best candidate. This method is the standard way for initializing clusters. Its disadvantage is that outliers can be chosen. On the other hand, multiple runs ensure that the chosen sample's quality improves.

Hand et al. propose an extension of Faber's method that starts with a random set of seeds [68]. They iteratively modify the partition by randomly moving some points to other clusters. The partition minimizing FW is chosen as the best candidate. To move each data point to another random cluster, a probability α, e.g., $\alpha = 0.3$, must be set. This method is actually only interesting if the parameter α is fixed for different datasets.

Bradley and Fayyad's method starts by randomly partitioning the dataset into J subsets [69]. Then, each subset is clustered with the K-Means algorithm using MacQueen2 initialization. MacQueen2 produces J sets of centers, each containing K points. Cluster centers are combined into a superset. Then, the superset is clustered by K-Means J times. Each time, K-Means is initialized with a different center set, and members of the center set that give the lowest FW are selected as final centers.

The PCA-Part method [70] uses a divisive hierarchical approach based on PCA [38]. The method starts with a single cluster containing the whole dataset. Then, it iteratively divides clusters with respect to FW. Clusters are divided into two subclusters by using a hyperplane that is orthogonal to the principal eigenvector of the cluster covariance matrix. The division process ends after K clusters are obtained.

The K-Means++ method selects centroids based on a random sampling with unequal probabilities [71]. First, it chooses at pure random an initial center $c_1 = x$ from the data point set X. Then, it selects the next center among the remaining points of X by random sampling with unequal probabilities. The probability of each point is proportional to its distance to the nearest center.

Finally, there are in the literature other methods bearing quadratic complexity [64, 65]. Among these methods, MaxMin (also called Maximin) [72] is particularly interesting. MaxMin first calculates all the paired distances between data points. Then, it chooses two centroids from the data points, which have the greatest distance to each other. Finally, the next centroid is the data point that is the farthest from its centroid. This approach helps decrease FW, which improves the homogeneity of clusters.

The choice of the two first centers makes MaxMin quadratic. Thus, two linear versions of MaxMin have been proposed. Gonzalez suggests to pick the first center c_1 randomly, and to choose the farthest point from c_1 as the second center c_2 [73]. Unfortunately, this version depends entirely on the random choice of c_1. Its disad-

vantage is that outliers can be chosen. In contrast, Katsavounidis et al. propose to consider the global mean of data as the first center c_1 [74]. Thus, only the distance of each point to the global mean has to be calculated to determine the second center c_2, which makes the method linear. Unfortunately, the appeal to a global center is not appropriate to Boolean data.

To summarize, Hand and Krzanowski rely on user-defined parameters [68] that may not be easy to set. MacQueen2, though easy to understand and implement, uses only one random sample. Faber improves MacQueen2's random sample through relaunches. In K-Means++, the random choice is replaced by a probabilistic choice. Moreover, cluster homogeneity is taken into account. However, since probabilistic selection does not always select a large-enough distance, several probabilistic samples are required and the set of best centers is selected from all relaunches.

In contrast, MaxMin constructs only one sample by decreasing FW, and is thus deterministic. Thus, we can be sure that the chosen center is the best. Yet, MaxMin can be less effective than K-Means++ in the presence of outliers.

7.2 Initialization Methods for Boolean Data

In the case of Boolean data, instead of summarizing each cluster by its gravity center, each cluster is summarized by its most central object, which is called a prototype. Then, instead of using the K-Means or Fuzzy C-Means iterative clustering algorithms that operate on continuous data, K-Medoids and Fuzzy K-Medoids must be employed. These methods only require to take the pairwise dissimilarities between objects into account, and not the original data.

Among previously presented initialization methods, only the ones that require the calculation of a gravity center must be avoided like Katsavounidis' variant of MaxMin, which chooses the global gravity center as the first center. In K-Medoids and Fuzzy K-Medoids calculation, the gravity center is replaced with prototype.

There are few articles specifically dedicated to the initialization of Fuzzy K-Medoids. Yet, Krishnapuram et al. discuss three possible methods [75].

- **Initialization 1** is similar to MacQueen2 and chooses the first set of K-Medoids at random. It can be applied repeatedly so that the best result is output. However, Its disadvantage is that outliers can be chosen.

- **Initialization 2** chooses the first medoid as the object that is the most central point in the dataset. The sums of the dissimilarities between each point and all other points are first calculated. The point having the smallest sum is considered the most central point. Then, medoids are added one by one by successively selecting the object that is the most dissimilar from previously selected ones.

- **Initialization 3** picks the first medoid randomly, to add a random component to the procedure. Other medoids are selected as in Initialization 2. According

to Krishnapuram, apart from the Initialization 1, the other two initialization methods work well in practice.

Eventually, Park and Jun propose a method that selects initial medoids as the K points whose sum of distance-ratio to all other points is minimum [76]. The distance-ratio between two points i and j is not a simple distance, but it is the profile-line coefficient $v_{ij} = d_{ij}/d_{i+}$ (where d_{i+} denotes the sum of d_{ij} for all $j = 1, 2, ...n$) issued from the pairwise distance matrix D (whose the general term is d_{ij}, $i = 1, 2, ..., n$, $j = 1, 2, ..., n$). Park and Jun consider that the K initial medoids are the K points having the smallest value of $v_j = v_{+j}$ (v_{+j} is the sum of v_{ij} for all $i = 1, 2, ..., n$ that corresponds to the v_j of Park). Park's strategy ensures that each chosen seed is on the whole near of the different objects taking into account the more or less eccentric character of the objects. Unfortunately, this strategy does not consider the inter-seeds distances and it cannot ensure that seeds are sufficiently distant from each other to make an efficient initialization.

7.3 MaxMin Linear Specification

Among these methods, MaxMin's simplicity and ability to build homogeneous clusters is very appealing. Yet, considering all paired distance between data points makes the method quadratic with respect to the number of data points. Thus, we present in this section an enhancement of MaxMin that makes it linear. Before introducing our changes, we first detail how MaxMin works (Algorithm 1).

Algorithm 1 MaxMin

Require: Set of data points $X = \{x_1, ..., x_n\}$
Require: Number of clusters K
  {Select the first two centroids c_1 and c_2}
  $c_1, c_2 \leftarrow \arg\max(d^2(x_i, x_j))\ i, j = 1, ..., n$
  $K^* \leftarrow 2$ {Number of seeds}
  {Find the remaining seeds}
  while $K^* < K$ **do**
    for all $x_i \neq c_{k^*}\ i = 1, ..., n,\ k^* = 1, ..., K^*$ **do**
      $d_m^2(x_i) \leftarrow \min(d^2(x_i, c_{k^*}))$
    end for
    $K^* \leftarrow K^* + 1$
    $c_{K^*} \leftarrow \arg\max(d_m^2(x_i))\ i = 1, ..., n$
  end while
  return $\{c_{k^*}\}\ k^* = 1, ..., K^*$

In our variant MaxMin Linear, we first calculate grand mean $\overline{x}$ (Equation 5.4). Then, we choose as first centroid the data point that is nearest to $\overline{x}$. The second

centroid is the data point that has the greatest distance to the first centroid. Thus, complexity remains linear with respect to the number of data points. Afterwards, the choice of the remaining centroids remains the same as in MaxMin. MaxMin Linear is formalized in Algorithm 2.

Algorithm 2 MaxMin Linear

Require: Set of data points $X = \{x_1, ..., x_n\}$
Require: Number of clusters K
  {Select the first two centroids c_1 and c_2}
  $\overline{x} \leftarrow \frac{1}{n} \sum_{i=1}^{n} x_i$
  for $i \leftarrow 1$ **to** n **do**
    $d_m^2(x_i) \leftarrow \min(d^2(\overline{x}, x_i))$
  end for
  $c_1 \leftarrow \arg\min(d_m^2(x_i))\ i = 1, ..., n$
  for $i \leftarrow 1$ **to** n **do**
    $d_m^2(x_i) \leftarrow \max(d^2(c_1, x_i))$
  end for
  $c_2 \leftarrow \arg\max(d_m^2(x_i))\ i = 1, ..., n$
  $K^* \leftarrow 2$ {Number of seeds}
  {Find the remaining seeds}
  while $K^* < K$ **do**
    for all $x_i \neq c_{k^*}\ i = 1, ..., n, k^* = 1, ..., K^*$ **do**
      $d_m^2(x_i) \leftarrow \min(d^2(x_i, c_{k^*}))$
    end for
    $K^* \leftarrow K^* + 1$
    $c_{K^*} \leftarrow \arg\max(d_m^2(x_i))\ i = 1, ..., n$
  end while
  return $\{c_{k^*}\}\ k^* = 1, ..., K^*$

As a final note, the use of MaxMin Linear is not limited to FCM on numerical data. It can also be used with Fuzzy K-Medoids for categorical data clustering. Thus, MaxMin Linear can serve in fuzzy clustering ensemble on heterogeneous data. This makes of MaxMin Linear a simple but noteworthy contribution.

7.4 Experimental Validation

In this section, we aim to compare MaxMin Linear to state-of-the-art initialization methods for FCM-like clustering algorithms, i.e., MacQueen2, Faber, K-Means++ and repeated K-Means++ (retaining the best result). These methods are indeed the most common linear methods and are good representatives for random, probability and distance-based methods (Section 7.1). Moreover, they do not require any parameterization.

7.4.1 Datasets

Initialization methods are compared on 15 commonly used real-life datasets (Table ***7.1***; IDs 1-15) from the UCI Machine Learning Repository[1] and seven artificial datasets (Table ***7.1***; IDs 16-22). Their characteristics are featured in Table ***7.1***.

***Table 7.1*:** Dataset features

ID	Datasets	# of data points	# of variables	# of clusters	Sources
1	Wine	178	13	3	UCI
2	Iris	150	4	3	UCI
3	Seeds	210	7	3	UCI
4	Original Wisconsin Breast Cancer (WBCD)	683	9	2	UCI
5	Wisconsin Diagnostic Breast Cancer (WDBC)	569	30	2	UCI
6	BUPA Liver Disorder (BUPA)	345	6	2	UCI
7	Pima	768	8	2	UCI
8	Glass	214	9	6	UCI
9	Vehicle	846	18	4	UCI
10	Segmentation	2310	19	7	UCI
11	Parkinson	150	22	2	UCI
12	Movement Libras	360	90	15	UCI
13	Ecoli	336	7	8	UCI
14	Yeast	1484	8	10	UCI
15	WineQuality-Red	1599	11	6	UCI
16	Bensaid	49	2	3	[77]
17	E1071-3	150	3	3	[78]
18	Ruspini	75	2	4	[22]
19	E1071-5	250	3	5	[78]
20	E1071-3-overlapped	150	3	3	[78]
21	Ruspini_noised	95	2	4	[22]
22	E1071-5-overlapped	250	3	5	[78]

In real-life datasets, the true number of clusters is assimilated to the number of labels. Although using the number of labels as the number of clusters is debatable, it is acceptable if the set of descriptive variables explains the labels well. In artificial datasets, the number of clusters is known by construction. Moreover, we created new artificial datasets by introducing overlapping and noise into some of the existing datasets, such as E1071-3 [78], Ruspini [22] and E1071-5 [78] (Table ***7.1***; IDs 17-19).

To create a new dataset, new data points are introduced and each must be labeled. To obtain a dataset with overlapping, we modify the construction of the E1071 artificial datasets. In the original datasets, there are three or five clusters of equal size (50). Cluster i is generated according to a Gaussian distribution $N(i; 0.3)$. To increase overlapping in the three clusters while retaining the same cluster size, we only change the standard deviation from 0.3 to 0.4. Then, there is no labeling problem.

[1] `http://archive.ics.uci.edu/ml/`

To introduce noise in a dataset, we add in each cluster noisy points generated by a Gaussian variable around each label gravity center. Noisy data are often generated by distributions with positive skewness. For example, in a two-dimensional dataset, for each label, we add points that are far away from the corresponding gravity center, especially on the right-hand side, which generally contains the most points. Then, we draw a random number r between 0 and 1. If $r \leq 0.25$, the point is attributed to the left-hand side. Otherwise, the point is attributed to the right-hand side. This method helps obtain noisy data that are $^1/_4$ times smaller and $^3/_4$ times greater, respectively, than the expected value for the considered label. This process is applied to the Ruspini dataset [22].

7.4.2 Experimental Settings

In our experiments, we parameterize the FCM algorithm as follows: default termination criterion $\epsilon = 0.0001$ and default fuzziness coefficient $m = 2$. We use these default settings as we are only interested in improving the initialization of FCM. All compared methods are written in Python version 2.7.4. Repeated K-Means++ runs are performed ten times.

7.4.3 Experimental Results

In our experiments, all compared initialization methods are run with all datasets. We account for the following comparison criteria: number of iterations and quality indices the Partition Coefficient Index V_{PC}, Chen and Linkens' index V_{CL}, FB, FW, FI, the fuzzy Ratio Index V_{FRatio}, Fukuyama and Sugeno's index V_{FS} and Xie and Beni's index V_{XB} (Section 8.1). In addition to these common indices, we also use Transformed Standardized Fuzzy Difference V_{TSFD} that is a new validity index we propose (Section 8.2). This index is defined by $TSFD = FB/FI$. Further, we rank the initialization methods with respect to all criteria.

Since presenting all results would take too much space, we only present three real-life datasets, i.e., WineQuality-Red (Tables ***7.2***, ***7.3*** and ***7.4***), Glass (Tables ***7.5***, ***7.6*** and ***7.7***) and Segmentation (Tables ***7.8***, ***7.9*** and ***7.10***), as well as three artificial datasets, i.e., Bensaid (Tables ***7.11***, ***7.12*** and ***7.13***), Ruspini_noised (Tables ***7.14***, ***7.15*** and ***7.16***) and E1071-5-overlapped (Tables ***7.17***, ***7.18*** and ***7.19***), respectively. Finally, the average ranking of initialization methods on all datasets is presented in Table ***7.20***.

From these experimental results, several observations can be drawn. In regard to the number of iterations, recall that Faber and K-Means++ $\times 10$ are relaunches of two stochastic initialization methods: MacQueen2 and K-Means++, respectively. With an average ranking of 1.68 (Table ***7.20***), MaxMin Linear outperforms all other methods, including single-run methods MacQueen2 (average ranking: 1.95) and K-Means++ (average ranking: 1.95).

***Table 7.2*:** Experiment results on WineQuality-Red (1/2)

Initialization method	# of iterations	V_{PC}	V_{CL}	FB	FW
MacQueen2	45	0.664	0.7455	110972.7	1224079.7
Faber	430	0.664	0.7455	**101440.4**	1224079.7
Kmeans++	37	0.616	0.7029	101440.5	1089058.1
K-Means++ ×10	393	0.664	0.7455	**101440.4**	1224073.7
MaxMin Linear	**34**	**0.665**	**0.7458**	110972.7	**1224384.8**

***Table 7.3*:** Experiment results on WineQuality-Red (2/2)

Initialization Method	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	1335052.363	11.0305	**0.9169**	-1113107.01	0.1621
Faber	1335052.363	11.0305	0.9148	-1113107.01	0.1621
K-Means++	1190498.537	10.7359	0.9148	-987617.57	0.2388
K-Means++ ×10	1335046.425	11.0304	0.9148	-1113101.04	0.1621
MaxMin Linear	**1335357.554**	**11.0332**	**0.9169**	**-1113412.13**	**0.1611**

***Table 7.4*:** Ranking of initialization methods on WineQuality-Red

Initialization Method	# of iteration	V_{PC}	V_{CL}	FB	FW	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	3	2	2	4	2	2	2	2	2	2
Faber	5	2	2	2	2	2	2	5	2	2
K-Means++	2	5	5	3	5	5	5	3	5	5
K-Means++ ×10	4	4	4	**1**	4	4	4	4	4	4
MaxMin Linear	**1**	**1**	**1**	5	**1**	**1**	**1**	**1**	**1**	**1**

***Table 7.5*:** Experiment results on Glass (1/2)

Initialization method	# of iterations	V_{PC}	V_{CL}	FB	FW
MacQueen2	**44**	0.493	0.570	452.6	**154.1**
Faber	456	0.493	0.570	452.6	**154.1**
Kmeans++	56	0.493	0.570	452.6	**154.1**
K-Means++ ×10	366	0.493	0.570	452.6	**154.1**
MaxMin Linear	68	**0.555**	**0.645**	**508.3**	162.9

***Table 7.6*:** Experiment results on Glass (2/2)

Initialization Method	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	606.8	2.94	0.74596	-298.5	2.358
Faber	606.8	2.94	0.74597	-298.5	2.358
Kmeans++	606.7	2.94	0.74593	-298.4	2.358
K-Means++ ×10	606.7	2.94	0.74604	-298.4	2.358
MaxMin Linear	**671.2**	**3.12**	**0.75725**	**-345.4**	**0.453**

Regarding clustering result quality, MaxMin Linear obtains the best average ranking for seven of the eight experimented quality indices (Table ***7.20***). Only the FB index yields a better result for the two multiple-runs methods, while the result of MaxMin Linear is similar to those of MacQueen2 and K-Means++. However, MaxMin

Table 7.7: Ranking of initialization methods on Glass

Initialization method	# of iteration	V_{PC}	V_{CL}	FB	FW	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	**1**	2	2	2	**1**	2	2	4	2	5
Faber	5	3	3	3	2	3	3	3	3	2
Kmeans++	2	5	5	5	4	5	5	5	5	4
K-Means++ ×10	4	4	4	4	3	4	4	2	4	3
MaxMin Linear	3	**1**	**1**	**1**	5	**1**	**1**	**1**	**1**	**1**

Table 7.8: Experiment results on Segmentation (1/2)

Initialization method	# of iterations	V_{PC}	V_{CL}	FB	FW
MacQueen2	103	0.381	0.476	12384361.4	5781042.6
Faber	731	0.398	0.488	14157566.6	5680259.6
Kmeans++	146	0.381	0.476	12388277.9	5781061.6
K-Means++ ×10	930	0.399	0.490	14254025.9	**5666840.5**
MaxMin Linear	**54**	**0.430**	**0.526**	**19234921.0**	6344612.7

Table 7.9: Experiment results on Segmentation (2/2)

Initialization Method	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	18165404.0	2.14	0.6818	-6603318.7	0.363
Faber	19837826.2	2.49	0.7137	-8477307.1	0.464
Kmeans++	18169339.6	2.14	0.6818	-6607216.3	0.361
K-Means++ ×10	19920866.4	2.52	0.7136	-8587185.5	**0.341**
MaxMin Linear	**25579533.7**	**3.03**	**0.7520**	**-12890308.3**	0.656

Table 7.10: Ranking of initialization methods on Segmentation

Initialization method	# of iteration	V_{PC}	V_{CL}	FB	FW	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	2	4	5	5	3	5	5	5	5	3
Faber	4	3	3	3	2	3	3	2	3	4
Kmeans++	3	5	4	4	4	4	4	3	4	2
K-Means++ ×10	5	2	2	2	**1**	2	2	4	2	**1**
MaxMin Linear	**1**	**1**	**1**	**1**	5	**1**	**1**	**1**	**1**	5

Table 7.11: Experiment results on Bensaid (1/2)

Initialization method	# of iterations	V_{PC}	V_{CL}	FB	FW
MacQueen2	25	0.682	0.690	19222.0	**7574.9**
Faber	229	0.682	0.690	19222.7	**7574.9**
Kmeans++	25	**0.756**	**0.781**	**23421.0**	7913.4
K-Means++	218	0.682	0.690	19224.2	**7574.9**
MaxMin Linear	**10**	0.755	**0.781**	23398.0	7913.4

Linear achieves the best trade-off between FB and FW, and thus maximizes the indices that take both FB and FW into account (V_{FRatio}, V_{FS} and V_{XB}). The best result for *MaxMin Linear* is obtained with V_{TSFD} (average ranking of 1.86; Table ***7.20***), the new index specially tailored for fuzzy clustering that we propose in

***Table 7.12*:** Experiment results on Bensaid (2/2)

Initialization method	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	26796.9	2.54	0.7173	-11647.1	0.276
Faber	26797.6	2.54	0.7173	-11647.8	0.276
Kmeans++	**31334.4**	**2.96**	**0.7475**	**-15507.6**	**0.068**
K-Means++	26799.1	2.54	0.7173	-11649.3	0.275
MaxMin Linear	31311.4	**2.96**	0.7473	-15484.6	0.069

***Table 7.13*:** Ranking of initialization methods on Bensaid

Initialization method	# of iterations	V_{PC}	V_{CL}	FB	FW	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	2	3	3	5	3	5	5	5	5	5
Faber	5	5	5	4	2	4	4	4	4	4
Kmeans++	2	**1**	**1**	**1**	4	**1**	**1**	**1**	**1**	**1**
K-Means++	4	4	4	3	**1**	3	3	3	3	3
MaxMin Linear	**1**	2	2	2	5	2	2	2	2	2

***Table 7.14*:** Experiment results on Ruspini_noised (1/2)

Initialization method	# of iterations	V_{PC}	V_{CL}	FB	FW
MacQueen2	9	0.775121	0.806518	219099.6	23421.0260
Faber	130	0.775125	0.806517	219100.8	23421.0258
Kmeans++	13	0.775122	0.806521	219101.1	23421.0258
K-Means++ ×10	105	**0.775128**	0.806518	219102.3	**23421.0256**
MaxMin Linear	**7**	0.775128	**0.806523**	**219105.4**	23421.0268

***Table 7.15*:** Experiment results on Ruspini_noised (2/2)

Initialization Method	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	242520.7	9.3548	0.903427	-195678.6	0.063680
Faber	242521.9	9.3549	0.903427	-195679.8	0.063681
Kmeans++	242522.1	9.3549	0.903427	-195680.0	0.063676
K-Means++ ×10	242523.3	9.3549	0.903426	-195681.3	0.063681
MaxMin Linear	**242526.4**	**9.3551**	**0.903429**	**-195684.4**	**0.063672**

***Table 7.16*:** Ranking of initialization methods on Ruspini_noised

Initialization Method	# of iteration	V_{PC}	V_{CL}	FB	FW	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	2	5	3	5	4	5	5	4	5	3
Faber	5	3	5	4	2	4	4	3	4	5
Kmeans++	3	4	2	3	3	3	3	2	3	2
K-Means++ ×10	4	**1**	4	2	**1**	2	2	5	2	4
MaxMin Linear	**1**	2	**1**	**1**	5	**1**	**1**	**1**	**1**	**1**

Section 8.2.

