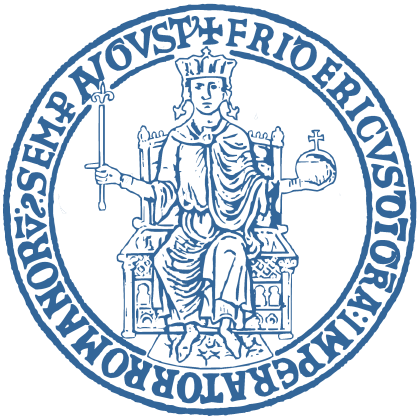
Archetypal analysis for histogram-valued data

An application to Italian school system in a benchmarking perspective

Francesco Santelli



A thesis presented for the PhD in
Social Sciences and Statistics

Department Of Social Sciences
University Federico II of Naples
XXXI Ciclo

Supervisors:
Prof. Francesco Palumbo
Prof. Enrica Morlicchio

Contents

Introduction

The present work set itself in the recent debate about the paradigm shift in statistical learning. Such new paradim is, in the last years, developed mainly from the birth of *Big Data*. Big Data re-framed the way in which scientific research and the constitution of knowledge itself is performed, changing also the nature of the information categorization process (Boyd and Crawford, 2012). According to recent estimates, data scientists claim that the volume of data would roughly double every two years thus reaching the 40 zettabytes (ZB) point by year 2020, considering that a zettabyte is a unit of measure, that amounts to one sextillion of bytes, or equivalently to one trillion of gygabytes, used in digital information system. Including in such estimate also the Internet of things (IoT) impact, it is likely that the data amount will reach 44 ZB by 2020. Big Data is not simply denoted by volume, but also from the variety and the complex nature of such data. The three defining characteristics of Big Data, as a matter of fact, are often assumed to be the three V: *volume* (the data growth and the constant increase of the run rates), *variety* (data come now from various and heterogeneous sources, and assume several natures), and *velocity* (the source speed of data flows is increased to the point that real-time data are available) (Zikopoulos, Eaton, et al., 2011). If taking under consideration only the amount of data, a particularly wide dataset made up by simple data points, can be considered as well as a Big Data case. More interesting, and at the same time more challenging, is the recent and common context in which data are retrieved from sources that are specifically suited for data with a more complex structure. Methodological and technical difficulties are a common problem in using statistical tools when facing increasing complexity in data. To give some examples, data from most used Social Networks (Facebook, Twitter, Instagram and so on) present various types of nature and attributes: associated with each post, there is the text contained in the post itself, the author, the hashtags, the date and the hour related to the post, the geographical location of the author, and even pictures or videos are allowed to be included in the posted content. Conceiving statistical tools able to properly analyze and interpret data so

complex and heterogenous, is a tough task. But the information potentially included in that composite data is often large enough to be advantageous, in terms of statistical learning, to design and to develop techniques adequate for Big, and complex, Data. For these reasons a recurring theme associated to the new paradigm is *Big Data opportunities* (Labrinidis and Jagadish, 2012): more data are constantly available, and these kind of data have not been easily analyzed before, opening a wide space of research and innovation ahead. This leads to propose new methods for the treatment of this type of data. The core of the contribution of this thesis goes to such direction.
Several domains have been deeply transformed by the arrival of Big Data, and some of them, like social network platforms (Facebook, Twitter, Instagram) could be seen as the perfect workshops in which new data structures are developed. Other domains, like Business Analytics (BA) or Information Technology (IT), have obtained great advantage from the wide use of Big Data. For some other fields, on the other hand, the connection with Big Data issues seems to be less obvious and less expected. With respect to such assertion, among more unexpected domains, it comes out as a particularly interesting case the Big Data radical shift in education (D. M. West, 2012). The onset of Big Data has affected both the pupils' learning process itself and the evaluation process of the educational system. For the former, many of the typical traditional tests provide just little immediate feedback to students, whom often fail to take full advantage of digital resources, so the improvement of the learning process is not as good as it could be. For the latter, most common way to evaluate school system is to assess pupils' skills and performances for what concerns several domains; traditional school evaluation can suffer from several limitations if disregarding Big Data opportunities. In the following, the focus will be on this particular aspect. At International level, the Programme for International Students Assessment (OECD-PISA), `https://www.oecd.org/pisa/`, is the organization that analyses, in a comparative fashion, pupil's skills in many Countries. In Italy, the school system evaluation is responsibility of the Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione (INVALSI), `http://www.invalsi.it/invalsi/index.php`. These, and other similar organizations, are able to gather a huge amount of information and data of different nature; most of the time, in a trade-off between complexity and analysis capability, the choice is to analyze data reduced to a simpler form. For example, several tests related to the same domain are reduced to their mean values for more advanced analysis, and the same happens in comparing school performances, losing so the internal variation of the phenomena. It is clear, therefore, that complexity is an opportunity but also a threat. Consistently, most of the statistical learning techniques used in this

context are aimed to reduce such complexity, finding salient units and/or clusters of units that can be properly described.
In the wide set of unsupervised statistical learning techniques, a specific role is played by Archetypal Analysis (Cutler and Breiman, 1994). Archetypes result as very useful salient units especially in account of their properties and location. They are extreme units, belonging to the convex hull of data cloud, defined as a linear combination of data units, meanwhile each data unit can be expressed as a linear combination of the identified archetypes. Therefore, they act as well separated units with extreme/peculiar behavior, suited for benchmarking purpose, as already proposed in (Porzio, Ragozini, and Vistocco, 2006, 2008). This techinque have been used as a statistical tool to achieve a benchmarking analysi,s in a quantitative internal perspective (Kelly, 2004; Smith, 1990), given that the aim, in the school system assessment, is to find excellence standards and worst performances in a public sector. Archetypes have been already proposed and discussed for complex data, especially in Symbolic Data Analysis framework (SDA) (Billard and Diday, 2006). Within this approach, the symbolic data table is an aggregation of simple points into hypercubes (broadly defined). This more complex data-matrix structure leads to define the *intent* as the set of characteristic descriptions, in the symbolic object, that defines a concept. Given the intent, the *extent* is the set of units in the data belonging to the concept according to the description (set of characteristics) and with respect to a rule of association. For this reason, in SDA, the core of the analysis is on the unit of second level (categories, classes or concepts) where units of the first level are aggregated into units of higher level. So, for example, based on given characterisics that are the intent, birds (individual units) can be grouped, using extent association, in species of birds (second level units). This approach allows to retain much more information from original data, allowing for complex unit to present an internal variation and structure. Several ontologies have been proposed for Symbolic Data (Noirhomme-Fraiture and Brito, 2011), and within this classification the focus will be, in this context, on Histogram Symbolic Object, and on the relationship between Histogram-valued data and Interval-valued data. In such SDA perspective, the proposed methodological approach refers to statistical learning techniques. The aim is to describe the aggregation of individual pupils' scores to an higher level, obtaining scores distributions rather than mean values for each school, and then analyze them by means of Histogram Archetypes, under the assumption that the loss of information to simplify data structure for further analysis is not always necessary and worthy in this case. Keeping the natural complexity of data is, from the viewpoint of this work, a potential added value to the interpretative power of the analysis.

A review about archetypal analysis with a focus on archetypes for complex data will be presented, for a proper discussion about the analytical development of the techniques. A particular emphasis will be given to the derivation of prototypes and archetypoids from archetypes. The methodological innovation, and the definition of the archetypes for histogram-valued data, will be presented after this section. Histogram-valued data have been already widely discussed and analysed in literature, especially for what concerns how to measure distance and/or dissimilarity among them. For this reason, a wide review about histogram distances/dissimilarities measures is presented in 2.1.2. Once a distance is defined and choosen, it is possible to develop and then perform clustering procedures for histogram-valued data, and a review about this topic is given in 2.2. Among all the available functions to calculate distance between histograms, particular emphasis will be given to the distances derived from the Wasserstein distance, that uses a function of centers and radii of histogram bins, in a similar way in which interval-valued data are expressed within SDA approach. This allows to exploit the intimate connection, between histogram-valued data and interval-valued data. The new proposed technique will be tested first on a toy example. In the following, the real data application is presented, using histogram archetypes identification as a tool to analyze data retrieved from INVALSI test. The archetypes identified will act as initial intents in the Symbolic Data Analysis approach, and the categorization of school-units in the space spanned by the archetypes will be the way in which extent allocation is performed. This work, seeks, first of all, to accomplish a task of practical nature: to create a space in which is possible to categorize Italian schools according to their reading/writing and mathematics skills using a distribution of pupils' scores rather than mean values. Then, for what concerns the methodological task, the aim is to to develop an extension of archetypal analysis (Cutler and Breiman, 1994) to deal with histogram-valued data as defined in Symbolic Data Analysis, creating so a proper tool to face the former real-data issue.
The work is structured as follows: in the first chapter an overview over statistical learning and its last changes in the Big Data era will be discussed, deepening the new role of educational assessment in recent years given the increasing data complexity. In the second chapter histogram-valued data will be reviewed in a SDA perspective, with particular emphasis on the wide set of available dissimilarity measures and on several unsupervised statistical learning techniques. The role of archetypes is analyzed and discussed in the third chapter, highlighting the usefulness of archetypal analysis is an useful tool in benchmarking evaluation. The extension of archetypal analysis to histogram-valued data is derived in the fourth chapter, presenting results based on a toy example to discuss properties, location and algorithm issues.