In conclusion, the results obtained with MaxMin Linear are a little better than those obtained with multiple-runs methods, but they require ten times fewer iterations. Moreover, MaxMin Linear is deterministic, whereas multiple-runs methods are

Table 7.17: Experiment results on E1071-5-overlapped (1/2)

Initialization method	# of iterations	V_{PC}	V_{CL}	FB	FW
MacQueen2	8	0.735646	0.762681	219.7337	48.715631
Faber	103	0.735645	0.762683	219.7358	48.715630
Kmeans++	12	0.735651	0.762685	219.7408	48.715632
K-Means++ ×10	113	0.735645	0.762683	219.7363	**48.715629**
MaxMin Linear	**7**	**0.735652**	**0.762688**	**219.7445**	**48.715629**

Table 7.18: Experiment results on E1071-5-overlapped (2/2)

Initialization Method	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	268.4494	4.5105	0.818530	-171.0181	0.11574
Faber	268.4514	4.5106	0.818535	-171.0202	**0.11569**
Kmeans++	268.4565	4.5107	0.818534	-171.0252	0.11575
K-Means++ ×10	268.4519	4.5106	0.818530	-171.0207	**0.11569**
MaxMin Linear	**268.4601**	**4.5108**	**0.818537**	**-171.0288**	0.11572

Table 7.19: Ranking of initialization methods on E1071-5-overlapped

Initialization method	# of iteration	V_{PC}	V_{CL}	FB	FW	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	2	3	5	5	4	5	5	5	5	4
Faber	4	5	4	4	3	4	4	2	4	**1**
Kmeans++	3	2	2	2	5	2	2	3	2	5
K-Means++ ×10	5	4	3	3	**1**	3	3	4	3	2
MaxMin Linear	**1**	**1**	**1**	**1**	2	**1**	**1**	**1**	**1**	3

Table 7.20: Average ranking of initialization methods on all datasets

Initialization method	# of iteration	V_{PC}	V_{CL}	FB	FW	FI	V_{FRatio}	V_{TSFD}	V_{FS}	V_{XB}
MacQueen2	1.95	3.36	3.55	3.86	3.41	3.41	3.41	3.04	3.41	3.55
Faber	4.45	2.73	2.82	1.73	2.73	2.73	2.73	3.27	2.73	2.91
K-Means++	1.95	3.86	3.68	3.86	3.86	3.86	3.86	3.54	3.86	3.36
K-Means++ ×10	4.41	2.68	2.55	**1.64**	2.86	2.86	2.86	3.22	2.86	2.82
MaxMin Linear	**1.68**	**2.27**	**2.32**	3.82	**2.05**	**2.05**	**2.05**	**1.86**	**2.05**	**2.27**

stochastic.

7.5 Summary

In this chapter, we propose a new, fast and easy to implement initialization method for FCM called MaxMin Linear. We experimentally compare MaxMin Linear to several initialization methods from the literature and shown that it outperforms existing methods on 22 synthetic and real-life datasets.

In addition, MaxMin Linear can be applied to algorithms other than FCM, such as Fuzzy K-Modes and Fuzzy K-Medoids, which apply on categorical data. In particular, MaxMin Linear allows decreasing the complexity of Park and Jun's FKM

implementation [76].

Chapter 8

A Visual Quality Index for Fuzzy C-Means

Contents

Clustering algorithms behave differently for different reasons. The first reason relates to dataset features such as geometry and the density distribution of clusters. The second reason is the choice of input parameters such as fuzziness coefficient m ($m = 1$ indicating that clustering is crisp and $m > 1$ that clustering becomes fuzzy), the number of clusters K and initialization method.

These parameters all affect the quality of clustering. To study how the choice of parameters impacts clustering quality, we need a quality criterion. For instance, when the dataset is well separated and has only two variables, a scatter plot can help determine the number of clusters. However, when the dataset has more than two variables, a good quality index is needed to compare various cluster configurations and choose the appropriate number of clusters.

Achieving a good clustering involves both minimizing intra-cluster distance (compactness) and maximizing inter-cluster distance (separability). A common issue in this process is that clusters are split up while they could be more compact. Many cluster quality indices have been proposed to address this problem for hard and fuzzy clustering, but none of them is always highly efficient [79].

Moreover, there is no real-life golden standard for clustering analysis, since various experts may have different points of views about the same data and express different constraints on the number and size of clusters. Thanks to a visual index (for example, the graph which considers the variations of the quality index according the number of clusters), different solutions can be presented with respect to the data. Thus, experts

can make a trade-off between their opinion and the best local solutions proposed by the visual index. Hence, in this chapter, we propose an innovative, visual quality index for FCM.

8.1 Fuzzy Clustering Quality Indices

According to Wang et al., there are two groups of quality indices [80]:

1. quality indices based only on membership values;
2. quality indices that associate membership values to cluster centers and data.

Quality indices in the first group notably include the Partition Coefficient Index V_{PC} [81] (Equation 8.1; $\frac{1}{K} \leq V_{PC} \leq 1$; to be maximized) and Chen and Linkens' index V_{CL} [82] (Equation 8.2; $0 \leq V_{CL} \leq 1$; to be maximized) takes both compactness (first term of V_{CL}) and separability (second term of V_{CL}) into consideration.

$$V_{PC} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^2 \tag{8.1}$$

$$V_{CL} = \frac{1}{n} \sum_{i=1}^{n} max_k(u_{ik}) - \frac{1}{c} \sum_{k=1}^{K-1} \sum_{j=k+1}^{K} \left[\frac{1}{n} \sum_{i=1}^{n} min(u_{ik}, u_{ij}) \right] \tag{8.2}$$

where $c = \sum_{k=1}^{K-1} k$.

The disadvantage of these indices is that they do not take x value into consideration.

Quality indices in the second group include an adaptation of the Ratio Index V_{FRatio} to fuzzy clustering [83] (Equation 8.3; $0 \leq V_{FRatio} \leq +\infty$; to be maximized), Fukuyama and Sugeno's index V_{FS} [84] (Equation 8.4; $-FI \leq V_{FS} \leq FI$; to be minimized), and Xie and Beni's index V_{XB} [85,86] (Equation 8.5; $0 \leq V_{XB} \leq FI/n * min\|x_j - v_k\|^2$; to be minimized).

$$V_{FRatio} = FB/FW \tag{8.3}$$

$$V_{FS} = FW - FB \tag{8.4}$$

$$V_{XB} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik}^m \|x_i - v_k\|^2}{n * min_{j,k} \|v_j - v_k\|^2} \tag{8.5}$$

These indices are built on the associated notions of fuzzy within and fuzzy between inertia. Their advantage is that they take fuzzy coefficient u_{ij} and x value into

consideration. These indices are well suited to hard clustering ($m > 1$) because in this case, $FI = FW + FB$ has a constant value. Unfortunately, in case of fuzzy clustering, FI is varying. For example, considering V_{FS}, $FS = FW - FB$, it would be logical to express $FW - FB$ in function of FI.

When the number of clusters increases, the value of quality indices mechanically increases, too. Then, the important question is: how useful is the addition of a new cluster? To answer this question, the most common solutions are penalization and the Elbow Rule [87].

The first way to penalize a quality index is to multiply it by a quantity that diminishes the index' value when the number of clusters increases. In this case, the main difficulty is to choose the penalty. For instance, the penalized version of V_{FRatio} is Calinski's index V_{FCH} [83] (Equation 8.6; $0 \leq V_{FCH} \leq +\infty$; to be maximized), where the penalty is based on both the number of clusters and data points.

$$V_{FCH} = \frac{FB/(K-1)}{FW/(n-K)} = \frac{n-K}{K-1}\frac{FB}{FW} \tag{8.6}$$

The second way to penalize a quality index is to evaluate index evolution relatively to the number of clusters, by considering the curve of the index' successive values. The most appropriate value of K can be determined visually with the help of either the Elbow Rule or an algebraic calculation [88].

To construct a visual determination of the Elbow Rule, K is represented on the horizontal axis and the considered quality index on the vertical axis. Then, we look for the value of K where there is a change in the curve's concavity. This change corresponds to the optimal number of clusters K.

To construct an algebraic determination, let i_K be the index value for K clusters. The variation of i_K before K and after K are compared. In case of a positive Elbow, the second difference $min_K((i_{K+1} - i_K) - (i_K - i_{K-1}))$ is minimized. Moreover, since the values before K and after K are used for calculation, the Elbow Rule can be applied to more than two clusters.

Among all above-described quality indices, there is no single quality index that gives the best result for any dataset. Thus, there is room for a new quality index that is specifically tailored for fuzzy validation and helps the user choose the value of K.

8.2 An Index Associated with a Visual Solution

The optimal number of clusters can be determined by considering the variation of a clustering validity index. There are two possible cases:

1. if the index is not monotonic with the number of clusters, we choose the number of clusters that optimizes the index;

2. if the index is monotonic, we may use a penalized version of the index.

Building a new quality index, we first consider FW to evaluate compactness and FB to evaluate separability. We can choose to calculate either $FB - FW$, which is similar to V_{FS} except for the sign, or $FB \div FW$, which is similar to V_{FRatio}. Unfortunately, $FI = FB + FW$ is not constant and $FB - FW \in [-FI, +FI]$. To take this particularity of fuzzy clustering into account, we propose to standardize $FB-FW$ by considering the Standardized Fuzzy Difference $SFD = (FB-FW) \div FI$ instead. Then, $SFD \in [-1, +1]$.

Moreover, adding a new cluster often improves clustering quality mechanically. Thus, many authors penalize the quality index with respect to K (the smaller n is, the greater the penalty), e.g., V_{FCH} (Section 8.1). To obtain a penalized index, we first linearly transform SFD so that its values belong to $[0, 1]$, thus obtaining the Transformed Standardized Fuzzy Difference $TSFD$ (Equation 8.7; $TSFD \in [0, 1]$; to be maximized).

$$TSFD = \frac{1 + SFD}{2} = \frac{FB}{FI} \tag{8.7}$$

Finally, by penalizing $TSFD$ as V_{FCH}, we obtain the Penalized Standardized Fuzzy Difference $PSFD$ (Equation 8.8; $PSFD \in [0, (n - K) \div (K - 1)]$; to be maximized).

$$PSFD = TSFD \times \frac{n - K}{K - 1} = \frac{FB - FW}{FI} \times \frac{n - K}{K - 1} \tag{8.8}$$

Instead of penalizing the quality index, another approach is to visualize the search for the best number of clusters K. The first solution is to apply the Elbow Rule to $TSFD$. $TSFD$ is plotted with respect to K in Figure ***8.1***(a). The drawback of this method is that the horizontal axis corresponds to an arithmetic scale of K values, which is not satisfying.

To fix this problem, we suggest to plot FB with respect to FI for considering each value of K, which we call Visual $TSFD$. Our aim is not to give an automatic solution, but to help the user visually choose the most appropriate value for K. The visualization we propose is shown in Figure ***8.1***(b), where the blue line plots $TSFD$ with respect to K, the full red line is the diagonal that corresponds to the best solutions ($FB = FI$) such that $TSFD = 1$, and the dashed red line connects the origin to each point associated with K values.

The smaller the angle between the full red line and the dashed red line, the better is the solution. As the value of K increases, the angle between the dashed red line and the diagonal decreases. Then, we choose the value of K beyond which the decrease becomes negligible. This value is considered as the optimal number of clusters. For example, in Figure ***8.1***(b), a first solution could be $K = 4$, a better solution $K = 6$ and it is not very interesting to consider $K > 6$.

In case of Boolean data, because gravity centers are replaced by prototypes, $FB + FW$ is not necessary equal to FI. Thus, in this case, $TSFD$ is defined by $TSFD = FB/(FW + FB)$.

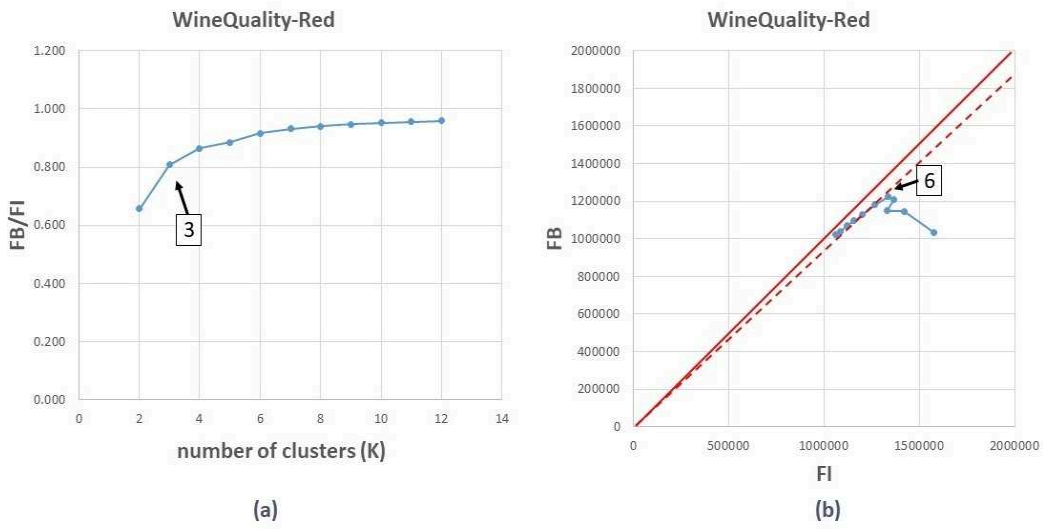

Figure 8.1*:** Comparison of Elbow Rule (a) and Visual $TSFD$ (b) on the WineQuality-Red dataset (see Table ***8.1)

8.3 Experimental Validation

In this section, we compare our proposals $TSFD$, $PSFD$, Visual $TSFD$ and the use of the Elbow Rule to state-of-the-art quality indices for FCM-like clustering algorithms, i.e., V_{PC}, V_{CL}, V_{FCH}, V_{FS} and V_{XB} (Section 8.1) that are indeed the most common quality indices applied on several numerical real-world and artificial datasets.

8.3.1 Datasets

Quality indices are compared on ten real-life datasets (Table ***8.1***; IDs 1-10) from the UCI Machine Learning Repository[1] and seven artificial datasets (Table ***8.1***; IDs 11-17). These datasets are those already used in Section 7.4.1, except the datasets having an optimal number of clusters $K = 2$, because computing the Elbow rule necessitates that $K \geq 3$.

8.3.2 Experimental Settings

In our experiments, the FCM algorithm is parameterized with its default settings: termination criterion $\epsilon = 0.0001$ and default fuzziness coefficient $m = 2$. We use these default settings because we are only interested in improving quality indices. All clustering quality indices are coded in Python version 2.7.4.

[1] `http://archive.ics.uci.edu/ml/`

8.3.3 Experimental Results

In these experiments, we test all quality indices against all datasets. Our comparison criterion is the number of clusters achieved by each quality index. We also rank the indices with respect to real-life and artificial datasets wins criteria.

Our experimental results are shown in Table ***8.1***, where the first four columns recall the datasets' features and the next columns the number of clusters computed with the tested indices. Moreover, since presenting all the results would take too much space, we retain only the indices which have the best result.

***Table 8.1*:** Quality index experiment results

ID	Datasets	# of data points	# of clusters	V_{PC}	V_{CL}	FB	V_{FCH}	V_{FS}	V_{XB}	Elbow V_{TSFD}	Visual V_{TSFD}
1	Wine	178	3	2	2	8	12	8	2	**3**	5
2	Iris	150	3	2	2	**3**	**3**	**3**	2	**3**	**3**
3	Seeds	210	3	2	**3**	**3**	**3**	**3**	2	**3**	**3**
4	Glass	214	6	2	2	12	12	12	2	4	5, 7
5	Vehicle	846	4	2	2	2	2	5	2	3	**4**, 5
6	Segmentation	2310	7	2	4	4	4	12	12	3	**7**, 8
7	Movement Libras	360	15	2	18	16	16	18	2	14	14, 16
8	Ecoli	336	8	2	3	3	3	12	3	3	3, 7
9	Yeast	1484	10	2	2	5	2	12	2	4	7, 8
10	WineQuality-Red	1599	6	2	2	**6**	7	**6**	2	3	**6**
11	Bensaid [77]	49	3	**3**	**3**	9	11	11	**3**	**3**	5
12	E1071-3 [78]	150	3	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
13	Ruspini [22]	75	4	**4**	**4**	**4**	**4**	**4**	**4**	3	**4**
14	E1071-5 [78]	250	5	2	**5**	4	**5**	**5**	2	3	**5**
15	E1071-3-overlapped	150	3	2	**3**	**3**	2	**3**	2	**3**	**3**
16	Ruspini_noised	95	4	**4**	12	**4**	**4**	**4**	**4**	**4**	**4**
17	E1071-5-overlapped	250	5	2	2	4	**5**	4	2	3	**5**
# of wins for real-life datasets				0	1	3	2	3	0	3	**5**
# of wins for artificial datasets				4	5	4	5	5	4	4	**6**
Total # of wins				4	6	7	7	8	4	7	**11**

As shown in Table ***8.1***, it is more difficult to predict an appropriate number of clusters for real-life datasets than for artificial datasets. Considering all indices, the average rate of success is indeed 21% in the case of real data, against 66% in the case of artificial data. Whatever the type of data, Visual $TSFD$ outperforms the other indices, with 5 wins out of 10 in the case of real datasets and 6 wins out of 7 in the case of artificial datasets. The worst results are obtained with V_{PC} and V_{XB} (0/10 and 4/7 wins each). The other indices achieve intermediary results.

In addition, when the value given by Visual $TSFD$ is erroneous, it is quite close to the expected K, in contrast to V_{FS}, our closest competitor (Table ***8.1***; datasets Wine, Glass, Segmentation, Ecoli and Bensaid). For example, the optimal number of clusters should be 6 for the Glass dataset. $V_{FS} = 12$ and Visual $TSFD$'s results are 5 and 7.

Furthermore, we compare in Figures ***8.2*** and ***8.3*** Visual $TSFD$ and the plot obtained with the Elbow Rule (which is labeled Elbow $TSFD$) with respect to K, on a sample of both real-life and artificial datasets bearing different characteristics, i.e., Glass, Vehicle, Ecoli, Ruspini, Ruspini_noised and E1071-5-overlapped (Table ***8.1***).