Data structure and data building procedure from INVALSI test is presented in the last chapter. Archetypes for histogram-valued data are so identified and then used as benchmarking tool for schools, based on distribution scores. In the Conclusions section, some hints about further developments, in particular for what concerns symbolic data archetypes, are proposed to deal with unanswered questions that this work still has left open.

Chapter 1

Statistical learning: challenges and perspectives in *Big Data* era

1.1 Statistical learning: definition and aim

The development of new statistical tools, as well as the improvement of technical/technological instruments, has brought growing interest in last years. Statistical Learning plays a main role in such scenario. It is at the intersection of statistics with other sciences in a multidisciplinary approach, since concepts, procedures and definitions are coming from several domains, even if converging somehow to similar results. This has lead to a very heterogeneous and diverse theoretical foundations in terms of thorough formalization. However, what all the Statistical Learning topics have in common is, for sure, the well-defined 4 phases (Berk, 2016) that can be assumed as the standard framework in which researchers perform their steps of analysis in order to extract useful information from data by means of statistical tools:

1 Data collection
It includes all the possible procedures in order to retrieve data. It is possible to collect new data with an ad-hoc survey, re-use old data, purchase data from other sources and so on.

2 Data management Often called also "Data wrangling". It consists of a series of actions or steps performed on data to organize, verify, transform, integrate, and extract even new data in an appropriate output format in order to perform the analysis (Singleton et al., 2005).

3 Data analysis it is the main core of the procedure, and aims to extract advantageous patterns from data, in terms of knowledge.

4 Interpretation of results It is related to the explanation, both from an analytic and from a substantial perspective, of the detected patterns,

These four steps are widely formalized in quantitative approaches broadly speaking, even if researchers belong to different fields. Researchers adopting different theoretical background have, likely, used divergent terminology for each part of each phase. Further, certain developments has been proposed by scientists working in industrial framework and business environment. They, in general, use terms that are not the same of the ones used by academics. This issue is an additional element that increases the heterogeneity into research phases definition and standardization. For the purpose of this work the first task is, due to this troublesome and somehow confusing framework, to define what is ***Statistical Learning*** in a quite accurate way and, as a consequence, to decide to what extent statistical techniques will be discussed in next sections and to understand the dynamic role of Statistical Learning in Big Data era. It will be defined mainly for its role in quantitative research and will be compared to the concepts of ***Data Mining*** and ***Machine Learning***.

- **Statistical Learning**
  As pointed out by several authors (e.g. Vapnik, 2000), a great revolution in statistics has happened starting from the 1960s. The Fisher's paradigm, developed in the 1920s, has the focus on the parameters estimation: the researcher has to know the exact number of these parameters to carry out a proper statistical analysis. The analysis about causal - effect relationships between variables can be faced, in this framework, by means of parameters estimation; parametric statistics is the way in which the effect size is calculated, given that predictors-responses relationship is assumed to be known. As a general idea, classic statistics, both frequentist and Bayesian, was considered first of all a branch of mathematics, that evolved as a sub-topic using as main theoretical framework the theory of probability, and as tool various optimization algorithms. Scientific community agrees about the most important proposals that reasonably started an innovation in statistics field:

  i Tikhonov Phillips regularization (Phillips, 1962) or, after the work of Arthur E. Hoerl (Hoerl and Kennard, 1970), ridge regression. It is the most commonly used method of regularization of ill-posed problems.

ii Development of non-parametric statistics methods (Conover, 1999), introduced by Parzen, Rosenblatt, and Chentsov. Inference procedures whose validity do not rely on a specific model for the theoretical population distribution are called distribution-free inference procedures. Non-parametric refers to the properties of the inference problem itself. The term distribution-free applies to the methodological properties that are involved in solving inference problem.

iii Formalization of the law of large numbers in functional space and its relation to the learning processes by Vapnik and Chervonenkis (1971).

iv Development of algorithmic complexity and its relationship with inductive statistical inference, mainly proposed by Kolmogorov, Solomonoff, and Chaitin (M. Li and Vitanyi, 2008).

New concepts and new techniques came out from these ideas, and the combination of the statistic domain with other fields created new approaches. But the main core of the statistical learning is unquestionably the following: *What can we learn from data? What do data tell us?*. As conclusive remark, moving from the classic paradigm to the modern approach to face statistic issues, a recent definition of Statistical Learning that summarizes its aims and its development as an analytical procedure, is proposed by Bousquet (2004):

> "The main goal of statistical learning theory is to provide a framework for studying the problem of inference, that is of gaining knowledge, making predictions, making decisions or constructing models from a set of data. This is studied in a statistical framework, that is there are assumptions of statistical nature about the underlying phenomena (in the way the data is generated)."

- **Data Mining**
  Data Minin aims to extract useful information from large data sets or databases (Hand, Mannila, and P. Smyth, 2001). It is a general and rough definition, therefore it includes elements from statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other areas. Lying at the intersection of all the previous domains, it has developed is own methods and working tools, but preserving some features from all the mentioned fields. Data Mining approach has been conceived to be used when the data set is

massive, complicated, and/or may have problematic issues (for example in case of more variables than observations). Sometimes, the acronym KDD (Knowledge Discovery in Database) is used as synonym of Data Mining. Often, Data Mining is associated with the so-called *big data* (Han, Pei, and Kamber, 2011). The term big data, used for the first time in 1941 according to the *Oxford English Dictionary*, was preceded by very large databases (VLDBs) which were managed using database management systems (DBMS). During the 1990s, it was proofed that digital storage was by far more cost-effective than paper storage (Morris and Truskowski, 2003). Since then, year by year, the storage capacity of electronic devices has increased quickly and sharply, big data have become more accessible and widespread, and Data Mining has established itself has a key role paradigm. In this sense, a first difference can be highlighted from Statistical Learning: while Data Mining is consistent and suitable almost exclusively when dealing with big data, Statistical Learning is a more flexible approach for what concerns data size, cause it is designed to extract usefull patterns from data even when dealing with a small dataset. Data that can be considered "small" due to their size, are indeed suited to be analyzed by means of classic statistical inference. Also the steps to carry out in order to perform an exhaustive analysis are different in their definition, even if they can be somehow interchangeable from a content point of view. According to a general and validated standard of the middle 90s (Fayyad, Piatetsky-Shapiro, and P. Smyth, 1996), these phases can be summarized in 5 big steps:

1 Data Selection
2 Pre-processing
3 Transformation
4 Proper Data Mining Analysis
5 Interpretation/evaluation

A more recent formalization (Han, Pei, and Kamber, 2011) is even more focused about the term *data*, giving it the maximum emphasis:

1 Data Cleaning
2 Data Integration
3 Data Selection
4 Data Transformation
5 Data Mining

6 Pattern Evaluation

7 Knowledge Presentation

Making a comparison between these Data Mining steps and the Statistical Learning ones, sum differences arise. In Data Mining steps no theoretical formalization or analytical hypothesis to be confirmed/disconfirmed are made. Data Mining analysis is, explicitly, approaching data with no previous formalized hypothesis from an analytical point of view, giving to this analysis an exclusively *exploratory* nature, while Statistical Learning can be both exploratory or *confirmatory*. On the other hand, the phase of data collection in Statistical Learning, is designed also taking into account the concepts of target population and, if necessary, several explicit research questions.

- **Machine Learning**
  Machine learning has been born as a field of computer science that gives computer systems the ability to "learn"; therefore, computers are able progressively to improve performance on a specific task or to achieve a given goal. It is done by means of an efficient utilization of data and without being explicitly programmed for these purpose (Samuel, 1959), but allowing computers to learn, gradually but automatically, without constant human interaction. Most of the time, the crucial part is to As well as for Statistical Learning and for Data Mining, also for Machine Learning there are some essential components that can be expressed explicitly as 3 main steps to carry out a Machine Learning procedure (Domingos, 2012):

  1 Representation
  A classifier, that is a function that transforms input data into output category, must be represented in a formal language that the computer system can handle properly. Formalize a set of classifiers that the learner (computer system) can learn is crucial. Usually, a classifier makes use of some sort of "training data", on which it trains its skills to figure out the best rule of classification. The representation space is also known as basically the space of allowed models (the hypothesis space).

  2 Evaluation
  An Evaluation function has the role to make an objective comparison between classifiers, in order to figure out which ones are good and which are performing poorly, and possibly to establish a ranking between them. This function is named, when discussed in

different contexts, as utility function, loss function, scoring function, or fitness function. Evaluation allows to test a chosen model even against data that has never been used for training. In this phase, at the end of the evaluation, additional *hyper* - parameters can be estimated. These hyper - parameters estimation is often called *tuning*, and it refers to some aspects of the procedure that are considered to be known in advance, in order to start the representation phase. A few parameters are so usually implicitly assumed fixed when machine learning procedure has started, and at this stage is a worthy to go back to the beginning and test those assumptions, and eventually try other values.