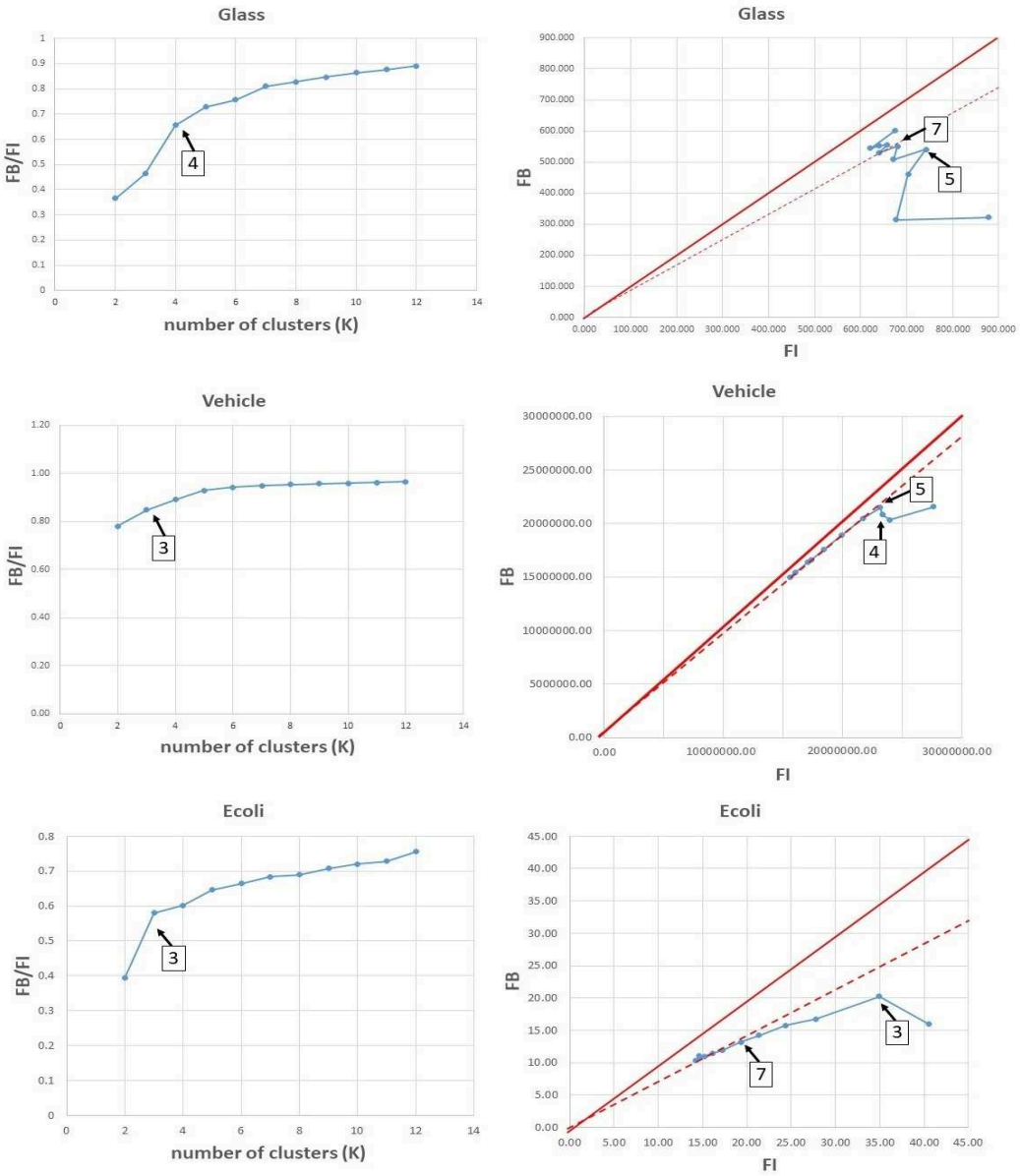

***Figure 8.2*:** Comparison of Elbow $TSFD$ and Visual $TSFD$ (1/2)

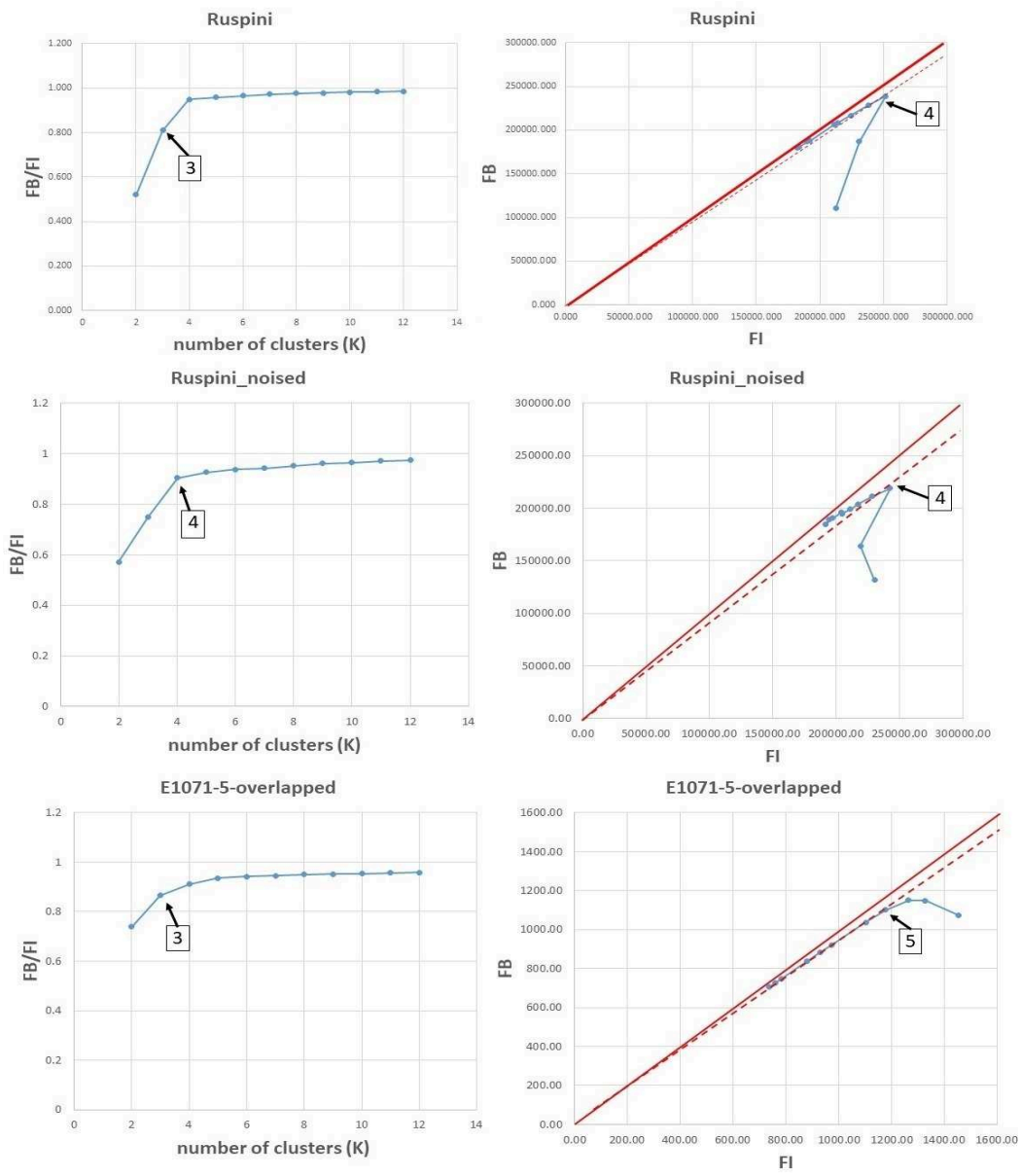

***Figure 8.3*:** Comparison of Elbow $TSFD$ and Visual $TSFD$ (2/2)

As is clearly visible from Figures ***8.2*** and ***8.3***, Visual $TSFD$ provides a better visual idea than Elbow $TSFD$ to determine K. Elbow $TSFD$ indeed highlights K values of 3 or 4, while the $TFSD$ blue plot systematically indicates larger and more relevant K values.

Eventually, since our work aims at real-life datasets, there is no ground truth nor golden standard for clustering analysis. In such a context, Visual $TSFD$ has the advantage of providing options to experts, instead of outputting a single K value. This makes our method more flexible than the existing ones in real-life scenarios.

8.4 Summary

In this chapter, we propose a novel quality index for FCM called Visual $TSFD$, which helps determine the number of clusters visually. We experimentally compare Visual $TSFD$ to state-of-the-art clustering quality indices and show that it outperforms existing indices on various datasets.

Furthermore, Visual $TSFD$ can also be used in the case of categorical data with Fuzzy K-Medoids [76]. Thus, Visual $TSFD$ allows dealing with heterogeneous datasets, which is particularly interesting in our applicative context.

As a result, our next step is to design a fuzzy clustering method based on Visual $TSFD$ that can deal with both numerical and categorical data.

Chapter 9

Disjoint Clustering on Archaeological and Archaeometric Data

Contents

As stated in Chapter 3, the archaeological and archaeometric data we are dealing with are of three types: chemical (numerical), descriptive (textual) and graphical (images). There has been a tremendous growth in the availability of archaeological and archaeometric data through the creation of ceramic databases. In the near future, such data will be considered as large (great number of objects) high-dimensional (great number of variables) data. In this chapter, our aim is to propose data mining techniques that are well-suited to deal with real-life, large, high-dimensional and heterogeneous data, while considering the specificities of data related to archaeological ceramic samples.

First, we present the hard clustering method used by archaeometry experts from the ArAr laboratory to construct groups of ceramic objects, which we will use as

reference. Then, we apply fuzzy clustering on chemical, description and image data separately (Figure ***9.1***), and compare the results with the expert-defined groups to examine their coherence.

Our experiment steps are shown in Table ***9.1***. The first step is data preprocessing such as for continuous data standardization or for Boolean data reduction. The second step is to determine the optimal number of clusters to use in FCM. The third step is to apply FCM algorithm to obtain results that are used to do interpretations by comparing with expert-defined groups at the end of the experiment.

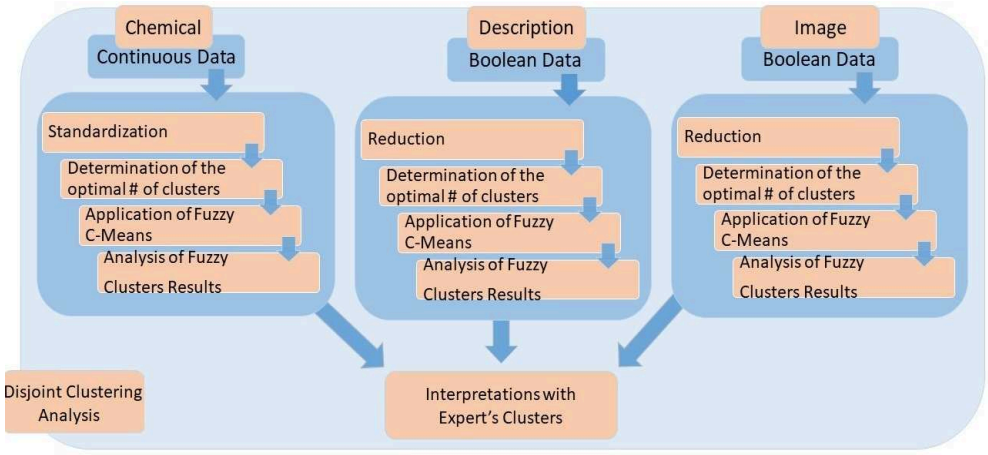

Figure 9.1: Disjoint clustering analysis scheme

9.1 Dataset and Expert-Defined Groups

Within the ArAr laboratory, archeometry experts define ceramic objects with respect to their chemical composition and other information, including descriptive and image data. To define groups of ceramic objects, experts first apply a hierarchical clustering algorithm (HCA; Section 5.1) to chemical data. They use the dendrogram output by HCA as a help to identify the structure of raw chemical data, which they further examine to define different groups of ceramics, also taking various factors into consideration, such as geological contexts, geochemical properties, analytical biases, as well as descriptive and image data [3].

The options chosen for HCA are to use standardized variables (Equation 5.8) and to calculate the dissimilarity between objects with the Euclidean distance. Two objects are clustered together when they minimize a given agglomeration criterion obtained by the centroid linkage method [10]. In this method, the distance between two clusters is calculated as the distance between the two mean vectors of the clusters. At each stage, the two clusters that have the smallest centroid distance are combined.

The overall time complexity of centroid linkage HCA is $O(n^2 \log n)$.

A dendrogram initially represents each object as a point on a horizontal baseline. Similar objects and groupings of objects are connected by a horizontal line whose intersection with the vertical axis gives the level of dissimilarity. Thus, HCA results are in the form of hierarchies. To identify clusters in a dendrogram, the usual procedure is to draw a horizontal line at a chosen threshold and to retain only the connections whose level is lower than the threshold. However, due to varying homogeneity in each group, ArAr experts do not use a single threshold for the whole dendrogram. They do not rely on the structure of the dendrogram only, but also use their knowledge (about the raw data, their properties and descriptive features) to choose the thresholds that are appropriate for each group.

Table ***9.1*** summarizes the features of our initial experimental dataset, which consists of 3300 ceramic objects. However, only 301 objects include all three types of descriptors (chemical, description and image data). Expert-defined groups are known for all these 301 objects.

***Table 9.1*:** Archaeological ceramic dataset features

Data	**#objects**	**#variables**	**#variables after binarization**	**#variables standardized ⋆ or reduced ‡**
Chemical	3141	24		17 ⋆
Description	3300	6	63	58 ‡
Image	415	30	150	119 ‡

Figure ***9.2*** displays the dendrogram obtained using HCA on the 301-object corpus. It is divided into three pieces to fit in the page. Expert-defined groups are underlined and their names indicated. It must be stressed that the initial expert-groups as defined in the ArAr laboratory contain more objects (a minimum of 15 to 20 in general). The expert-groups appear reduced in size in our corpus due to the fact that not all the objects had images data attached to them, and only those which had some were considered.

The dendrogram shows that most samples are well grouped. It is clearly visible that objects in the same group are consecutive, while only a few samples lie out of the expert-defined groups. Only five objects that are in a marginal position do not integrate the expert-defined groups: one object labeled 1 (Lapithos group) and four objects labeled 2 (Plaka_LRA1 group). Group Plaka_LRA1 is split, with four samples in one group, while the other samples appear as marginal to other groups.

We can also notice that not all groups bear the same homogeneity. Some groups are connected at a higher level, while other groups are connected at a lower level according to expert choice. If clusters are determined automatically from the dendrogram, a global threshold is chosen and the resulting clusters are different than expert-defined groups. The experts may propose different levels that are better suited to the nature of the studied objects.

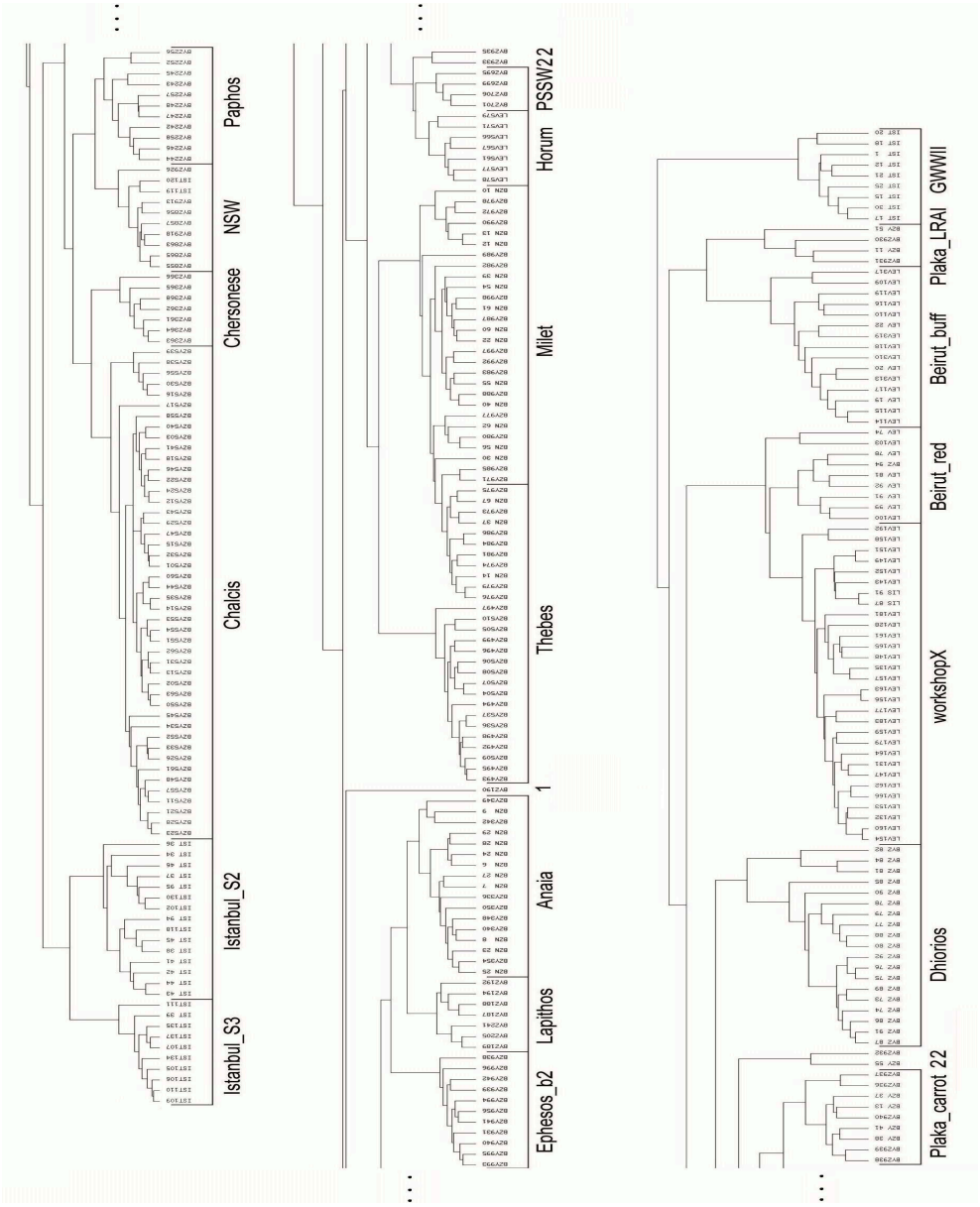

***Figure 9.2*:** Dendrogram obtained by HCA clustering (S.Y. Waksman)

9.2 Chemical Data

9.2.1 Fuzzy C-Means Parameters

This experiment consists of several steps (Figure ***9.1***) that are detailed below.

Data Preprocessing To give each variable an equivalent weight, chemical variables are standardized using Equation 5.8.

Determination of the Optimal Number of Clusters Before FCM is applied to standardized chemical data, the optimal number of clusters is determined by our Visual $TSFD$ method (Chapter 8). The resulting plot (Figure ***9.3***) shows that the best K value for chemical data is clearly 20, since the following K values do not change the angle between the dashed line and the full line.

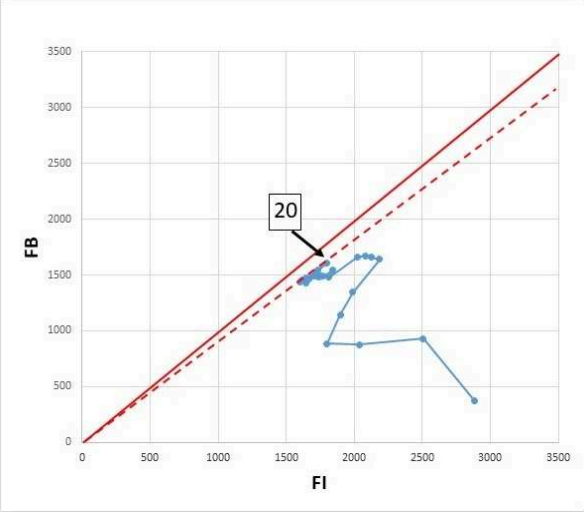

***Figure 9.3*:** Cluster number determination (chemical data)

9.2.2 Fuzzy C-Means Results

We apply FCM to standardized chemical data with $K = 20$ to obtain the fuzzy coefficient matrix. Then, we identify for each object its first three largest fuzzy coefficients and sort the objects according to the identifiers of the three closest centers (Table ***9.2***). Analyzing this table reveals the limited diversity of centers of order two and three associated with each first center.

Considering only the main center associated with each object, a new table is constructed which gives for each main center the list of groups having at least one object associated with the considered center (Table ***9.3***). Analyzing this table allows to evaluate the greater or lesser homogeneity of each center in terms of expert-defined groups.

Lastly, we present the main fuzzy clustering centers that are associated with expert-defined groups and their frequency (Table ***9.4***). Analyzing this table reveals the different main centers associated with each group.

***Table 9.2*:** Centers of order 1–3 in chemical dataset

Cntr 1	**Cntr 2**	**Cntr 3**	**Expert-group**	**# of objects**
c1	c8	c18	Plaka_LRA1	3
c2	c13	c17	GWWII	8
c2	c13	c19	GWWII	1
c3	c9	c20	Dhiorios	3
c4	c19	c12	Beirut_red	8
c5	c10	c8	Thebes	1
c5	c10	c15	Thebes	16
c6	c14	c8	Beirut_buff	8
c6	c14	c18	Beirut_buff	2
c7	c11	c16	Chalcis	4
c7	c11	c20	Chalcis	5
c7	c11	c20	Chersonese	1
c7	c16	c11	Chalcis	8
c7	c16	c17	Chalcis	14
c7	c16	c20	Chersonèse	5
c7	c16	c20	Chalcis	8
c7	c17	c16	Chalcis	4
c7	c20	c11	Lapithos	1
c7	c20	c16	Chalcis	3
c7	c20	c16	Chersonese	1
c8	c5	c1	Plaka_LRA1	1
c8	c10	c5	Plaka_LRA1	1
c8	c18	c1	Plaka_LRA1	2
c8	c18	c10	Horum	2
c8	c18	c10	Plaka_LRA1	1
c8	c18	c10	PSSW	4
c8	c18	c15	Horum	5
c9	c3	c17	Dhiorios	1
c9	c3	c20	Dhiorios	4
c9	c16	c3	Dhiorios	1
c9	c16	c17	Dhiorios	3
c9	c17	c16	Dhiorios	1
c9	c20	c3	Dhiorios	2
c9	c20	c16	Dhiorios	3
c9	c20	c17	Dhiorios	1
c10	c15	c5	Milet	10
c10	c15	c8	Milet	5
c10	c15	c16	Milet	2
c10	c15	c18	Milet	2
c11	c7	c20	Ephesos_b2	3
c11	c20	c7	Ephesos_b2	7
c11	c20	c15	Ephesos_b2	1
c12	c19	c4	workshopX	19
c12	c19	c13	workshopX	1
c13	c17	c7	Istanbul_S3	1
c13	c17	c12	Istanbul_S3	2
c13	c17	c16	Istanbul_S3	3
c13	c17	c19	Istanbul_S3	4
c14	c6	c8	Beirut_buff	1
c14	c6	c18	Beirut_buff	4
c15	c10	c8	Milet	9
c15	c10	c11	Milet	1
c15	c10	c16	Milet	2
c15	c10	c20	Milet	8
c16	c7	c17	NSW	1
c16	c7	c17	Paphos	7
c16	c7	c18	NSW	4
c16	c7	c18	Paphos	1
c16	c7	c20	NSW	2
c16	c17	c7	Paphos	1
c16	c18	c7	NSW	3
c16	c18	c10	Paphos	2
c17	c7	c13	Istanbul_S2	2
c17	c7	c16	Istanbul_S2	3
c17	c7	c18	Istanbul_S2	2
c17	c13	c7	Istanbul_S2	2
c17	c13	c16	Istanbul_S2	5
c17	c16	c7	Istanbul_S2	1
c18	c8	c10	Plaka_carrot	3
c18	c8	c16	Plaka_carrot	4
c18	c8	c17	Plaka_carrot	1
c18	c16	c8	Plaka_carrot	1
c19	c12	c4	Beirut_red	1
c19	c12	c4	workshopX	10
c20	c7	c11	Lapithos	1
c20	c11	c7	Anaia	5
c20	c11	c7	Lapithos	1
c20	c11	c15	Anaia	5
c20	c15	c7	Anaia	2
c20	c15	c7	Lapithos	1
c20	c15	c7	Lapithos	1
c20	c15	c10	Anaia	1
c20	c15	c10	Anaia	1
c20	c15	c10	Anaia	1
c20	c15	c10	Anaia	1
c20	c15	c16	Anaia	1
c20	c16	c7	Lapithos	2
c20	c16	c15	Lapithos	1
Total				301

9.2.3 Comparison with expert-defined groups

Results from Section 9.2.2 show that there is a strong link between chemical data clustering and the ceramic groups defined by experts. We can consider different types of experts-defined groups depending on the number of main centers associated with each given expert-defined group (see Table ***9.4***) and on the number of heterogeneous

***Table 9.3*:** Centers of order 1 in chemical dataset

Center	Expert-group	Frequency	
c1	Plaka_LRA1	3	3
c2	GWWII	9	9
c3	Dhiorios	3	3
c4	Beirut_red	8	8
c5	Thebes	17	17
c6	Beirut_buff	10	10
c7	Chalcis	46	
	Chersonese	7	
	Lapithos	1	54
c8	Horum	7	
	Plaka_LRA1	5	
	PSSW	4	16
c9	Dhiorios	16	16
c10	Milet	19	19
c11	Ephesos_b2	11	11
c12	workshopX	20	20
c13	Istanbul_S3	10	10
c14	Beirut_buff	5	5
c15	Milet	20	20
c16	Paphos	11	
	NSW	10	21
c17	Istanbul_S2	15	15
c18	Plaka_carrot	9	9
c19	Beirut_red	1	
	workshopX	10	11
c20	Anaia	17	
	Lapithos	7	24
		Total	301

centers among them (see Table ***9.3***), as summarized in Table ***9.5***. For example, let us consider Plaka_LRA1. There are 2 main centers, c1 and c8, associated with Plaka_LRA1 (see Table ***9.4***), while only 1 of these 2 centers, i.e., c8 is heterogeneous. Thus, Plaka_LRA1 belongs to the type 2.1. Considering different types of groups helps to better organize the experts-defined groups.

Based on the group types from Table ***9.5***, Ephesos_b2, GWWII, Istanbul_S2, Istanbul_S3, Plaka_carrot and Thebes are expert-defined groups of type 1.0 for chemical data (Table ***9.4***). Thus, if we know the main center of any object, we can use this information to find the expert-defined group and *vice versa*.

Group type 2.0 corresponds to expert-defined groups Beirut_buff, Dhiorios and Milet. This type of group is similar to type group 1.0, because both centers are exclusive.

Group type 1.1 corresponds to expert-defined groups Anaia, Chalcis, Chersonese, Horum, NSW, Paphos and PSSW. Here, it is not possible to guess the expert-defined

***Table 9.4*:** Main centers in chemical dataset

Expert-group	**# of objects**	**Center**	**Group type**
Anaia	17	c20(17/24)	1.1
Beirut_buff	15	c6(10/10), c14(5/5)	2.0
Beirut_red	9	c4(8/8), c19(1/11)	2.1
Chalcis	46	c7(46/54)	1.1
Chersonese	7	c7(7/54)	1.1
Dhiorios	19	c9(16/16), c3(3/3)	2.0
Ephesos_b2	11	c11(11/11)	1.0
GWWII	9	c2(9/9)	1.0
Horum	7	c8(7/16)	1.1
Istanbul_S2	15	c17(15/15)	1.0
Istanbul_S3	10	c13(10/10)	1.0
Lapithos	8	c7(1/54), c20(7/24)	2.2
Milet	39	c10(19/19), c15(20/20)	2.0
NSW	10	c16(10/21)	1.1
Paphos	11	c16(11/21)	1.1
Plaka_carrot	9	c18(9/9)	1.0
Plaka_LRA1	8	c1(3/3), c8(5/16)	2.1
PSSW	4	c8(4/16)	1.1
Thebes	17	c5(17/17)	1.0
workshopX	30	c12(20/20), c19(10/11)	2.1
Total	301		

***Table 9.5*:** Types of expert-defined groups

# of centers	**# of non-exclusive centers**	**Description**
1	**0**	One main center. This center is exclusive
1	**1**	One main center. This center is non-exclusive
2	**0**	Two main centers. These two centers are exclusive
2	**1**	Two main centers. One of these two centers is non-exclusive
2	**2**	Two main centers. These two centers are non-exclusive
3	**0**	Three main centers. These three centers are exclusive
3	**1**	Three main centers. One of these three centers is non-exclusive
3	**2**	Three main centers. Two of these three centers are non-exclusive
3	**3**	Three main centers. These three centers are non-exclusive

group of an object from the center, which is heterogeneous. For example, let us consider expert-defined groups Paphos and NSW. All 11 objects of Paphos and all 10 objects of NSW are connected with the same main center c16. Although it is

not possible to identify the group by knowing the center, since all the objects of Paphos and NSW lie in the same center, we expect that these two expert-defined groups are close from a chemical point of view. Their association is well explained by the structure of the dendrogram resulting from HCA, where these two expert-defined groups are consecutive (Figure ***9.2***). The same case appears for the expert-defined groups Chalcis and Chersonese: all the objects of Chalcis and Chersonese are related to the same main center, c7.

Group type 2.1 includes the Beirut_red, Plaka_LRA1 and workshopX expert-defined groups. One of the centers is clearly related to one expert-defined groups (c6 Beirut_red and c1 for workshopX), while the other center is more ambiguously connected. Among these three groups, with the exception of only one sample, Beirut_red can be assimilated to group type 1.0 and workshopX to group type 2.0. Only Plaka_LRA1 really corresponds to a group of type 2.1, since Plaka_LRA1 is split into two different centers. Three objects of Plaka_LRA1 are sufficiently different from five other objects of Plaka_LRA1, which are close to the expert-defined group Horum.

There is only one group of type 2.2, Lapithos. This group is also more difficult to analyze, i.e., one out of eight Lapithos objects is marginal, while the remaining objects lie in the same center (c20), where they are associated with the expert-defined group Anaia (Figure ***9.2***). Lapithos and Anaia are actually close to each other from the chemical point of view.

In conclusion, we note that it is not so interesting to take the second and third centers into account. If we add the second main center, analyzing the result is more complicated, while these additional centers do not bring enough information to enhance the outcome. Thus, we consider only the first main center. However, this is presumably the particular case of this dataset. Other real-life datasets may not behave the same.

Secondly, it seems that some of the chemical clusters obtained (by considering only the largest membership value) are too small in size to be reliable (Table ***9.3***). Thus, there is a need for more information to achieve results better connected to expert-defined groups.

9.3 Description Data

9.3.1 Fuzzy C-Means Parameters

Figure ***9.1*** presents the steps involved in description data analysis, which are detailed below.

Data Preprocessing Description data are in Boolean format, which induces a sparse high-dimensional matrix. Thus, we reduce the matrix using the MCFA method implemented in the XLSTAT data analysis application (version 2017.4)[1]. MCFA finds

[1] `https://www.xlstat.com/en/`

the axes with the greatest projected variance. In our case, description data give 11 axes (Table ***9.6***) that represent about 75% of the variance.

***Table 9.6*:** Description data reduction results

Results	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**
Eigenvalue	0.112	0.101	0.076	0.076	0.061	0.054	0.052	0.047	0.044	0.044	0.037
Inertia (%)	11.183	10.067	7.621	7.561	6.125	5.379	5.218	4.680	4.439	4.370	3.712
Cumulative %	11.183	21.250	28.871	36.432	42.557	47.936	53.154	57.834	62.273	66.642	70.354
Adjusted inertia	0.009	0.007	0.004	0.004	0.002	0.001	0.001	0.001	0.001	0.001	0.000
Adjusted inertia (%)	22.51	17.51	8.75	8.57	4.87	3.36	3.07	2.20	1.85	1.76	0.99
Cumulative %	22.51	40.02	48.77	57.34	62.21	65.57	68.64	70.84	72.69	74.45	**75.45**

Determination of the Optimal Number of Clusters Before FCM is applied to reduced description data, we find the optimal number of clusters with Visual $TSFD$. The resulting plot (Figure ***9.4***) shows that the best K value for description data is clearly 12.

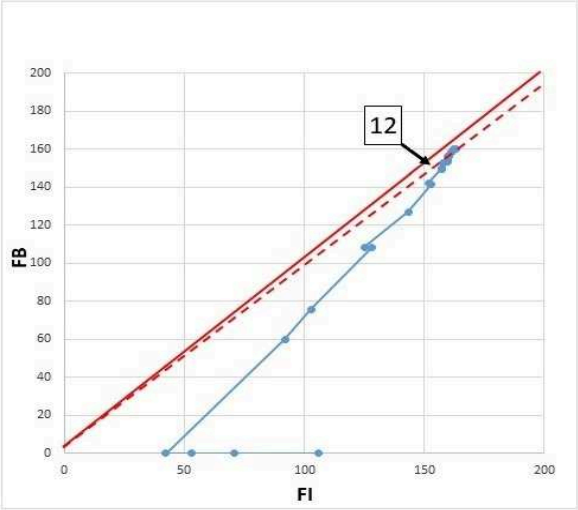

***Figure 9.4*:** Cluster number determination (description data)

9.3.2 Fuzzy C-Means Results

We apply FCM to the reduced description data with $K = 12$ to obtain the fuzzy coefficient matrix. Then, we identify for each object its first three greatest fuzzy

coefficients and sort the objects according to the identifiers of the three closest centers (Table ***9.7***) as we did for chemical data.

Table 9.7: Centers of order 1–3 in description dataset

Cnrt 1.	**Cnrt 2**	**Cnrt 3**	**Expert-groups**	**# of objects**	**Cnrt 1**	**Cnrt 2**	**Cnrt 2**	**Expert-groups**	**# of objects**
c1	c4	c12	Chersonese	3	c7	c11	c12	Thebes	2
c1	c12	c11	Chersonese	4	c7	c12	c9	Chalcis	3
c2	c11	c12	Ephesos_b2	2	c7	c12	c11	Chalcis	6
c2	c12	c11	Ephesos_b2	5	c8	c6	c12	Dhiorios	19
c3	c10	c11	Horum	1	c9	c6	c7	Beirut_buff	5
c3	c11	c10	Horum	4	c9	c6	c7	Beirut_red	5
c3	c11	c12	Horum	1	c9	c6	c11	Beirut_buff	10
c3	c12	c10	Horum	1	c9	c6	c11	Beirut_red	4
c4	c12	c11	Plaka_carrot	9	c10	c11	c7	Istanbul_S2	8
c4	c12	c11	Plaka_LRA1	8	c10	c11	c12	GWWII	8
c5	c8	c11	Lapithos	4	c10	c11	c12	Istanbul_S2	4
c5	c11	c7	Paphos	3	c10	c11	c12	Istanbul_S3	10
c5	c11	c7	Lapithos	3	c10	c11	c12	NSW	2
c5	c11	c8	Paphos	7	c10	c12	c11	GWWII	1
c5	c11	c8	Lapithos	1	c10	c12	c11	Istanbul_S2	3
c5	c11	c12	Paphos	1	c11	c7	c6	NSW	8
c6	c9	c11	workshopX	24	c11	c9	c6	PSSW	3
c6	c11	c10	PSSW	1	c11	c10	c12	Anaia	4
c6	c11	c12	workshopX	6	c11	c12	c10	Anaia	13
c7	c11	c9	Chalcis	7	c11	c12	c10	Milet	14
c7	c11	c10	Chalcis	21	c12	c7	c11	Chalcis	6
c7	c11	c10	Thebes	15	c12	c11	c10	Ephesos_b2	4
c7	c11	c12	Chalcis	3	c12	c11	c10	Milet	25
					Total				301

Thus, we also construct a new table by considering only the first main center (order one) associated with each object (see Table ***9.8***). The analysis of this table shows how each main center is diverse in terms of experts-defined groups (between 1 and 4 groups for each center).

Lastly, we present the different main fuzzy clustering centers which are associated with each experts-groups and its frequency for description data (Table ***9.9***). This table illustrates the diversity of main centers associated with each group (1 or 2 centers for each group).

9.3.3 Comparison with Expert-Defined Groups

Our first remark is that the number of clusters for description data ($K = 12$) is different from that of expert-defined groups, which is 20. However, there is a strong link between these expert-defined groups (Table ***9.5***) and our clustering results.

For instance, Chersonese, Dhiorios and Horum expert-defined groups correspond to type 1.0 (Table ***9.9***). As for chemical data, if we know the main clustering center of any object (c1, c3 and c8), we can use this information to find the expert-defined groups and *vice versa*.

***Table 9.8*:** Centers of order 1 in description dataset

Center	Expert-group	Frequency	
c1	Chersonese	7	7
c2	Ephesos_b2	7	7
c3	Horum	7	7
c4	Plaka_carrot	9	
	Plaka_LRA1	8	17
c5	Paphos	11	
	Lapithos	8	19
c6	PSSW	1	
	workshopX	30	31
c7	Chalcis	40	
	Thebes	17	57
c8	Dhiorios	19	19
c9	Beirut_buff	15	
	Beirut_red	9	24
c10	GWWII	9	
	Istanbul_S2	15	
	Istanbul_S3	10	
	NSW	2	36
c11	Anaia	17	
	Milet	14	
	NSW	8	
	PSSW	3	42
c12	Ephesos_b2	4	
	Milet	25	
	Chalcis	6	35
		Total	301

Groups of type 1.1 are Anaia, Beirut_buff, Beirut_red, GWWII, Istanbul_S2, Istanbul_S3, Lapithos, Paphos, Plaka_carrot, Plaka_LRA1, Thebes and workshopX. Here, it is not possible to know to what group an object belongs, as centers are heterogeneous. For example, all objects in the expert-defined groups Beirut_buff and Beirut_red correspond to the same fuzzy center. These two expert-defined groups are very close to each other from the description point of view. However, we can distinguish them through chemical clustering results (Table ***9.4***). Expert-defined groups Istanbul_S2, Istanbul_S3 and GWWII, on one hand, and Plaka_carrot, Plaka_LRA1, Paphos and Lapithos, on the other hand, are similar cases.

In most of these cases, the similarity between clusters is due to similar description data related to geographical features: same provenance (site, town, country) and, in some cases, same origin. Expert-defined groups Istanbul_S2, Istanbul_S3 and GWWII correspond to objects found in the same town (Istanbul) and the first two groups also originate from Istanbul. Objects of the Plaka_carrot and Plaka_LRA1 expert-defined

Table 9.9: Main centers in description dataset

Expert-groups	**# of objects**	**Center**	**Group type**
Anaia	17	c11(17/42)	1.1
Beirut_buff	15	c9(15/24)	1.1
Beirut_red	9	c9(9/24)	1.1
Chalcis	46	c7(40/57), c12(6/35)	2.2
Chersonese	7	c1(7/7)	1.0
Dhiorios	19	c8(19/19)	1.0
Ephesos_b2	11	c2(7/7), c12(4/35)	2.1
GWWII	9	c10(9/36)	1.1
Horum	7	c3(7/7)	1.0
Istanbul_S2	15	c10(15/36)	1.1
Istanbul_S3	10	c10(10/36)	1.1
Lapithos	8	c5(8/19)	1.1
Milet	39	c12(25/35), c11(14/42)	2.2
NSW	10	c11(8/42), c10(2/36)	2.2
Paphos	11	c5(11/19)	1.1
Plaka_carrot	9	c4(9/17)	1.1
Plaka_LRA1	8	c4(8/17)	1.1
PSSW	4	c11(3/42), c6(1/31)	2.2
Thebes	17	c7(17/57)	1.1
workshopX	30	c6(30/31)	1.1
Total	301		

groups have the same provenance (the site of Plaka), but differ in origin. Finally, expert-defined groups Paphos and Lapithos correspond to objects with provenance and origin in the same country, Cyprus, but in different sites.

Among the other expert-defined groups, workshopX is of type 1.1, but can be assimilated to type 1.0 with only one exception. Thebes is a true type 1.1 because all its objects lie in a single center that is shared with expert-defined groups Chalcis. This shows that these two groups are close to each other from the description point of view.

There is only one group of type 2.1, Ephesos_b2. 7 out of 11 Ephesos_b2 objects are associated with the main center c2, which is an exclusive center, while the 4 remaining objects are associated with c12, together with 6 objects of Chalcis and 25 objects of Milet.

Expert-defined groups of type 2.2 are Chalcis, Milet, NSW and PSSW. This type is also difficult to analyze because each group splits into two main centers and each of them is heterogeneous. The most heterogeneous case is Milet's, which splits in two large parts, 25 objects with c12 and 14 objects with c11, these two centers being non-exclusive.

As in the case of chemical data, we note that the second and third centers do not bring much additional information. Thus, we only consider the first main center.

Finally, the proximity between expert-defined groups can vary according to the point of view, i.e., chemical or description data. It is particularly interesting to

examine whether expert-defined groups that cannot be distinguished using chemical data only can be distinguished using description data. For instance, expert-defined groups of type 1.1, Paphos and NSW (associated with chemical center c16), Chalcis and Chersonese (associated with chemical center c7), Horum and PSSW (associated with chemical center c8), Lapithos and Anaia (associated with chemical center c20) can be distinguished with description analysis results.

9.4 Image Data

9.4.1 Fuzzy C-Means Parameters

Figure ***9.1*** presents the steps we follow for image data analysis.

Data Preprocessing As chemical data, image data are in a Boolean format. Thus, we also use MCFA to reduce them. Here, 6 axes (Table ***9.10***) represent around 70% of the variance.

Table 9.10: Image data reduction results

Result	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**
Eigenvalue	0.091	0.071	0.060	0.055	0.041	0.037
Inertia (%)	9.067	7.109	5.982	5.467	4.053	3.731
Cumulative %	9.067	16.176	22.159	27.626	31.679	35.411
Adjusted inertia	0.007	0.004	0.003	0.002	0.001	0.001
Adjusted inertia (%)	27.67	16.06	10.81	8.75	4.22	3.41
Cumulative %	27.67	43.73	54.54	63.30	67.52	**70.93**

Determination of the Optimal Number of Clusters Searching for the optimal number of clusters with Visual $TSFD$ outputs the plot from Figure ***9.5***, which shows that the best K value for image data is 16.

9.4.2 Fuzzy C-Means Results

We apply the FCM algorithm to reduced image data, which helps identify, for each object, its first three greatest fuzzy coefficients, and sort the objects according to the identifiers of the three closest centers (Tables ***9.11*** and ***9.12***). We may notice that the maximum fuzzy coefficient membership value is very small and very similar to the second and third fuzzy coefficient membership values. Thus, interpretation is not easy.

9.4.3 Comparison with Expert-Defined Groups

Clustering results from image data clustering are not correlated with the experts-defined groups. Most groups are split into many main centers. There may be several reasons, such as pictures not taken under the same conditions, colors not calibrated, unsatisfactory image treatment and selected features badly connected with chemical

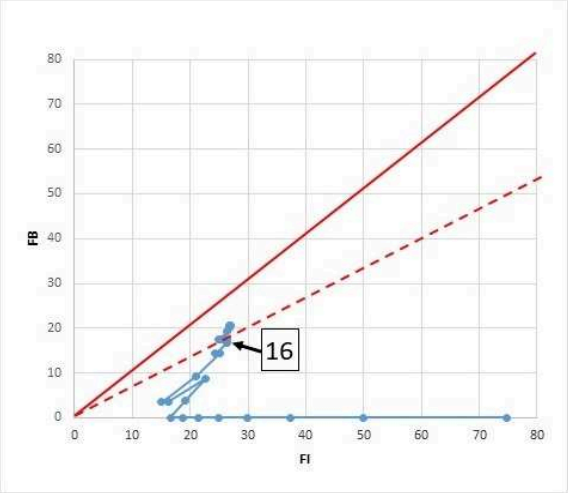

***Figure 9.5*:** Cluster number determination (image data)

and description data. For example, Figure ***9.6*** shows the similarity between fabric images of samples from different groups. It clearly shows that samples belonging to different expert-defined groups may have similar features such as color and size of inclusions. It creates difficulty to separate them by clustering methods.