3 Optimization
This last step is the phase when one can searches for the space of represented models to obtain better evaluations. The choice of optimization technique is crucial to the efficiency of the algorithm; it is the strategy how it is expected to reach the best model.

Given these features, Machine Learning approach is able to elaborate several powerful and useful tools to improve performances of computer systems overall. Therefore, several findings coming from machine learning field can be exploited also by researchers using a different approach, like Statistical Learning or Data Mining.
Traditionally, there have been two fundamentally different types of tasks in Machine Learning (Chapelle, Schölkopf, and Zien, 2009). The first one is *supervised learning*. Within this approach, given a set of input variables, the aim is to use such variables to obtain a good prevision about a set of output variables. The utopian goal to be achieved is to approximate the mapping function so well that once input variables are given, no errors occur in finding output variables values. Output variables are often called also targets or labels. It is called supervised learning because the whole process of the algorithm learning directly from the training set is made in a similar way in which teacher supervises a learning process. He knows the right answer (output), so the algorithm iteratively makes predictions based on the input using training data, and it is corrected by the teacher. Learning process stops definitively when the algorithm achieves a level performance that suits with a predetermined threshold. When the output is a set of continuous observations, the task leads to an analysis that falls in the regression family; when the output is a set of discrete - categorical observations, the task is solved as a classification problem.
On the other hand, in *unsupervised learning*, no output is provided

as intended in supervised learning. Thus, the goal for unsupervised learning is to explicitly understand and model the underlying structure (in terms of distribution or patterns) in data. All the observations are considered to be input, and no learning process is carried out to improve algorithm performance, given that there no teacher interference in suggesting the correct output. Computational procedures are left to their own to extract useful information and interesting structure to interpret from data. Some authors put the emphasis on the opportunity to exploit this kind of approach to figure out the random variables that has likely generated the observed data. Other techniques are quantile estimation, clustering, outlier detection, and dimensionality reduction. As last remark, is important to point out a pretty recent branch; *Semi-Supervised Learning* (SSL) (X. Zhu, 2006). It is, very intuitively, half-way between Supervised Learning and Unsupervised Learning. In this case, the algorithm faces some observation with no labels, usually in a pretty large dataset, while some others are target units provided with labels. Data matrix is therefore divided into two parts: a certain number of observations providing output labels, and certain observations where examples are without labels. So, SSL can be assumed to be a mixture of both approaches. A good example of real data scenario, is a images archive in which only few of the pictures provide a label, (e.g. mouse, cat, bird, dog) and a large part of them are not labelled. Due to this data structure, a mixture of supervised and unsupervised techniques can be used in this case.

Table 1.1 summarizes the most important features of the three previously described frameworks, stressing out similarities, differences and interconnections.

The tool proposed in this work, can be framed in the wide family of Statistical Learning. Since it does not require labelled units, it will be compared to other methods in the groups of Unsupervised Statistical Learning. Therefore we will focus, in the following, only on techniques belonging to this group, in order to make an exhaustive comparison with Archetypical Analysis for Histogram Data. Therefore, a section aimed to describe and introduce what are distributional data 2, as particular case of symbolic data analysis (SDA), is presented in the following section 1.4. Further, it will be deepen the relationship between interval-valued data and histogram-valued data within SDA approach, to exploit their intimate connection to discuss the statistical development of archetypal analysis.

Table 1.1 Comparison between Statistical Learning, Data Mining and Machine Learning

Analysis	***Main Aims***	***Data Size*** and ***Approach***	***Steps***	***Subgroups***
Statistical Learning	Provide a framework for studying problem as in classical inferential perspective but no only, with the aim of extract knowledge from a set of data	Both small and Big data - Both confirmatory and exploratory	Data collection, data management, data analysis, interpretation	Frequentist, Bayesian
Data Mining	Extract useful information from large datasets or databases	Only big data - Only exploratory (extract knowledge from data)	Data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge presentation	No real subgroups, but several techniques, such as Text Mining, Web Mining etc.
Machine Learning	Give computer systems ability to "learn"; develop algorithms; recently going closer to SL	Only big data - Exploratory (no theoretical background)	Representation, Evaluation, Optimization	Supervised, Unsupervised, Semi-Supervised