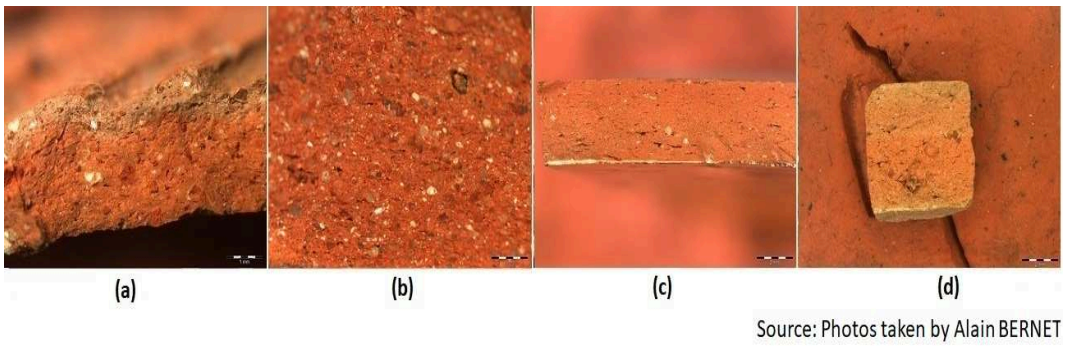

***Figure 9.6*:** Fabric images of samples belonging to different expert-defined groups: LIS87-workshopX (a), LEV81-Beirut_red (b), BZY340-Anaia (c), BZY497-Thebes (d)

***Table 9.11*:** Centers of order 1–3 in image dataset (1/2)

Cntr 1	**Cntr 2**	**Cntr 3**	**Expert-groups**	**# of objects**	**Cntr 1**	**Cntr 2**	**Cntr 3**	**Expert-groups**	**# of objects**
c1	c9	c4	Dhiorios	2	c6	c10	c12	Chalcis	1
c2	c11	c7	Beirut_red	1	c6	c10	c12	NSW	1
c2	c11	c7	Istanbul_S2	1	c6	c10	c13	Chalcis	1
c2	c11	c7	workshopX	4	c6	c10	c13	Istanbul_S2	1
c2	c11	c8	workshopX	3	c6	c10	c13	Milet	1
c3	c6	c10	Chalcis	1	c6	c10	c13	NSW	1
c3	c6	c10	Plaka_carrot	1	c6	c10	c13	Plaka_LRA1	1
c3	c6	c12	Plaka_carrot	1	c6	c10	c13	Thebes	1
c3	c10	c6	Chalcis	2	c6	c10	c14	Chalcis	1
c3	c10	c6	Milet	1	c6	c10	c14	GWWII	1
c3	c10	c6	Thebes	1	c6	c10	c14	Thebes	2
c3	c10	c13	Istanbul_S2	1	c6	c12	c10	Chalcis	1
c3	c10	c15	Istanbul_S2	2	c6	c12	c10	GWWII	1
c3	c12	c13	Anaia	1	c6	c14	c7	Chalcis	1
c3	c13	c15	Dhiorios	1	c6	c14	c10	Chalcis	1
c3	c14	c6	Chalcis	3	c6	c14	c10	GWWII	3
c3	c14	c6	Istanbul_S2	1	c6	c14	c10	Istanbul_S2	1
c3	c14	c7	Istanbul_S3	1	c7	c2	c11	GWWII	1
c3	c15	c5	Dhiorios	1	c7	c6	c12	Chalcis	1
c3	c15	c10	Beirut_buff	1	**c7**	c6	c14	Beirut_red	1
c3	c15	c10	Dhiorios	1	c7	c6	c14	Dhiorios	1
c3	c15	c10	Istanbul_S2	1	c7	c8	c14	Chersonese	1
c3	c15	c10	Istanbul_S3	1	c7	c8	c16	Anaia	1
c3	c15	c10	Thebes	1	c7	c8	c16	Chalcis	1
c3	c15	c12	Dhiorios	1	c7	c8	c16	Chersonese	3
c3	c15	c12	workshopX	1	c7	c8	c16	Dhiorios	1
c4	c9	c1	Plaka_carrot	1	c7	c14	c6	Anaia	3
c5	c8	c7	Dhiorios	1	c7	c14	c6	Chalcis	1
c5	c8	c7	workshopX	1	c7	c14	c16	Chalcis	1
c5	c8	c12	Anaia	1	c7	c14	c16	NSW	1
c5	c8	c12	Dhiorios	1	c7	c14	c16	Thebes	1
c5	c8	c12	Ephesos_b2	2	c7	c16	c8	Paphos	1
c5	c8	c12	Istanbul_S3	3	c8	c12	c6	NSW	1
c5	c8	c12	Milet	1	c8	c12	c13	Paphos	1
c5	c8	c12	Paphos	1	c8	c16	c5	Milet	1
c5	c8	c12	workshopX	3	c8	c16	c7	Anaia	1
c5	c8	c16	Istanbul_S3	1	c8	c16	c7	Ephesos_b2	1
c5	c8	c16	Plaka_LRA1	1	c8	c16	c12	Anaia	2
c5	c12	c7	workshopX	1	c8	c16	c12	Chersonese	1
c5	c12	c8	Ephesos_b2	4	c8	c16	c12	Ephesos_b2	1
c5	c12	c8	Milet	2	c8	c16	c12	Horum	1
c5	c12	c8	NSW	2	c8	c16	c12	Istanbul_S3	2
c5	c12	c8	Plaka_carrot	1	c8	c16	c12	NSW	1
c5	c12	c8	Plaka_LRA1	2	c9	c1	c4	Milet	1
c5	c12	c8	Thebes	1	c9	c1	c4	NSW	1
c5	c12	c8	workshopX	8	c9	c1	c4	Plaka_LRA1	2

9.5 Summary

In this chapter, we cluster a dataset from the ArAr laboratory, which consists of chemical, description and image data characterizing ceramics. We apply FCM (Sec-

***Table 9.12*:** Centers of order 1–3 in image dataset (2/2)

Cntr 1	Cntr 2	Cntr 3	Expert-groups	# of objects	Cntr 1	Cntr 2	Cntr 3	Expert-groups	# of objects
c10	c6	c9	Paphos	1	c13	c16	c8	Anaia	1
c10	c6	c13	Beirut_buff	1	c13	c16	c8	Beirut_buff	1
c10	c6	c13	Chalcis	4	c13	c16	c12	Milet	1
c10	c6	c13	Istanbul_S2	1	c14	c6	c7	Beirut_red	3
c10	c6	c13	Istanbul_S3	1	c14	c6	c7	Chalcis	15
c10	c6	c13	Thebes	5	c14	c6	c7	GWWII	1
c10	c13	c6	Anaia	1	c14	c6	c10	Chalcis	1
c10	c13	c6	Beirut_buff	6	c14	c6	c10	Istanbul_S2	1
c10	c13	c6	Chalcis	3	c14	c7	c6	Beirut_buff	1
c10	c13	c6	GWWII	1	c14	c7	c6	Beirut_red	2
c10	c13	c6	Horum	2	c14	c7	c6	Beirut_red	3
c10	c13	c6	Istanbul_S3	1	c14	c7	c6	Istanbul_S2	1
c10	c13	c6	Milet	6	c15	c3	c7	Chalcis	1
c10	c13	c6	PSSW	1	c15	c3	c8	Chersonese	1
c10	c13	c6	Thebes	2	c15	c3	c8	Dhiorios	1
c10	c13	c12	Dhiorios	1	c15	c3	c8	Istanbul_S2	1
c10	c13	c12	Istanbul_S2	1	c15	c3	c8	Lapithos	4
c10	c13	c16	Beirut_buff	1	c15	c3	c8	PSSW	1
c11	c2	c5	workshopX	1	c15	c3	c13	Dhiorios	1
c11	c2	c6	Istanbul_S2	1	c15	c3	c16	Paphos	1
c11	c2	c7	Plaka_carrot	1	c15	c5	c3	Dhiorios	1
c11	c2	c7	workshopX	3	c15	c8	c3	Dhiorios	3
c11	c2	c12	Plaka_carrot	1	c15	c8	c16	Ephesos_b2	1
c11	c2	c12	workshopX	2	c16	c7	c6	Beirut_buff	1
c11	c2	c14	workshopX	1	c16	c7	c6	Beirut_red	1
c12	c5	c6	Plaka_carrot	1	c16	c7	c6	Horum	1
c12	c5	c6	Plaka_LRA1	1	c16	c7	c6	Thebes	1
c12	c6	c13	Chalcis	1	c16	c7	c8	Milet	1
c12	c6	c13	Plaka_carrot	1	c16	c8	c7	Dhiorios	1
c12	c6	c13	workshopX	2	c16	c8	c7	Lapithos	1
c12	c6	c13	Chersonese	1	c16	c8	c7	PSSW	1
c12	c6	c13	Milet	1	c16	c8	c13	Anaia	3
c12	c13	c6	Chalcis	1	c16	c8	c13	Beirut_buff	1
c12	c13	c6	Paphos	1	c16	c8	c13	Dhiorios	1
c12	c13	c6	Plaka_LRA1	1	c16	c8	c13	Ephesos_b2	1
c12	c13	c8	GWWII	1	c16	c8	c13	Lapithos	2
c12	c13	c8	Milet	1	c16	c8	c13	Milet	8
c12	c13	c8	Plaka_carrot	1	c16	c8	c13	NSW	2
c12	c13	c10	Paphos	1	c16	c8	c13	Paphos	2
c13	c10	c8	Paphos	1	c16	c13	c6	Paphos_Lemba	1
c13	c10	c12	Beirut_buff	1	c16	c13	c6	Anaia	1
c13	c10	c12	Milet	9	c16	c13	c6	Ephesos_b2	1
c13	c10	c16	Anaia	2	c16	c13	c6	Lapithos	1
c13	c10	c16	Milet	1	c16	c13	c10	Beirut_buff	1
c13	c12	c10	Milet	3	c16	c13	c10	Chalcis	1
c13	c12	c10	Thebes	2	c16	c13	c10	Horum	2
c13	c12	c10	Beirut_red	1	c16	c13	c10	Istanbul_S2	1
c13	c12	c10	Horum	1	c16	c13	c10	PSSW	1
c13	c12	c10	Milet	1	**Total**				301

tion 5.1.2.1) to each type of data separately and compare the results with ceramic groups defined by archaeometry experts. The results obtained with chemical and

description data show that the method we propose is feasible and the link of each dataset with the opinions of experts is good, particularly in the case of chemical data, as could be expected.

In our experiments, we notice that the results obtained with either chemical or description data, the second and third centers obtained through clustering do not bring much additional information. The analysis is thus based on the first main fuzzy center, although clustering results would have been different with crisp clustering. Yet, fuzzy clustering provides clusters that are more compatible with the expert-defined groups. However, in chemical data results, some of the clusters have a very small size, while this is not the case with the description data results.

Eventually, an important point is that we compare our method, which is automatic and whose complexity is $O(n)$, to the experts' grouping method, which is partially manual and whose complexity is $O(n^2 \log n)$.

Chapter 10

Multiple Clustering on Archaeological and Archaeometric Data

Contents

In Chapter 9, we separately cluster the different types of data regarding 301 ceramic objects. The results obtained from chemical and description data, although different, are relatively easy to interpret. However, clustering results on image data is much less understandable.

Thus, in this chapter, we aim to combine chemical and description data to obtain a better match with expert-defined groups. Moreover, multiple clustering also simulates

the collaboration between researchers or laboratories using different criteria to cluster objects, e.g., one laboratory clustering chemical data (archaeometry laboratory) and another one clustering descriptive data (archaeology laboratory).

We first discuss several possible strategies to carry out multiple clustering analysis, and then detail the most relevant, i.e., ensemble clustering and combined partition clustering. Finally, we compare multiple clustering results to expert-defined groups.

10.1 Multiple Clustering Strategies

10.1.1 Mixing All Variables

To mix all variables, we can only consider variables that are all of the same type. At first glance, chemical variables are numerical, while description variables are Boolean, but the description data matrix is actually a sparse high-dimensional Boolean matrix that we reduce to a numerical matrix with MCFA. However, it is not appropriate to mix chemical variables with variables corresponding to the first factors of the correspondence analysis on description data.

10.1.2 Mixing All Centers

It would be interesting to mix all fuzzy centers by considering the two fuzzy clustering matrices as a single (whole) data matrix. Then, rows would correspond to objects and columns to variables. In our case, the fuzzy coefficient matrix for chemical variables (Section 9.2.1) corresponds to the first 20 variables ($K = 20$). Likewise, the fuzzy coefficient matrix for description variables (Section 9.3.1) corresponds to the following 12 variables ($K = 12$). Overall, we have 32 variables. Unfortunately, this matrix presentation does not give the same weight to chemical data and description data, which introduces a bias for the distance calculation.

10.1.3 Creating a Committee of Fuzzy Clusterers (Ensemble Approach)

Knowing that each separate analysis gives a fuzzy clustering, we create a committee of fuzzy clusterers. In our case, we have two clusterers, one for chemical data and one for description data. The advantage of this method is twofold. First, we do not have any problem with data type, since each is processed according to its properties. Second, the synthesis of clusterings preserves the fuzzy property.

10.1.4 Combined Partition Clustering

Knowing that the second and third centers do not add much information in the experiments from Chapter 9, an option is to "defuzzify" the partition obtained with each type of data (chemical and descriptive), and then to cross-tabulate the partitions.

In case of hard clustering, each clusterer gives a partition. With only 2 or 3 clusterers, an option is to cross tabulate the obtained partitions. To apply this option to fuzzy-clustering, it is enough to defuzzify each fuzzy clustering result (Section 10.8).

We will apply the last two presented multiple clustering strategies: creating a committee of fuzzy clusterers (Section 10.2 to 10.7) and combined partition clustering (Section 10.8).

10.2 Fuzzy Clustering Ensemble Schemes

We focus in this section on fuzzy clustering ensemble schemes that assemble both numerical and categorical real-life data. Avogadri and Valentini's scheme indicates how to deal with ensemble clustering problems by advising crisp and fuzzy approaches [89]. It can be divided into six steps, for which we review the solutions proposed in the literature.

Step 1 Reduce the Dimension of Data High dimensional data raise various issues: difficult calculations, risk of redundant variables, noisy features and risk of non-pertinent variables. Possible solutions depend on the nature of variables. To reduce the dimension of continuous data, a straightforward solution is to apply PCA (Section 5.2).

To reduce the dimension of Boolean data, feature selection and feature transformation by finding subspaces of data are two suitable solutions. Moreover, to deal with sparse data, Boolean data variables can be reduced with MCFA (Section 5.2).

Step 2: Generation of Multiple Fuzzy Clusterings There are different means of obtaining multiple fuzzy clustering. The first solution is to generate different views on the data by using resampling or random projections, and then applying the same fuzzy clustering algorithm to each view. The second solution is to apply different clustering algorithms on the same dataset, either by varying the parameters of a given algorithm or by using different algorithms. The objects being described by different types of variables, the third solution is to construct a dataset for each type of variables. Then, a fuzzy clustering algorithm is applied to each dataset. Each fuzzy clustering gives a membership coefficient matrix whose rows indicate the list of fuzzy coefficients of each object associated with each center. For instance, FCM can be used to generate multiple fuzzy clusterers [89, 90].

Step 3: Crispization of Base Clusterings Crispization (hardening) is an option before moving to Step 4 [89]. "Defuzzifying techniques", i.e., hard clustering or α-cut, are applied to the fuzzy clusterings obtained in the generation of multiple clusterings. Hard clustering and α-cut are formalized in Equations 10.1 and 10.2, respectively.

$$X_{ik}^{H} = \begin{cases} 1 & \Leftrightarrow argmax_s U_{is} = k \\ 0 & \text{otherwise} \end{cases} \tag{10.1}$$

$$X_{ik}^{\alpha} = \begin{cases} 1 & \Leftrightarrow U_{ik} \geq \alpha \\ 0 & \text{otherwise} \end{cases} \tag{10.2}$$

where U is the fuzzy membership matrix. X_{ik} is a function related to cluster k. For hard clustering, $X_{ik} = 1$ if the i^{th} example belongs to the k^{th} cluster. Otherwise, $X_{ki} = 0$.

For α-cut, $X_{ik} = 1$ if the i^{th} example membership value to the k^{th} cluster is greater than a given threshold α for the k^{th} cluster. Otherwise, $X_{ik} = 0$, $1 \leq k \leq K$, $1 \leq i \leq n$ and $0 \leq \alpha \leq 1$.

Step 4: Aggregation Fuzzy aggregation may be achieved by combining membership coefficients matrices to obtain a $n \times n$ similarity matrix. This matrix is generated by applying different fuzzy t-norms to the membership function of each pair of examples (e.g., minimum [91], algebraic product [92], Lukasewicz's t-norms [93] and drastic product [92]) [89]. In case of crispization, it is possible to build a similarity matrix based on Boolean values using each pair of examples. Then, the algebraic product form of $t - norms$ is used.

Step 5: Clustering in the Embedded Similarity Space Once the global similarity matrix is obtained, fuzzy C-Means may be applied to the rows of this matrix [89, 94].

Step 6: Consensus Clustering Eventually, the consensus clustering is represented by a consensus membership matrix in the fuzzy case [89]. If crispization was used (Step 3), final results are crisp.

10.3 Proposed Ensemble Fuzzy Clustering Method

In Chapter 9, we conduct two fuzzy clustering analyses carried out separately on chemical and description data, respectively. These are our two clusterers, which each output a membership matrix. Each matrix is used to construct pair-wise dissimilarity matrix by using the Manhattan distance $d_1(i, i')$ between the rows of the considered membership matrix U [95]. If points i and i' are considered, the Manhattan distance between i and i', denoted by $d_1(i, i')$, is calculated as given in Equation 10.3.

$$d_1(i, i') = \sum_{j=1}^{K} |u_{ij} - u_{i'j}| \tag{10.3}$$

There are several ways to associate these pairwise matrices. The simplest is to calculate the average pairwise dissimilarity matrix. Considering n observations, t_i, $i = 1, 2, ..., n$, the general formula for the mean of order h is given in Equation 10.4.

$$\bar{t}_h = \frac{1}{n}(\sum_{i=1}^{n} {t_i}^h)^{1/h} \tag{10.4}$$

Particular cases are $h \to -\infty$ (minimum), $h = -1$ (harmonic mean), $h \to 0$ (geometric mean) , $h = 1$ (arithmetic mean), $h = 2$ (quadratic mean) and $h \to +\infty$ (maximum). It is known that t_h is increasing with h from $h-\infty$ (minimum) to $h+\infty$ (maximum).

We prefer to give more weight to the clustering that gives the lowest dissimilarity between two points, we calculate not only the average dissimilarity matrix, but also the harmonic and minimum pairwise dissimilarity matrices (Figure ***10.1***).

Lastly, we construct the fuzzy clustering from the global dissimilarity matrix, and we have to use the Fuzzy K-Medoids method (FKM; Section 5.1.2.2) because we work only with dissimilarity matrix. We can not calculate centroids. Thus, we are using prototypes instead of centroids.

10.4 Average Pairwise Dissimilarity-Based Clustering

10.4.1 Fuzzy K-Medoids Parameters

In this subsection, we precise the construction of the average pairwise dissimilarity matrix and the way to determine the optimal number of clusters.

Average Pairwise Dissimilarity Matrix Construction The global pairwise dissimilarity matrix is calculated as the arithmetic average of both the chemical and the description pairwise dissimilarity matrices.

Determination of the Optimal Number of Clusters The average pairwise dissimilarity matrix only stores distances between objects. Data values being unknown, the calculation of centers is impossible. Thus, the FKM must be used. By using Visual $TSFD$ in conjunction with FKM, we obtain the plot from Figure ***10.2***, which shows that the best K value is 20.

10.4.2 Fuzzy K-Medoids Results

The FKM algorithm is applied to the average pairwise dissimilarity matrix with $K = 20$ in order to obtain the fuzzy coefficient matrix. We use our MaxMin Linear method (Section 7.3) for all initializations.

From the fuzzy coefficients matrix, we can identify for each object the three centers having the largest fuzzy coefficients value and sort the objects according to the identifiers of these centers (Table ***10.1***).

***Table 10.1*:** Centers of order 1–3 in average pairwise dissimilarity matrix

Cntr 1	**Cntr 2**	**Cntr 3**	**Expert-group**	**# of objects**	**Cntr 1**	**Cntr 2**	**Cntr 3**	**Expert-group**	**# of objects**
c1			workshopX	1	c13			GWWII	1
c1	c19	c9	workshopX	4	c13	c10	c5	GWWII	1
c1	c19	c16	workshopX	19	c13	c10	c20	GWWII	3
c2			Chalcis	1	c13	c20	c10	GWWII	2
c2	c12	c9	Thebes	6	c14			Paphos	1
c2	c12	c18	Chalcis	7	c14	c8	c15	Paphos	6
c2	c18	c12	Chalcis	11	c14	c8	c20	Paphos	1
c3			Dhiorios	1	c14	c15	c8	Paphos	3
c3	c8	c15	Dhiorios	5	c15			Chersonese	1
c3	c9	c6	Dhiorios	9	c15	c8	c14	Chersonese	5
c3	c9	c8	Dhiorios	3	c15	c8	c20	Chersonese	1
c4			Beirut_buff	1	c16			Beirut_red	1
c4	c9	c16	Beirut_buff	3	c16	c4	c8	Beirut_buff	1
c4	c16	c9	Beirut_buff	6	c16	c4	c8	Beirut_red	1
c5			Istanbul_S3	1	c16	c4	c14	Beirut_buff	2
c5	c13	c9	Istanbul_S3	5	c16	c8	c14	Beirut_red	2
c5	c13	c20	Istanbul_S3	3	c16	c14	c4	Beirut_buff	1
c6			Plaka_carrot	1	c16	c14	c8	Beirut_buff	1
c6	c9	c8	Plaka_carrot	3	c16	c14	c8	Beirut_red	1
c6	c9	c14	Plaka_carrot	3	c16	c19	c8	Beirut_red	2
c6	c9	c14	Plaka_LRA1	2	c16	c19	c10	Beirut_red	2
c6	c14	c8	Plaka_carrot	2	c17			Milet	1
c6	c14	c8	Plaka_LRA1	4	c17	c7	c11	Milet	12
c6	c14	c9	Plaka_LRA1	2	c17	c10	c7	Milet	2
c7			Milet	1	c17	c10	c11	Anaia	3
c7	c17	c11	Milet	15	c17	c10	c16	PSSW	2
c8			Ephesos_b2	1	c17	c10	c20	Anaia	1
c8	c14	c15	Dhiorios	1	c17	c10	c20	Ephesos_b2	3
c8	c15	c14	Ephesos_b2	6	c17	c11	c7	Milet	4
c9			Horum	1	c17	c11	c10	Anaia	4
c9	c6	c11	Horum	1	c17	c20	c10	Ephesos_b2	1
c9	c8	c11	Horum	1	c18			Chalcis	1
c9	c8	c14	Horum	2	c18	c2	c12	Chalcis	4
c9	c14	c8	Horum	1	c18	c10	c12	Chalcis	6
c10			NSW	1	c18	c10	c17	Chalcis	13
c10	c17	c9	Horum	1	c18	c10	c19	Chalcis	1
c10	c17	c11	Anaia	2	c18	c10	c20	Chalcis	1
c10	c17	c16	PSSW	1	c18	c12	c10	Chalcis	1
c10	c17	c18	NSW	2	c19			workshopX	1
c10	c17	c19	PSSW	1	c19	c1	c16	workshopX	4
c10	c17	c20	Lapithos	3	c19	c16	c10	workshopX	1
c10	c17	c20	NSW	1	c20			Istanbul_S2	1
c10	c18	c17	NSW	3	c20	c5	c13	Istanbul_S3	1
c10	c18	c19	NSW	1	c20	c8	c15	Istanbul_S2	1
c10	c20	c17	Lapithos	5	c20	c10	c17	Istanbul_S2	2
c11			Anaia	1	c20	c13	c5	NSW	2
c11	c7	c17	Milet	1	c20	c13	c8	Istanbul_S2	1
c11	c17	c7	Milet	3	c20	c13	c10	GWWII	2
c11	c17	c8	Anaia	3	c20	c13	c15	Istanbul_S2	1
c11	c17	c10	Anaia	3	c20	c13	c16	Istanbul_S2	1
c12			Thebes	1	c20	c15	c8	Istanbul_S2	4
c12	c2	c9	Thebes	3	c20	c15	c13	Istanbul_S2	1
c12	c2	c10	Thebes	1	c20	c16	c8	Istanbul_S2	1
c12	c2	c15	Thebes	3	c20	c16	c13	Istanbul_S2	1
c12	c2	c18	Thebes	2	c20	c17	c10	Istanbul_S2	1
c12	c10	c18	Thebes	1	Total				301

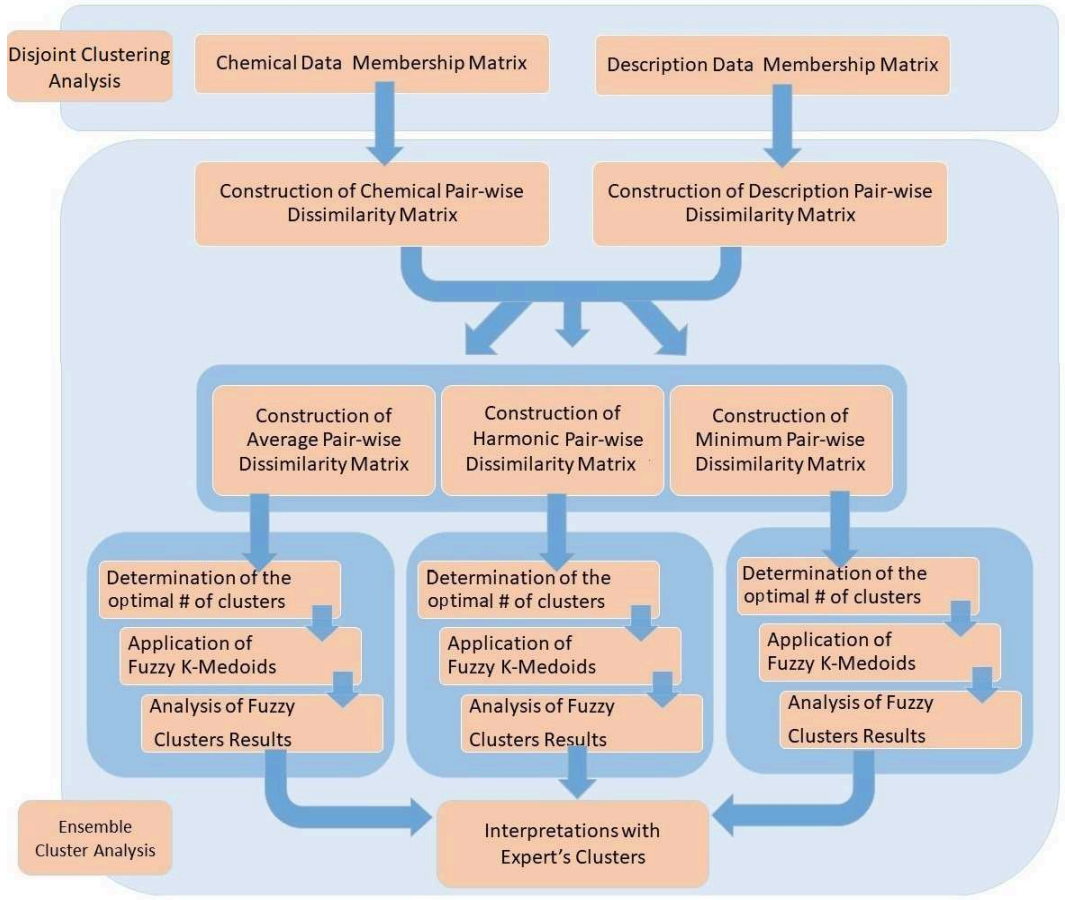

***Figure 10.1*:** Ensemble clustering scheme

The analysis of Table ***10.1*** shows the low diversity of the centers of order 2 and 3 associated with each first center. Thus, we construct a new table by considering only the first main fuzzy cluster center (order 1) associated with each object (Table ***10.2***). It is important to notice that, since we use FKM, centers are prototypes, not centers of gravity. Thus, for each center, there is only one main center and maximum membership coefficient value is 1. Other coefficient values are 0. Hence, for each prototype, only one center has to be considered.

Lastly, we present different main fuzzy clustering centers that are associated with each expert-defined groups, as well as their frequency (Table ***10.3***).

Table 10.2: Centers of order 1 in average pairwise dissimilarity matrix

Center	Expert-group	Frequency	
c1	workshopX	24	24
c2	Chalcis	19	
	Thebes	6	25
c3	Dhiorios	18	18
c4	Beirut_buff	10	10
c5	Istanbul_S3	9	9
c6	Plaka_carrot	9	
	Plaka_LRA1	8	17
c7	Milet	16	16
c8	Ephesos_b2	7	
	Dhiorios	1	8
c9	Horum	6	6
c10	NSW	8	
	Anaia	2	
	PSSW	2	
	Lapithos	8	
	Horum	1	21
c11	Anaia	7	
	Milet	4	11
c12	Thebes	11	11
c13	GWWII	7	7
c14	Paphos	11	11
c15	Chersonese	7	7
c16	Beirut_red	9	
	Beirut_buff	5	14
c17	Milet	19	
	Anaia	8	
	PSSW	2	
	Ephesos_b2	4	33
c18	Chalcis	27	27
c19	workshopX	6	6
c20	Istanbul_S2	15	
	Istanbul_S3	1	
	NSW	2	
	GWWII	2	20
		Total	301

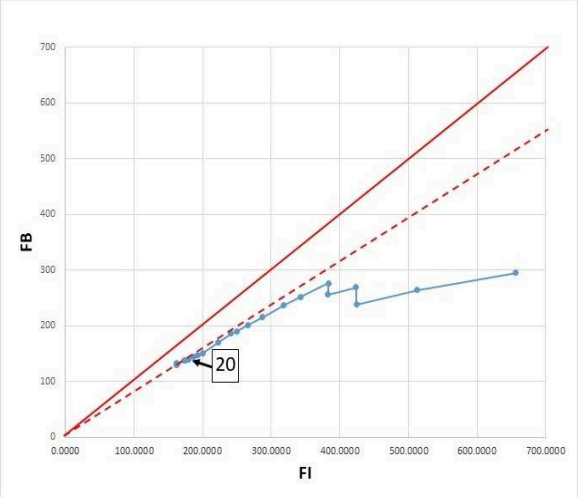

***Figure 10.2*:** Cluster number determination (average pairwise dissimilarity matrix)

10.5 Harmonic Pairwise Dissimilarity-Based Clustering

10.5.1 Fuzzy K-Medoids Parameters

In this subsection, we precise the construction of the harmonic pairwise dissimilarity matrix and the way to determine the optimal number of clusters.

Harmonic Pairwise Dissimilarity Matrix Construction The pairwise dissimilarity matrices are used to calculate the harmonic pairwise dissimilarity matrix as given in the Equation 10.4 with the case $h = 1$.

Determination of the Optimal Number of Clusters As the average, the harmonic pairwise dissimilarity matrix only stores distances between objects. Thus, we use again Visual $TSFD$ combined with FKM to find the optimal number of clusters. The resulting plot (Figure ***10.3***) shows that the best K value for the harmonic pairwise dissimilarity matrix is 20.

10.5.2 Fuzzy K-Medoids Results

The FKM algorithm is applied to the harmonic pairwise dissimilarity matrix with $K = 20$ in order to obtain the fuzzy coefficients matrix. Then, we identify each object's first three largest fuzzy coefficients and sort the objects according to the identifiers of the three centers having the largest fuzzy coefficients (Table ***10.4***).

Next, only the first main center (order 1) is considered for each object and the corresponding table is constructed (Table ***10.5***).

Lastly, we present different main fuzzy clustering centers that are associated with

***Table 10.3*:** Main centers in average pairwise dissimilarity matrix

Expert-group	**# of objects**	**Center**	**Group type**
Anaia	17	c10 (2/21), c11(7/11), c17(8/33)	3.3
Beirut_buff	15	c4(10/10), c16(5/14)	2.1
Beirut_red	9	c16(9/14)	1.1
Chalcis	46	c2(19/25), c18(27/27)	2.1
Chersonese	7	c15(7/7)	1.0
Dhiorios	19	c3(18/18), c8(1/8)	2.1
Ephesos_b2	11	c8(7/8), c17(4/33)	2.2
GWWII	9	c13(7/7), c20(2/20)	2.1
Horum	7	c9(6/6), c10(1/21)	2.1
Istanbul_S2	15	c20(15/20)	1.1
Istanbul_S3	10	c5(9/9), c20(1/20)	2.1
Lapithos	8	c10(8/21)	1.1
Milet	39	c7(16/16), c11(4/11), c17(19/33)	3.2
NSW	10	c10(8/21), c20(2/20)	2.2
Paphos	11	c14(11/11)	1.0
Plaka_carrot	9	c6(9/17),	1.1
Plaka_LRA1	8	c6(8/17)	1.1
PSSW	4	c10(2/21), c17(2/33)	2.2
Thebes	17	c2(6/25), c12(11/11)	2.1
workshopX	30	c1(24/24), c19(6/6)	2.0
Total	301		

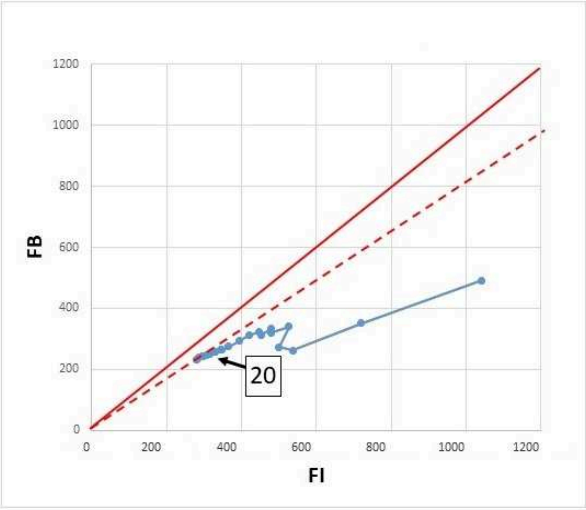

***Figure 10.3*:** Cluster number determination (harmonic pairwise dissimilarity matrix)

each expert-defined group, as well as their frequency for the harmonic pairwise dissimilarity matrix (Table ***10.6***).

***Table 10.4*:** Centers of order 1–3 in harmonic pairwise dissimilarity matrix

Cntr 1	**Cntr 2**	**Cntr 3**	**Expert-group**	**# of objects**
c1			workshopX	1
c1	c20	c2	workshopX	1
c1	c20	c14	workshopX	22
c2			Plaka_LRA1	1
c3			Dhiorios	1
c3	c15	c2	Dhiorios	1
c4			Ephesos_b2	1
c4	c2	c3	Ephesos_b2	4
c4	c16	c14	Ephesos_b2	2
c5			Beirut_buff	1
c5	c14	c2	Beirut_buff	9
c5	c14	c16	Beirut_buff	5
c6			Chalcis	1
c6	c12	c16	Chalcis	3
c6	c12	c18	Chalcis	5
c6	c16	c12	Chalcis	1
c6	c16	c18	Chalcis	6
c6	c18	c16	Chalcis	8
c6	c18	c20	Chalcis	2
c7			Horum	1
c7	c2	c14	Horum	2
c7	c14	c16	Horum	1
c7	c16	c14	Horum	1
c7	c16	c18	Horum	1
c8			Istanbul_S3	1
c8	c13	c14	Istanbul_S3	4
c8	c13	c16	Istanbul_S3	3
c8	c14	c13	Istanbul_S3	2
c9			Paphos	1
c9	c14	c16	Paphos	6
c9	c18	c16	Paphos	3
c10			Chersonese	1
c10	c14	c2	Chersonese	3
c10	c16	c14	Chersonese	2
c11			Milet	1
c11	c19	c16	Milet	26
c11	c19	c18	Milet	7
c12			Thebes	1
c12	c6	c16	Thebes	8
c12	c14	c16	Thebes	5
c12	c16	c18	Thebes	1
c12	c18	c16	Thebes	1
c13			GWWII	1
c13	c8	c16	Istanbul_S2	1
c13	c16	c8	Istanbul_S2	1
c13	c16	c18	Istanbul_S2	1
c13	c16	c18	NSW	2
c13	c16	c20	GWWII	1
c13	c18	c16	Istanbul_S2	9
c13	c18	c20	GWWII	6
c13	c20	c18	GWWII	1
c14			Beirut_red	1
c14	c5	c2	Beirut_red	1
c14	c16	c18	Beirut_red	1
c14	c16	c20	Beirut_red	2
c14	c20	c16	Beirut_red	1
c15			Dhiorios	1
c15	c3	c2	Dhiorios	16
c16			Ephesos_b2	1
c16	c19	c18	Ephesos_b2	3
c17			Plaka_carrot	1
c17	c2	c3	Plaka_carrot	8
c17	c2	c3	Plaka_LRA1	7
c18			NSW	1
c18	c9	c16	Paphos	1
c18	c6	c20	Chalcis	3
c18	c13	c20	Istanbul_S2	1
c18	c16	c6	Chalcis	1
c18	c16	c10	Chersonese	1
c18	c16	c19	Horum	1
c18	c16	c20	Chalcis	3
c18	c16	c20	NSW	1
c18	c19	c20	Anaia	1
c18	c19	c20	Lapithos	3
c18	c20	c6	Chalcis	1
c18	c20	c13	Istanbul_S2	2
c18	c20	c16	Chalcis	7
c18	c20	c16	NSW	3
c18	c20	c16	Thebes	1
c18	c20	c19	Chalcis	5
c18	c20	c19	Lapithos	5
c18	c20	c19	NSW	3
c18	c20	c19	PSSW	4
c19			Anaia	1
c19	c11	c16	Milet	5
c19	c16	c11	Anaia	1
c19	c16	c18	Anaia	3
c19	c18	c16	Anaia	11
c20			workshopX	1
c20	c1	c18	workshopX	4
c20	c14	c18	Beirut_red	1
c20	c18	c13	Beirut_red	1
c20	c18	c13	workshopX	1
c20	c18	c14	Beirut_red	1
Total				301

***Table 10.5*:** Centers of order 1 in harmonic pairwise dissimilarity matrix

Center	Expert-group	Frequency	
c1	workshopX	24	24
c2	Plaka_LRA1	1	1
c3	Dhiorios	2	2
c4	Ephesos_b2	7	7
c5	Beirut_buff	15	15
c6	Chalcis	26	26
c7	Horum	6	6
c8	Istanbul_S3	10	10
c9	Paphos	10	10
c10	Chersonese	6	6
c11	Milet	34	34
c12	Thebes	16	16
c13	GWWII	9	
	Istanbul_S2	12	
	NSW	2	23
c14	Beirut_red	6	6
c15	Dhiorios	17	17
c16	Ephesos_b2	4	4
c17	Plaka_carrot	9	
	Plaka_LRA1	7	16
c18	NSW	8	
	Paphos	1	
	Chalcis	20	
	Istanbul_S2	3	
	Chersonese	1	
	Horum	1	
	Anaia	1	
	Lapithos	8	
	Thebes	1	
	PSSW	4	48
c19	Anaia	16	
	Milet	5	21
c20	workshopX	6	
	Beirut_red	3	9
		Total	301

***Table 10.6*:** Main centers in harmonic pairwise dissimilarity matrix

Expert-group	# of objects	Center	Group type
Anaia	17	c18(1/48), c19(16/21)	2.2
Beirut_buff	15	c5(15/15)	1.0
Beirut_red	9	c14(6/6), c20(3/9)	2.1
Chalcis	46	c6(26/26), c18(20/48)	2.1
Chersonese	7	c10(6/6), c18(1/48)	2.1
Dhiorios	19	c3(2/2), c15(17/17)	2.0
Ephesos_b2	11	c4(7/7), c16(4/4)	2.0
GWWII	9	c13(9/23)	1.1
Horum	7	c7(6/6), c18(1/48)	2.1
Istanbul_S2	15	c13(12/23), c18(3/48)	2.2
Istanbul_S3	10	c8(10/10)	1.0
Lapithos	8	c18(8/48)	1.1
Milet	39	c11(34/34), c19(5/21)	2.1
NSW	10	c13(2/23), c18(8/48)	2.2
Paphos	11	c9(10/10), c18(1/48)	2.1
Plaka_carrot	9	c17(9/16)	1.1
Plaka_LRA1	8	c2(1/1), c17(7/16)	2.1
PSSW	4	c18(4/48)	1.1
Thebes	17	c12(16/16), c18(1/48)	2.1
workshopX	30	c1(24/24), c20(6/9)	2.1
Total	301		

10.6 Minimum Pairwise Dissimilarity-Based Clustering

10.6.1 Fuzzy K-Medoids Parameters

In this subsection, we precise the construction of the minimum pairwise dissimilarity matrix and the way to determine the optimal number of clusters.

Minimum Pairwise Dissimilarity Matrix Construction The pairwise dissimilarity matrices are used to calculate the minimum pairwise dissimilarity matrix as given in the Equation 10.4 with the case $h \to -\infty$.

Determination of the Optimal Number of Clusters Again, Visual $TSFD$ and FKM are combined to find the optimal number of clusters. The resulting plot (Figure ***10.4***) shows that the best K value for the minimum pairwise dissimilarity matrix is 20.

10.6.2 Fuzzy K-Medoids Results

The FKM algorithm is applied to the minimum pairwise dissimilarity matrix with $K = 20$ in order to obtain the fuzzy coefficients matrix. Then, each object's first three largest fuzzy coefficients are identified and are sorted the objects according to the identifiers of the three centers having the largest fuzzy coefficients (Table ***10.7***).

Next, a new table is constructed by considering only the first main fuzzy cluster center (order 1) associated with each object (Table ***10.8***).