1.2 Statistical learning methodological challenges for Big Data

The development of the data-driven decision-making (DDDM) (Provost and Fawcett, 2013) approach involves making decisions that are based on objective data analyses, in the sense that business governance decisions are undertaken only if backed up with verifiable data. This approach excludes so the influence of intuitive decisions or decisions based on observation alone. Big Data are the core of several DDDM analyses, increasing the size and the heterogeneity of available data to be used in such analyses. With such vast amounts of data now available, and estimating an exponential increase of that amount in the next years, companies in almost every sector are deepening techniques and methodologies to exploit data information for competitive advantage (Mallinger and Stefl, 2015). So, bbbb the paradigm shift in statistical learning has been encouraged not only by academics and researchers, but also by the expectation of businessmen and entrepreneurs, that have been involved in this fast and radical change, and are facing this increasing amount of data playing an active role. In literature, the wide use of big data analysis in order to obtain predictive power and so gaining advantages in terms of revenues, is often called as Big Data Predictive Analytics (BDPA) (Dubey et al., 2017). To deal with the recent availability of big data, new perspectives about data analysis have been inspired from several fields, leading to the development of new (or modified) statistical techniques. Therefore, new theoretical framework has been developed, as well as new ways to accomplish practical analyses. Thus, both innovations, theoretical and practical, have lead to a paradigm shift in data analytics. As pointed out by Sinan Aral in Cukier, 2010.

> "Revolutions in science have often been preceded by revolutions in measurement."

Therefore, it is reasonable to assume that the creation of new ways to store, measure, analyse, interpret (in one word, conceive) data is the basis that stays at the foundations of a new paradigm (Kitchin, 2014). Some authors have claimed that this new paradigm is the establishment of "the end of the theory" (Anderson, 2008). This debate is centered around the data-driven approach as a result of big data era, with a reborn of an *empiricism* that does allow researchers to mine and retrieve information from data without making explicit theoretical hypotheses. This is an epistemological controversial, that points out how statistics as a whole has to face a new paradigm. According

to Kitchin, 2014, the strong emphasis on *big-data-driven* analyses has several properties, or features, that help to overcome several shortcomings of traditional deductive approach:

- Given that $n \to \infty$, in big data framework it is possible to represent a whole domain and, consistently, obtain results belonging to the entire population.
- There is less space for a priori hypothesis, and statistical models are less useful in this context.
- Analysis is without sampling design and/or questionnaire design potential biases. Data are able to have a meaningful and truthful trait by themselves.
- To the extreme, data meaning transcends context or domain-specific knowledge, thus anyone familiar with statistics and visualization could be able to interpret them in an exhaustive way. If the analysis is carried out taking into account mainly information included in data and retrieved directly from them, it means that context knowledge takes second place.

Thus, given all these factors, some questions come out about the role and the validity of the techniques embedded in the original inductive statistics toolbox. A major case is the wide discussion over the validity of the p-value in big data scenario. Proper *ad-hoc* developments and possible solutions can be found in Hofmann, 2015. Null hypothesis significance testing (NHST) is the classical way in which inductive statistics confirms or disconfirms underlying hypothesis. The controversy around it, and as a result around p-value issues, has been broadly discussed already in Nickerson, 2000. In case of $n \to \infty$, in the new paradigm some authors suggest to move from classical NHST interpretation of confidence interval and significance in general into confidence intervals for effect sizes, which are considered presumably the measures of maximum information density (Howell, 2011). Classical statistical inference, in such paradigm, seems so to have lost the predominant role with its traditional tools. Generally speaking, the interest has moved towards techniques that are able to *reduce* the data complexity, rather than found significant relationships or differences. This is mainly due to the fact that in ase of big data analysis the following conditions are true: target population is not defined in advance as in the traditional way, sampling theory has lost its dominant role as in classical inference procedures, p-value and NHST have to be used carefully (due to the fact that when $n \to \infty$ and so even very

small discrepancies should be significant), data sources and data nature are numerous and heterogeneous. As also pointed out in Franke et al., 2016, it is almost obvious that in such scenario the focus of statistical analyses is on dimensionality reduction and in finding salient units to summarize main patterns. This work tries to give a contribution in such direction.

1.3 Big Data: a very brief history

As claimed by Zikopoulos, Eaton, et al., 2011:

> "In short, the term Big Data applies to information that cant be processed or analyzed using traditional processes or tools. Increasingly, organizations today are facing more and more Big Data challenges."

Which kind of organizations are we talking about? Who was the first company to make aware use of big data? Who was the first scholar to come out with a formal definition of big data? The answers to these questions are neither obvious nor easy. Probably, not even so useful. As said, the paradigm shift has been pretty fast and driven by a key development of available technology, which has affected everybody, academy and companies. A brief history of big data is useful to figure out the general path, and to understand what are the new challenges in terms of fields of application. According to some authors, for example Barnes, 2013, the issues related to modern big data framework and paradigm shift are, overall, a prosecution of past debates around several statistical themes. He has framed its speculations in his own geographic field research. By the way, given that big data are usually generated continuously, quickly and in large numbers, some fields can be considered as the most important natural sources of big data. In the following, 5 of the most relevant sources for big data are summarized:

- Media technology. Data are complex (images, videos, sounds) and are generated very fast by electronic devices. A key role is played by social media platforms (Facebook, Twitter, Instagram and so on).
- Cloud platforms. These platforms are designed for storing massive amount of data. Cloud platforms can be private, public or third party.
- Web in general. This is probably the most obvious, still the most important. Most of the data available in the net are free and retrievable.
- Internet of Things (IoT). Data generated from the IoT devices and their interconnections. IoT is a system of interrelated computing devices,

mechanical and digital machines, that are provided with the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.

- Databases. In traditional form and in recent form. Structured or unstructured. For the most part, structured data are related to information with a very high degree of organization, whereas unstructured data is essentially the opposite.

In general, electronic devices like phones or IoT devices, produce massive, dynamic flows of heterogeneous, fine-grained, relational data. From a general perspective, it has been natural for companies acting in these fields to experiment an initial connection with big data analytics; academics involved in researches in that fields, as well, have conceived the new paradigm way earlier than others. Big data have, from that moment on, moved and influenced more fields year by year. Experts familiar with big data analytics have been able to apply the big data opportunities to new domains of application. For example, studies to improve benchmarking in medical sector have been proposed in Jee and G.-H. Kim, 2013. The main purpose of the study is to explore how and when use big data in order to effectively reduce healthcare concerns; especially for what concerns the selection of appropriate treatment paths, improvement of healthcare systems, and so on. For other public sectors, a wide review can be found in (G.-H. Kim, Trimi, and Chung, 2014), with a particular emphasis on enhancing government transparency and balancing social communities. Governments and official institutions have availability of a huge amount of data, especially collected in traditional forms, i.e. census data collection, but also an increasing capability to collect and analyze data that come from more recent sources, such as administrative sources. Data of different form (traditional, structured, unstructured, semi-structured, complex and so on), even if gathered from different sources, can belong to the same public sector, and their simultaneous analysis in order to obtain an exhaustive information retrieve process, is a hard challenge to face. The efficient use of big data analytics could provide sustainable solutions for the present state of art, and suggest future decisions to undertake, making a more aware use of information from policy makers. In the next section, challenges of the use of big data in educational sector will be deepen.

1.3.1 Big data in education and massive testing

In the contemporary era of big data, authors interested about the relationships between the paradigm shift regarding innovative data analytics and educational system, often refers to the the concept of *big educational data*.

In Macfadyen, Dawson, et al., 2014, Learning Analytics (LA) is defined as the possibility of implementing assessments and feedbacks in real-time evaluation systems. Learning analytics provides higher education helpful insights that could advice strategic decision-making regarding resources distribution to obtain educational best-performances. Further, the aim in this context is to process data at scale that are focused mainly on improvement of pupils' learning and to the development of self regulated learning skills, under the assumption that to improve cognitive skills it is necessary to customize learning process with respect to teachers and students needs and requirements.However, also in this field, to accomplish these kinds of aims, a shift in culture is needed: from assessment - for - accountability to assessment for learning (Hui, G. T. Brown, and S. W. M. Chan, 2017). In the former, the evaluation is made because it is just a duty to accomplish, also because there is a law that makes it mandatory, and so people involved in a given system are aware that they *have* to carry out a process of evaluation. In the latter, efforts are made to use evaluation findings to undertake new policies, with the aim to improve future learning processes in school system. The new possibility to make use of big data analytics tools has become the major innovation in order to reach that cultural change. In Manyika et al., 2011 it has been outlined how data in general expands the capacity and ability of organizations, even public sectors, to make sense of complex environments, and educational system belongs without any doubt to the group of complex environments. Due to budget restrictions and increasing heterogeneity in learners, scholastic programmes and teachers' background, several authors has claimed that using Learning Analytics in big data era is not a potential advantageous option but indeed an imperative that each organization has to pursue (Macfadyen and Dawson, 2012). This kind of approach will lead to optimize educational systems, making an efficient use of funds allocated for schools, highlighting good and bad practices mining information from data (Mining, 2012). A key role in that sense is played by Learning Management Systems (LMSs) (M. Brown, 2011). Several researchers and technical reports corroborates how learning management systems have the ability to increase student sense of community (both at scholastic and university level). Further, they can help to provide support in learning communities and enhance student engagement and success.
One of the most important and well-known source of big data in education is worldwide the PISA assessment in OECD organization framework [1]. Its main aim is to assess, by means of standardized and comparable tests, pupils' learning/cognitive skills across the world. The core of PISA tasks can