***Table 10.7*:** Centers of order 1–3 in minimum pairwise dissimilarity matrix

Cntr 1	**Cntr 2**	**Cntr 3**	**Expert-group**	**# of objects**
c1			Plaka_LRA1	1
c1	c15	c14	Plaka_LRA1	6
c1	c15	c14	Plaka_carrot	9
c2			Dhiorios	1
c2	c15	c14	Dhiorios	15
c3			workshopX	1
c3	c14	c15	workshopX	21
c3	c15	c14	workshopX	2
c3	c19	c18	workshopX	4
c4			Ephesos_b2	1
c4	c15	c14	Ephesos_b2	4
c4	c18	c14	Ephesos_b2	2
c5			Beirut_buff	1
c5	c14	c15	Beirut_buff	9
c5	c14	c18	Beirut_buff	5
c6			Paphos	1
c6	c14	c15	Paphos	6
c6	c18	c14	Paphos	2
c6	c18	c19	Paphos	1
c6	c19	c18	Paphos	1
c7			Horum	1
c7	c14	c18	Horum	2
c7	c15	c14	Horum	2
c7	c18	c19	Horum	1
c8			Chalcis	1
c8	c18	c14	Chalcis	3
c8	c18	c20	Chalcis	10
c8	c20	c18	Chalcis	2
c8	c20	c19	Chalcis	7
c9			Chersonese	1
c9	c14	c15	Chersonese	3
c9	c14	c18	Chersonese	2
c9	c18	c19	Chersonese	1
c10			Anaia	1
c10	c17	c18	Anaia	2
c10	c17	c18	Milet	4
c10	c17	c19	Anaia	3
c10	c17	c20	Anaia	1
c10	c18	c19	Anaia	1
c10	c18	c20	Ephesos_b2	1
c10	c19	c17	PSSW	2
c10	c19	c18	Anaia	1
c10	c19	c18	Horum	1
c10	c19	c20	Anaia	3
c10	c19	c20	PSSW	2
c10	c20	c19	Anaia	5
c11			Thebes	1
c11	c14	c18	Thebes	5
c11	c18	c14	Thebes	7
c11	c18	c8	Thebes	1
c11	c19	c18	Thebes	2
c12			Istanbul_S3	1
c12	c13	c14	Istanbul_S3	6
c12	c13	c16	Istanbul_S3	1
c12	c14	c13	Istanbul_S3	1
c13			NSW	1
c13	c10	c19	Istanbul_S2	2
c13	c12	c16	Istanbul_S3	1
c13	c16	c18	NSW	1
c13	c16	c18	Istanbul_S2	3
c13	c16	c19	Istanbul_S2	4
c13	c19	c16	Istanbul_S2	6
c14			Beirut_red	1
c14	c5	c15	Beirut_red	1
c14	c18	c19	Beirut_red	4
c14	c19	c10	Beirut_red	1
c14	c19	c18	Beirut_red	1
c15			Dhiorios	1
c15	c2	c14	Dhiorios	2
c15	c14	c1	Plaka_LRA1	1
c16			GWWII	1
c16	c13	c10	GWWII	5
c16	c13	c14	GWWII	1
c16	c13	c18	GWWII	1
c16	c13	c19	GWWII	1
c17			Milet	1
c17	c10	c18	Milet	10
c17	c10	c19	Milet	24
c18			Ephesos_b2	1
c18	c10	c19	Ephesos_b2	2
c18	c14	c19	workshopX	1
c19			NSW	1
c19	c10	c11	Thebes	1
c19	c10	c16	Beirut_red	1
c19	c10	c18	NSW	2
c19	c10	c20	NSW	5
c19	c16	c18	workshopX	1
c20			Chalcis	1
c20	c8	c19	Chalcis	6
c20	c10	c19	Chalcis	4
c20	c10	c19	Lapithos	8
c20	c19	c8	Chalcis	1
c20	c19	c10	Chalcis	10
c20	c19	c18	Chalcis	1
Total				301

***Table 10.8*:** Centers of order 1 in minimum pairwise dissimilarity matrix

Center	Expert-group	Frequency	
c1	Plaka_carrot	9	
	Plaka_LRA1	7	16
c2	Dhiorios	16	16
c3	workshopX	28	28
c4	Ephesos_b2	7	7
c5	Beirut_buff	15	15
c6	Paphos	11	11
c7	Horum	6	6
c8	Chalcis	23	23
c9	Chersonese	7	7
c10	Anaia	17	
	Milet	4	
	PSSW	4	
	Horum	1	
	Ephesos_b2	1	27
c11	Thebes	16	16
c12	Istanbul_S3	9	9
c13	Istanbul_S2	15	
	NSW	2	
	Istanbul_S3	1	18
c14	Beirut_red	8	8
c15	Dhiorios	3	
	Plaka_LRA1	1	4
c16	GWWII	9	9
c17	Milet	35	35
c18	Ephesos_b2	3	
	workshopX	1	4
c19	NSW	8	
	Thebes	1	
	Beirut_red	1	
	workshopX	1	11
c20	Chalcis	23	
	Lapithos	8	31
		Total	301

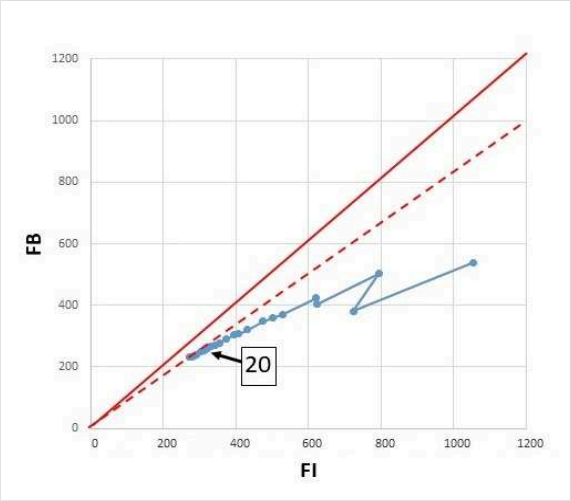

***Figure 10.4*:** Cluster number determination (minimum pairwise dissimilarity matrix)

Lastly, we present different main fuzzy clustering centers that are associated with each expert-defined groups and its frequency for the minimum pairwise dissimilarity matrix (see Table ***10.9***).

10.7 Comparison with Expert-Defined Groups

To construct the pairwise dissimilarity matrix associated with the clustering ensemble, we use different ways to average the two initial pairwise dissimilarity matrices computed from chemical and description data: arithmetic mean, harmonic mean and minimum. While the average calculation gives an equal weight to both initial dissimilarity matrices, the harmonic mean gives an advantage to small values of the two distances we consider. This advantage increases in case of calculations using the minimum value. We use the majority rule to guess the expert-defined group of an object based on its main centers. It appears that the most coherent results with respect to expert-defined groups are achieved with the minimum dissimilarity matrix (Table ***10.10***).

Based on the results achieved with the minimum dissimilarity matrix, if only the first main centers are considered (Table ***10.8***), there are 13 first centers out of 20 (about 2 out of 3), which correspond to clusters including only one expert-defined group. For instance, if the first center is c2, we can be sure that it is expert-defined group Dhiorios, and if the first center is c9, the expert-defined group is Chersonese.

In addition, Table ***9.5*** indicates the types of expert-defined groups (as introduced

Table 10.9: Main centers in minimum pairwise dissimilarity matrix

Expert-group	**# of objects**	**Center**	**Group type**
Anaia	17	c10(17/27)	1.1
Beirut_buff	15	c5(15/15)	1.0
Beirut_red	9	c14(8/8), c19(1/11)	2.1
Chalcis	46	c8(23/23), c20(23/31)	2.1
Chersonese	7	c9(7/7)	1.0
Dhiorios	19	c2(16/16), c15(3/4)	2.1
Ephesos_b2	11	c4(7/7), c18(3/4), c10(1/27)	3.2
GWWII	9	c16(9/9)	1.0
Horum	7	c7(6/6), c10(1/27)	2.1
Istanbul_S2	15	c13(15/18)	1.1
Istanbul_S3	10	c12(9/9), c13(1/18)	2.1
Lapithos	8	c20(8/31)	1.1
Milet	39	c17(35/35), c10(4/27)	2.1
NSW	10	c19(8/11), c13(2/18)	2.2
Paphos	11	c6(11/11)	1.0
Plaka_carrot	9	c1(9/16)	1.1
Plaka_LRA1	8	c1(8/16), c15(1/4)	2.2
PSSW	4	c10(4/27)	1.1
Thebes	17	c11(16/16), c19(1/11)	2.1
workshopX	30	c3(28/28), c18(1/4), c19(1/11)	3.2
Total	301		

Table 10.10: Prediction results for arithmetic, harmonic and minimum pairwise dissimilarity matrix-based clustering

Method	**# of good predictions**	**Rate of good predictions**
Average	245	0.813
Harmonic	247	0.820
Minimum	265	0.880

in Section 9.2.3). Based on the results obtained with the minimum pair-wise dissimilarity matrix, expert-defined groups Beirut_buff, Chersonese, GWWII and Paphos are of type 1.0. With this type of groups, if we know the main clustering center of any object, we can use this information to find the expert-defined group and *vice versa*.

Type 1.1 groups are Anaia, Istanbul_S2, Lapithos, Plaka_carrot and PSSW. Unlike groups of type 1.0, it is not possible to know what group an object belongs to, as centers are heterogeneous, e.g., 17 objects of expert-defined group Anaia and 4 objects of expert-defined group PSSW belong to cluster center c10. Also, all objects of Plaka_carrot and eight out of nine objects of Plaka_LRA1 are related to each other, since both lie in the same cluster center c1.

Type 2.1 groups are Beirut_red, Chalcis, Dhiorios, Horum, Istanbul_S3, Milet and Thebes. One of the two centers is exclusive, whereas the other is heterogeneous. Among these groups, with the exception of only one object, Beirut_red, Horum, Istanbul_S3 and Thebes can be assimilated to groups of type 2.0.

Only expert-defined groups Chalcis, Dhiorios and Milet really are of type 2.1, since each group is split into two different fuzzy centers. Chalcis is evenly split in two, with 23 objects associated with c8, which is exclusive, and 23 objects associated with c20, which also incorporates 8 objects from group Lapithos. For Dhiorios, 16 out of 19 objects are associated with c2, which is exclusive, while the remaining 3 objects are associated with another center. As for Milet, 35 objects out of 39 are associated with center c17, which is exclusive, while the 4 remaining objects are with c10, which gathers all the 17 Anaia objects.

Type 2.2 groups are Plaka_LRA1 and NSW. This type is not easy to analyze. In the case of NSW, 8 objects out of 10 are associated with center c19, and the remaining two objects take place in another center, c13, which gathers all the 15 objects of Istanbul_S2.

Type 3.2 groups are Ephesos_b2 and workshopX. WorkshopX can actually be assimilated to a group of type 1.0, with the exception of 2 objects. Ephesos_b2 is truly a group of type 3.2. Seven objects out of eleven are associated with one center, c4, while four objects are associated with c18, and the remaining objects with c10.

At this point, we need to focus on heterogeneous groups. For instance, clusters c10 and c20 include objects of different expert-defined groups (Table ***10.8***). By considering the membership coefficient of each expert-defined group, Chalcis and Lapithos can be distinguished, e.g., all the membership coefficients of Lapithos are lower than 0.08. At the same time, membership coefficients of Chalcis are often larger than Lapithos'.

Furthermore, we may consider the example of expert-defined groups Plaka_carrot and Plaka_LRA1, which are associated with the same center, c1. Since we know that Plaka_carrot and Plaka_LRA1 are coming from the same location, it is no surprise that these two expert-defined group lie in the same cluster. This similarity, based on the description data, is emphasized by using the minimum calculation. To better distinguish these two groups, we should have selected a larger number of axes in

the factorial analysis. It is indeed only the 15^{th} factorial axis that can distinguish Plaka_carrot and Plaka_LRA1.

In conclusion, the clustering ensemble method gives interesting results. It associates description and chemical information and it outputs more homogeneous cluster sizes than disjoint clusterings. However, clustering ensemble has two drawbacks: (1) the rate of good predictions achieved by clustering ensemble is similar to that of clustering chemical data only; (2) the overall time complexity of the method is $O(n^2)$ due to the generation of the pairwise dissimilarity matrices.

10.8 Combined Partition Clustering

One advantage of fuzzy clustering is that it can be "defuzzified". To fall back to a hard clustering, the first main centers can simply be taken into account, e.g., clusters are considered by using only the maximum coefficient membership values, or alternatively by using the two or three largest coefficients.

In our case, the second and third centers do not bring much information, so that we only take into consideration the maximum coefficient membership values and Table ***10.11*** [89]. The 301 studied objects being described according to their first chemical center and their first description center, Table ***10.11*** gives the number of objects in each cell. In addition, it provides the names of expert-defined groups appearing in each cell. For instance, considering the intersection of chemical row c3 and description column c8, Table ***10.11*** indicates Dhrs:3, which is the number of objects corresponding to this association of chemical and description centers in expert-defined group Dhiorios. At the end of the same line (c3), the total number of objects in chemical group c3 (3) is indicated. Moreover, we reordered rows and columns to keep close one to another different parts of each expert-defined groups. Ultimately, the combined partition distinguishes 33 combined clusters (cells of the Table ***10.11***) which are all pure in terms of groups defined by experts.

The advantage of this cross-tabulation is that chemical and description-based clusterings can be presented in the same table in a synthetic and efficient way. The main outcome is that we never get two expert-defined groups in the same cell. This means that if we know the main chemical and description centers of a new object, then we know the expert-defined group of the object. For example, for an object belonging to chemical center c3 and description center c8, the expert-defined group is Dhiorios. The rate of good predictions of expert-defined groups using the combination of chemical and description fuzzy clustering is 100%. Can this error rate be considered as a generalized error rate? From one side, we do not use the labels of objects during the construction of the combined clustering. Thus, the error rate we obtained thanks to the combined clustering can be considered as a generalized error rate. On the other side, in supervised learning, neither the labels nor the exogenous variables concerning the objects of the test set are used to construct the classifier. To use the same procedure in our case, it would be necessary to distinguish among the objects

***Table 10.11*:** Combined Partition for chemical and description clusters. Experts-groups with their corresponding abbreviations are: Anaia (Ana), Beirut_buff (BeiB), Beirut_red (BeiR), Chalcis (Chls), Chersonese (Chsn), Dhiorios (Dhrs), Ephesos_b2 (Ephs), GWWII (Gww), Horum (Hrm), Istanbul_S2 (IstS2), Istanbul_S3 (IstS3), Lapithos (Lpth), Milet (Mlt), NSW (Nsw), Paphos (Pphs), Plaka_carrot (PlkC), Plaka_LRA1 (PlkL), PSSW (Pssw), Thebes (Thbs), workshopX (Wrks)

Chem./ Descr.	**c8**	**c4**	**c3**	**c5**	**c7**	**c2**	**c11**	**c12**	**c10**	**c6**	**c1**	**c9**	
c3	Dhrs: 3												3
c9	Dhrs: 16												16
c8		PlkL: 5	Hrm: 7				Pssw: 3			Pssw: 1			16
c1		PlkL: 3											3
c18		PlkC: 9											9
c11						Ephs: 7		Ephs: 4					11
c5					Thbs: 17								17
c20				Lpth: 7			Ana: 17						24
c2									Gww: 9				9
c7				Lpth: 1	Chls: 40			Chls: 6			Chsn: 7		54
c10							Mlt: 5	Mlt: 14					19
c15							Mlt: 9	Mlt: 11					20
c13									IstS3: 10				10
c17									IstS2: 15				15
c16				Pphs: 11			Nsw: 8		Nsw: 2				21
c12										Wrks: 20			20
c19										Wrks: 10		BeiR: 1	11
c14												BeiB: 5	5
c6												BeiB: 10	10
c4												BeiR: 8	8
	19	17	7	19	57	7	42	35	36	31	7	24	301

a learning set which allows to construct the combined partition and a test set to evaluate the obtained combined partition.

Conversely, it is interesting to search the combined clusters which correspond to the groups defined by experts. There are 9 groups defined by experts which correspond exclusively to one chemical center and one description center (see Table ***10.12***), i.e, only one combined cluster (one cell of the cross tabulation). It means that for each of these groups, all the objects are at once in same chemical cluster and in the same description cluster.

***Table 10.12*:** Doubly homogeneous groups

Expert-group	**Chem. center**	**Descr. center**	**Frequency**
Anaia	c20	c11	17
Chersonese	c7	c1	7
GWWII	c2	c10	9
Horum	c8	c3	7
Istanbul S2	c17	c10	15
Istanbul S3	c13	c10	10
Paphos	c16	c5	11
Plaka carrot	c18	c4	9
Thebes	c5	c7	17

Considering other groups, 4 of them are homogeneous with respect to chemical data clustering only, while 6 of them are homogeneous with respect to description data clustering only. Groups homogeneous with respect to chemical data clustering only (see Table ***10.13***) are Chalcis, Ephesos_b2, NSW and PSSW. Each of these groups is split into two description clusters.

***Table 10.13*:** Groups homogeneous with respect to chemical data clustering only

Expert-group	**Chem. center**	**Descr. center**	**Frequency**
Chalcis	c7	c7, c12	46
Ephesos_b2	c11	c2, c12	11
NSW	c16	c11, c10	10
PSSW	c8	c11, c6	4

Groups homogeneous with respect to description data clustering only (see Table ***10.14***) are Beirut_buff, Beirut_red, Dhiorios, Lapithos, Plaka_LRA1, WorkshopX. Each of these 6 groups is split into two chemical clusters.

***Table 10.14*:** Groups homogeneous with respect to description data clustering only

Expert-group	**Chem. center**	**Descr. center**	**Frequency**
Beirut_buff	c6, c14	c9	15
Beirut_red	c4, c19	c9	9
Dhiorios	c3, c9	c8	19
Lapithos	c7, c20	c5	8
Plaka LRA1	c1, c8	c4	8
WorkshopX	c12, c19	c6	30

It remains 1 group, Milet, which is doubly heterogeneous (see Table ***10.15***). This group is heterogeneous with respect to both chemical clustering (split into 2 different chemical clusters) and description clustering (split into 2 description clusters), which gives 4 combined clusters for Milet.

***Table 10.15*:** Doubly heterogeneous groups

Expert-group	**Chem. center**	**Descr. center**	**Frequency**
Milet	c10, c15	c11, c12	39

In conclusion, the combined partition method gives very good results, while bearing a complexity of $O(n)$. With our corpus of data, it was sufficient to consider the first main center only, but in other cases it would be possible to consider the second center as well. However, this would have the drawback of increasing the complexity of the combined clustering model.

10.9 Summary

When designing our ensemble clustering method, we propose different ways to calculate the average dissimilarity distance in order to give more or less weight to the low values of the dissimilarity when calculating the average. The best results are obtained with the minimum dissimilarity. Even if it does not improve prediction results compared to clustering chemical data only, it gives a more satisfying clustering in terms of clusters size.

Thus, we propose the construction of a combined partition clustering, which has three advantages: (1) its complexity is linear; (2) it is entirely coherent with expert-defined groups; and (3) the results can be presented in a synthetic way using a cross-tabulation which reports on the homogeneity of each group in terms of associated chemical and description clustering centers. In our case study, each cell in the cross-table corresponds to a single expert-defined group. Thus, this knowledge can be used to determine the expert-defined group a sample belongs to, provided its chemical and description clustering results are known.

Chapter 11

Conclusion and Perspectives

11.1 Conclusion

The first part of this thesis introduces a selection of existing ceramic databases as well as some archaeological and archaeometric data warehouses. Then, we propose a new ceramic database model called Ceramo 3.0. This database stores complex archaeological and archaeometric data related to the ceramics studied in the ArAr laboratory. It is further used to source a data warehouse to allow on-line analytical processing (OLAP).

Thus, this part of the thesis contributes to complex ceramic data modeling, storing, navigation and observation. Moreover, the models we propose can easily be adapted to other application domains, e.g., economics and medicine, which share similar data modeling and analysis problems. These contributions were published in the proceedings of the *9^{th} International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT 2015)* [96].

In the second part of the thesis, which focuses on clustering complex archaeological data, we first survey clustering methods, including data mining methods applied to archaeological and archaeometric studies. We evaluate several elements that help achieve good clustering, i.e., clustering initialization and the discovery of an optimal number of clusters. Next, we propose a fuzzy approach that opens new discussions with experts in archaeology and archaeometry. We start by using image segmentation and color detection methods to capture the color, size and frequency of inclusions from fabric images, in order to enrich cluster analysis.

Then, we propose a new linear initialization method called MaxMin Linear, which increases the performance of fuzzy clustering and outperforms state-of-the-art initialization techniques. This contribution was published in the proceedings of the *14^{th} International Conference on Machine Learning and Data Mining (MLDM 2018)* [97].

We also propose a new visual quality index called Visual $TSFD$ that provides the optimal number of clusters, still for enhancing clustering quality. We compare Visual $TSFD$ with existing quality indices on several real-life and artificial datasets. Considering experiment results, whatever the type of data, Visual $TSFD$ outperforms all other indices. This contribution was published in the proceedings of the *14^{th} International Conference on Artificial Intelligence Applications and Innovations (AIAI 2018)* [98].

Further, we perform two kinds of experiments, first by clustering the different types of data from Ceramo 3.0 separately, and then by performing multiple clustering. We compare these new analyses to the hard clustering combined with human expertise that is currently used to construct expert-defined groups of ceramics in the ArAr laboratory. These groups are composed of ceramic samples that have similar characteristics, which we consider as references (ground truth) in our experiments.

In the disjoint clustering experiment, we apply fuzzy clustering onto chemical, description and image data separately, to examine the coherence of our results with respect to expert-defined groups. The results we obtained, especially with chemical and description data, show that our method is feasible and that the link of each dataset with the opinions of experts is good.

Then, we introduce two multiple clustering strategies, i.e, a fuzzy clustering ensemble and combined partition clustering. We propose a novel fuzzy clustering ensemble scheme that calculates different ways of finding the average distance between ceramic objects with quadratic complexity. Even though this scheme does not enhance prediction results with respect to chemical data clustering, it outputs more homogeneous cluster sizes, as in expert-defined groups. Yet, it is computationally costly.

In contrast, combined partition clustering has a linear complexity. It also presents results in a synthetic way in a cross-tabulation table that helps determine the expert-defined group a sample belongs to.

11.2 Perspectives

Although there is already a long tradition in the fields of archaeology and archaeometry of developing IT and statistical tools, this thesis was challenging in its interdisciplinary character. In this work, we somehow simulate processes involved in interdisciplinary research, when crossing viewpoints on the same objects or categories of objects characterized and defined according to different disciplines. We also dealt with data having a heterogeneous character - numerical, text, images. Improvements could certainly be obtained in the way the last two categories were dealt with, especially for images data. For example, commercial software programs such as Olympus Stream Essentials [1] could be used for fabric images.

Still, the point is that the methodology we developed in this thesis can potentially be applied to a large variety of heterogeneous data. This perspective is important in a context of growing availability of various kind of data, especially through the Internet.

Our work however firstly stresses the importance of working on well balanced corpuses, in order to obtain larger clusters of more even size. We would also need to better evaluate the performance our method of combined partitioning when generalizing it. To do so, we could organize a k-fold cross-validation. For example, for a 2-fold

[1] `https://www.olympus-ims.com/en/microscope/stream2/`

cross-validation, we first need to randomly split the dataset into two equal-sized sets. The first set is used as a training set, and the second to validate the first set. Then, the dataset is used by training on the second set and validating on the first. Finally, the generalized error rate is calculated by averaging the two error rates obtained. We can use the error rate of all datasets to evaluate the method. The value of k could also be 3 or 5 depending on the decision of experts.

To apply cross-validation, we need to be able to insert a new object. In the case of hard clustering, the basic procedure is to insert a new object in the cluster whose center is the nearest to the object considered. It would be interesting to propose a new insertion procedure, which would be well suited to the case of fuzzy clustering.

Moreover, we need to improve the method of choosing the main centers associated with a given object, by taking into account not only the order, but also the value of each fuzzy coefficient.

Finally, when navigating data with OLAP, we analyze data with classical, numerical aggregation functions such as sum, average, maximum, etc. It would be interesting to take textual data into account as well, as there are challenges ahead to efficiently aggregate textual data. An interesting lead might be to require the help of human experts [99].

Appendices

Appendix A

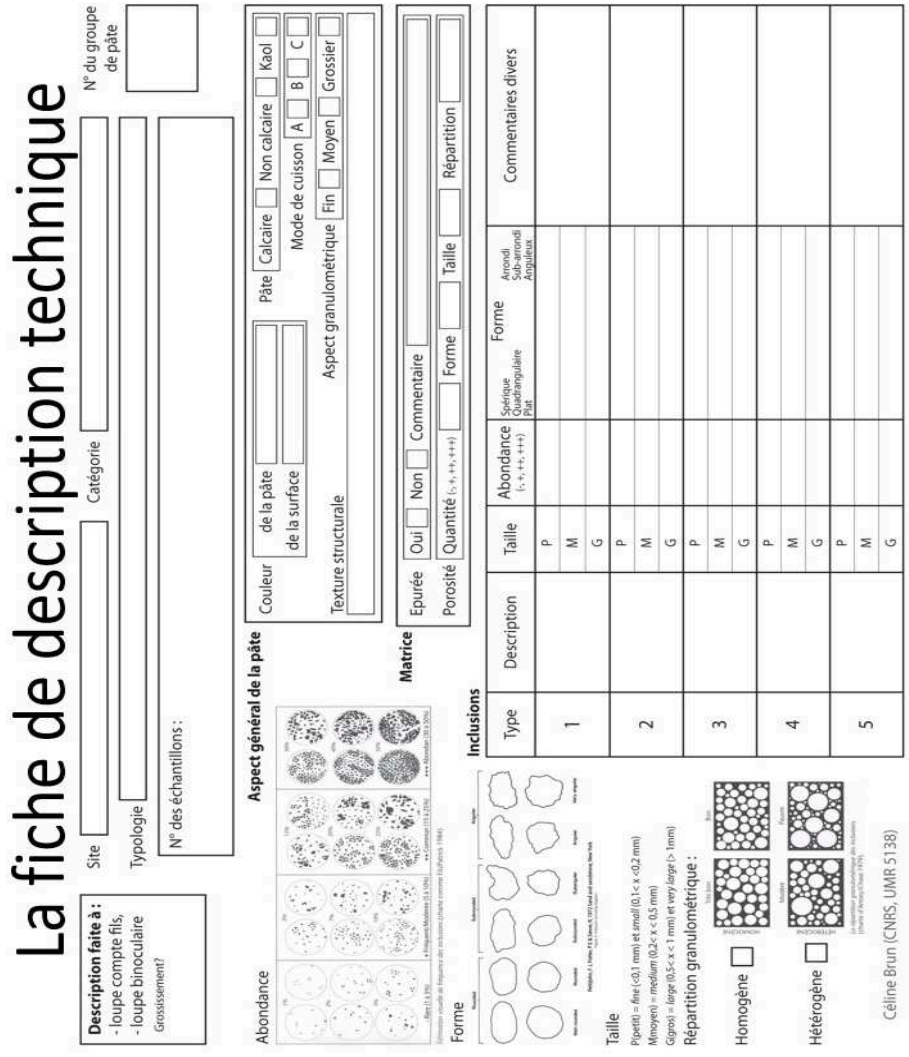

***Figure* A.1:** Fabric description sheet used in the ArAr laboratory (C. Brun)

Appendix B
List of Publications

Conference Proceeding

1. Öztürk A., Lallich S., Darmont J., Waksman S. Y., *MaxMin Linear Initialization for Fuzzy C-Means*, 14$^{\text{th}}$ International Conference on Machine Learning and Data Mining (MLDM 2018), Springer: Machine Learning and Data Mining in Pattern Recognition, July 2018 (to appear).

2. Öztürk A., Lallich S., Darmont J., *A Visual Quality Index for Fuzzy C-Means*, In: Iliadis L., Maglogiannis I., Plagianakos V. (eds) Artificial Intelligence Applications and Innovations. AIAI 2018. IFIP Advances in Information and Communication Technology, Springer, Cham, vol 519 p. 546-555, (2018) (DOI: 10.1007/978-3-319-92007-8_46)

3. Öztürk A., Eyango L., Waksman S. Y., Lallich S., Darmont J., *Warehousing Complex Archaeological Objects*, 9$^{\text{th}}$ International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT 2015), Springer: Lecture Notes in Artificial Intelligence, vol. 9405 p.226-239 (2015) (DOI: 10.1007/978-3-319-25591-0_17)

Appendix C
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[2].

La base de données du MURR Archaeometry Laboratory[3] construite à l'Université

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

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

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.

Couvrant un large éventail de périodes et de régions, la base de données Ceramo du Laboratoire d'Archéologie et d'Archéométrie (ArAr) de Lyon[4] 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 [5], 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 [11]. 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 [15].

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

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 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) [96].

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 [22], 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) [37] 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) [39], 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 [10] [3].

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[5] 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 [62], qui repose sur un clustering obtenu à l'aide des K-Means.

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

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 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[6], 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

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

utilisées pour les versions floues de ces algorithmes.

La méthode la plus simple est celle proposée par MacQueen [28], qui propose d'utiliser les K premiers points de données comme centroïdes. Mais une telle procédure 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) [72] 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 [73]. 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 [74]. 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) [97].

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 [80]. 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 [76]. 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 In-

novations (AIAI 2018) [98].

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] Clive Orton and Mike Hughes. *Pottery in archaeology*. Cambridge University Press, 2013.

[2] Maurice Picon. L'analyse chimique des céramiques: bilan et perspectives. In *Archeometria della Ceramica. Problemi di Metodo, Atti 8 Simposio Internazionale della Ceramica (Rimini 1992)*, pages 3–26, 1993.

[3] 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.

[4] UMR 5138. CNRS - Université de Lyon, Ceramo, forthcoming at http://www.arar.mom.fr/ceramomdatabase/, 2018.

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

[6] Roberta Tomber and John Dore. *The national Roman fabric reference collection: a handbook*. Museum of London, Archaeology Service, 1998.

[7] Patrick Quinn, Dominic Rout, Luke Stringer, Timothy Alexander, Alasdair Armstrong, and Sam Olmstead. Petrodatabase: an on-line database for thin section ceramic petrography. *Journal of Archaeological Science*, 38(9):2491–2496, 2011.

[8] A. Hein and V. Kilikoglou. ceradat—prototype of a web-based relational database for archaeological ceramics. *Archaeometry*, 54(2):230–243, 2012.

[9] Michael D. Glascock, Robert J. Speakman, and Hector Neff. Archaeometry at the University of Missouri research reactor and the provenance of obsidian artefacts in North America. *Archaeometry*, 49(2):343–357, 2007.

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

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

[12] S. Musco, A. Salvatori, M. Mazzei, and C. D'Agostini. Lapis Pallens: Integrated Research on Ancient Roman Quarries of red tuff of Aniene river known as Latomie di Salone (Rome). *5th International Congress on Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin*, (49), 2011.

[13] Eric Paquet and H.L. Viktor. Long-term preservation of 3-d cultural heritage data related to architectural sites. *the ISPRS Working Group*, 4, 2005.

[14] Alain Casali, Sébastien Nedjar, Rosine Cicchetti, Lotfi Lakhal, and Noël Novelli. Lossless reduction of datacubes using partitions. *International Journal of Data Warehousing and Mining*, 5(1):18, 2009.

[15] 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.

[16] Pedro Nogueira, Patricia Moita, Rui Boaventura, Jorge Pedro, Jaime Maximo, Luis Almeida, Susana Machado, Rui Mataloto, André Pereira, S Ribeiro, et al. A spatial data warehouse to predict megaliths slabs sources: mixing geochemistry, petrology, cartography and archaeology for spatial analysis. pages 310–313, 2015.

[17] Jérôme Darmont, Omar Boussaid, Jean-Christian Ralaivao, and Kamel Aouiche. An architecture framework for complex data warehouses. In *7th International Conference on Enterprise Information Systems (ICEIS 2005), Miami, USA*, pages 370–373. INSTICC, May 2005.

[18] Maurice Picon. *Introduction à l'étude technique des céramiques sigillées de Lezoux*, volume 2. Centre de Recherches sur les techniques Gréco-romaines, 1973.

[19] Maurice Picon. Les modes de cuisson, les pâtes et les vernis de la Graufesenque: une mise au point. In M., GENIN, A., VERNHET (dir.). *Céramiques de la Graufesenque et autres productions d'époque romaine. Nouvelles recherches. Hommages à Bettina Hoffmann. Eds. M. Mergoil, Montagnac (Archéologie et histoire romaine 7)*, pages 139–163, 2002.

[20] Yona Waksman and Véronique François. Vers une redéfinition typologique et analytique des céramiques du type "Zeuxippus Ware". *Bulletin de correspondance hellénique*, 128(2.1):629–724, 2004.

[21] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

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

[23] Richard O. Duda, Peter E. Hart, David G. Stork, et al. *Pattern classification*, volume 2. Wiley New York, 1973.

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

[25] Erich Schikuta. Grid-clustering: An efficient hierarchical clustering method for very large data sets. In *Pattern Recognition, the 13th International Conference On*, volume 2, pages 101–105. IEEE, 1996.

[26] Wei Wang, Jiong Yang, Richard Muntz, et al. Sting: A statistical information grid approach to spatial data mining. In *VLDB*, volume 97, pages 186–195, 1997.

[27] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice Hall, Inc., 1988.

[28] 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.

[29] Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program PAM). *Finding groups in data: an introduction to cluster analysis*, pages 68–125, 1990.

[30] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In *the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD)*, pages 21–34. Singapore, 1997.

[31] James C. Bezdek, Robert Ehrlich, and William Full. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203, 1984.

[32] Raghu Krishnapuram, Anupam Joshi, and Liyu Yi. A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In *Fuzzy Systems Conference Proceedings, 1999 IEEE International*, volume 3, pages 1281–1286, 1999.

[33] Joseph C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Taylor & Francis*, 3(3):32–57, 1973.

[34] Linda Ellis. *Archaeological method and theory: an encyclopedia*. Routledge, 2003.

[35] Charlotte Hug, Rebecca Deneckere, and Aymen Ammar. Intentional process modeling of statistical analysis methods. *the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology CAA 2014 - 21st Century Archaeology/F. Giligny, F. Djindjian, L. Costa, P. Moscati, S. Robert (eds.)*, pages 481–488, 2014.

[36] Clive Orton. *Mathematics in archaeology*. Cambridge University Press Cambridge, 1982.

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

[38] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

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

[40] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[41] Giligny François. Intrasite spatial analysis applied to the Neolithic sites of the Paris Basin: from the archaeological feature to global analysis. *The 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology CAA 2014 - 21st Century Archaeology/F. Giligny, F. Djindjian, L. Costa, P. Moscati, S. Robert (eds.)*, pages 497–507, 2014.

[42] David J. Pasta. Learning when to be discrete: continuous vs. categorical predictors. In *SAS Global Forum*, volume 248, pages 1–10, 2009.

[43] Ming-Yi Shih, Jar-Wen Jheng, and Lien-Fu Lai. A two-step method for clustering mixed categroical and numeric data. *Tamkang Journal of Science and Engineering*, 13(1):11–19, 2010.

[44] Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4):393–423, 2002.

[45] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings 1995*, pages 194–202. Elsevier, 1995.

[46] Zengyou He, Xiaofei Xu, and Shengchun Deng. Clustering mixed numeric and categorical data: A cluster ensemble approach. *arXiv preprint cs/0509011*, 2005.

[47] Charu C. Aggarwal. *Data mining: the textbook*. Springer, 2015.

[48] Michal R. Chmielewski and Jerzy W. Grzymala-Busse. Global discretization of continuous attributes as preprocessing for machine learning. *International Journal of Approximate Reasoning*, 15(4):319–331, 1996.

[49] Glenn W. Milligan and Martha C. Cooper. A study of standardization of variables in cluster analysis. *Journal of Classification*, 5(2):181–204, 1988.

[50] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[51] Alexander Topchy, Anil K. Jain, and William Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, 2005.

[52] Martin Kampel and Robert Sablatnig. Color calibration for pre-classification of pottery. In *the Czech Pattern Recognition Workshop*, pages 185–189, 2000.

[53] Martin Kampel, Robert Sablatnig, and Emanuele Costa. *Classification of archaeological fragments using profile primitives*. In Computer Vision, Computer Graphics and Photogrammetry – a Common Viewpoint, the 25th Workshop of the Austrian Association for Pattern Recognition, 2001.

[54] Kristina Adler, Martin Kampel, Raimund Kastler, Martin Penz, Robert Sablatnig, Katerina Schindler, and Srdan Tosovic. Computer aided classification of ceramics-achievements and problems. In *The 6th International Workshop on Archaeology and Computers*, pages 3–12, 2001.

[55] Anna O. Shepard. *Ceramics for the Archaeologist*. Carnegie Institution of Washington Washington, DC, 1956.

[56] Nikhil R. Pal and Sankar K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[57] Hugo Hedberg. A survey of various image segmentation techniques. *Dept. of Electroscience*, 118, 2010.

[58] Muhammad Waseem Khan. A survey: Image segmentation techniques. *International Journal of Future Computer and Communication*, 3(2):89, 2014.

[59] Nida M. Zaitoun and Musbah J. Aqel. Survey on image segmentation techniques. *Procedia Computer Science*, 65:797–806, 2015.

[60] Irwin Sobel and Gary Feldman. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, pages 271–272, 1968.

[61] Heng-Da Cheng, X. H. Jiang, Ying Sun, and Jing Li Wang. Color image segmentation: advances and prospects. *Pattern recognition*, 34(12):2259–2281, 2001.

[62] 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.

[63] Douglas Steinley and Michael J. Brusco. Initializing k-means batch clustering: A critical evaluation of several techniques. *Journal of Classification*, 24(1):99–121, 2007.

[64] Ranjan Maitra, Anna D. Peterson, and Arka P. Ghosh. A systematic evaluation of different methods for initializing the k-means clustering algorithm. *IEEE Transactions on Knowkedge and Data Engineering*, pages 522—-537, 2011.

[65] M. Emre Celebi, Hassan A. Kingravi, and Patricio A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, 2013.

[66] Marija J. Norušis. *IBM SPSS Statistics 19 statistical procedures companion*. Prentice Hall, 2012.

[67] Vance Faber. Clustering and the continuous k-means algorithm. *Los Alamos Science*, 22(138–144), 1994.

[68] David J. Hand and Wojtek J. Krzanowski. Optimising k-means clustering results with standard software packages. *Computational Statistics & Data Analysis*, 49(4):969–973, 2005.

[69] Paul S. Bradley and Usama M. Fayyad. Refining initial points for k-means clustering. In *ICML*, pages 91–99, 1998.

[70] Ting Su and Jennifer G. Dy. In search of deterministic methods for initializing k-means and gaussian mixture clustering. *Intelligent Data Analysis*, 11(4):319–338, 2007.

[71] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *The eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

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

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

[74] 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.

[75] Raghu Krishnapuram, Anupam Joshi, Olfa Nasraoui, and Liyu Yi. Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE transactions on Fuzzy Systems*, 9(4):595–607, 2001.

[76] 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.

[77] Amine M. Bensaid, Lawrence O. Hall, James C. Bezdek, Laurence P. Clarke, Martin L. Silbiger, John A. Arrington, and Reed F. Murtagh. Validity-guided (re) clustering with applications to image segmentation. *IEEE Transactions on Fuzzy Systems*, 4(2):112–123, 1996.

[78] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang, Chih-Chen Lin, and Maintainer David Meyer. Package 'e1071'. https://cran.r-project.org/web/packages/e1071/index.html, Version 1.6-8, 2017.

[79] Nikhil R. Pal and James C. Bezdek. Correction to" on cluster validity for the fuzzy c-means model"[correspondence]. *IEEE Transactions on Fuzzy Systems*, 5(1):152–153, 1997.

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

[81] James C. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3):58–73, 1973.

[82] Min-You Chen and Derek A. Linkens. Rule-base self-generation and simplification for data-driven fuzzy models. In *Fuzzy Systems, 2001. The 10th IEEE International Conference on*, volume 1, pages 424–427. IEEE, 2001.

[83] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[84] Yoshiki Fukuyama and M. Sugeno. A new method of choosing the number of clusters for the fuzzy c-mean method. In *5th Fuzzy Syst. Symp., 1989*, pages 247–250, 1989.

[85] Xuanli Lisa Xie and Gerardo Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991.

[86] Nikhil R. Pal and James C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy systems*, 3(3):370–379, 1995.

[87] Raymond B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[88] Evgenia Dimitriadou, Sara Dolničar, and Andreas Weingessel. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1):137–159, 2002.

[89] Roberto Avogadri and Giorgio Valentini. Ensemble clustering with a fuzzy approach. In *Supervised and Unsupervised Ensemble Methods and their Applications, Oleg Okun, Giorgio Valentini (Eds)*, pages 49–69. Springer, 2008.

[90] Witold Pedrycz. Collaborative fuzzy clustering. *Pattern Recognition Letters*, 23(14):1675–1686, 2002.

[91] Kurt Gödel and Anzeiger Akademie der Zum intuitionistischen Aussagenkalkül. Wissenschaften wien, math.-naturwissensch. *Klasse*, 69:65–66, 1932.

[92] Petr Hájek. *Metamathematics of fuzzy logic*, volume 4. Springer Science & Business Media, 1998.

[93] J. Lukasiewicz. On three-valued logic. ruch filozoficzny, 5,(1920), english translation in borkowski, l.(ed.) 1970. jan lukasiewicz: Selected works, 1920.

[94] Mohammad Ahmadzadeh, Zahra Azartash Golestan, Javad Vahidi, and Babak Shirazi. A graph based approach for clustering ensemble of fuzzy partitions. *J. Math. Comput. Sci*, 6:154–165, 2013.

[95] Dibya Jyoti Bora and Anil Kumar Gupta. Effect of different distance measures on the performance of k-means algorithm: an experimental study in matlab. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 5(2):2501–2506, 2014.

[96] 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.

[97] 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.

[98] 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.

[99] Mustapha Bouakkaz, Youcef Ouinten, Sabine Loudcher, and Yulia Strekalova. Textual aggregation approaches in olap context: A survey. *International Journal of Information Management*, 37(6):684–692, 2017.