[1] https://www.oecd.org/pisa/

be summarized with the words of OECD secretary-general Angel Gurra [2]:

> "Quality education is the most valuable asset for present and future generations. Achieving it requires a strong commitment from everyone, including governments, teachers, parents and students themselves. The OECD is contributing to this goal through PISA, which monitors results in education within an agreed framework, allowing for valid international comparisons. By showing that some countries succeed in providing both high quality and equitable learning outcomes, PISA sets ambitious goals for others."

These ambitious goals can be seen as benchmarking best-performances to look forward. As well as scholastic performances in terms of proficiency by itself, PISA tests data are also addressed to figure out how specific sociological, cultural, economical and demographic variables are able to affect the overall pupils' results.
From all these hints and previous researches, it is clear that:

i Big data are available also in educational system, both from official institutions (such data from Minister of Education) as well as from tests to assess pupils' skills.

ii New statistical learning paradigms go in parallel with new cultural framework in education: from assessment - for - accountability to assessment - for - learning.

iii It is becoming crucial to adopt decisions based on big data analytics, and expectations are that policy makers use findings from big data in conscious way. From policy makers perspective, it is not only advantageous to use such findings, but somehow mandatory nowadays.

iv Learning processes can be improved if results are correctly used and interpreted, leading to a customization in learning processes.

v Proficiency tests like PISA, the one carried out by OECD, but many others all around the world, are crucial sources of big data in education, and proper tools should be created and checked to analyze such tests.

This work aims to address the last item since it presents a new tool to study complex data in education, showing it in action on real data and in particular to the Italian case of proficiency scores grouped by school.

[2] https://www.oecd.org/pisa/pisaproducts/37474503.pdf

1.4 Introduction to Symbolic Data Analysis

If data are made up by n objects or individuals, where each generic unit i is defined by a set of collected values from different variables of size k, with a generic variable j, data matrix has the classic structure $\mathbf{X}_{(n,k)}$ (1.1). In contrast, symbolic data with measurements on k random variables are k-dimensional hypercubes (or hyperrectangles) in $\mathcal{R}^k$, or a Cartesian product of k distributions, broadly defined (Billard and Diday, 2006). A single point unit is therefore a special, and the simplest, case of symbolic data, that leads to the described classic form of matrix (1.1) SDA provides a framework for the representation and interpretation of data that comprehends inherent variability. Units under analysis in this approach, usually called entities, are therefore not single elements, but groups (or clusters, or set of units) gathered taking into account some given criteria. This leads to consider that there is an internal variation within each variable for each group. Furthermore, when dealing with concepts, such as animal species, pathologies description, athletes types, and so on, data involve an intrinsic variability that can not be neglected.
Each observation in SDA has, thus, a more complex structure, with this internal variation that has to be taken into account; while dealing with simple points, only variation *between* observations is the core of the analysis. In SDA approach, the intrisic and comprehensive structure of observations leads to deal with *within* variation, as additional source of variability. From an interpretative point of view, symbolic objects plays a key role in statistics for complex data, cause they are suited to model *concepts*. This is a notion that has been developed inside Formal Concept Analysis (FCA), that is a framework laying mainly on the borders of Ontology and Information Systems, of which the first founder is considered Rudolf Wille in the early 80s (Wille, 1982). The original motivation of FCA was the aim to find for real-world situations and contexts a confirm of mathematical order theory. FCA deals with structured data which describe relationship between units and a peculiar set of attributes/characteristics. FCA aims to produce two different kinds of output from the input data. One is a concept lattice, that is an agglomeration of concepts described in formal way contained in the data, usually hierarchically ordered using subconcept-superconcept relation (Belohlavek, 2008). Such formal concepts are intended as representation of, basically, natural concepts that human beings have in mind in an intuitively way, such as “mammal organism”, “electric car”, “number divisible by 2 and 5”, and so on. The second finding of FCA is the attribute implications, that describes the way in which a particular dependency, included in data, comes from formal concepts; e.g., “respondent with age under 15 are at high

schools", "all mammals in data have 4 feet", and so on. Many authors tried to figure out a valide hierarchical structure for the framework (i.e. Priss, 2006); to be more precise, in FCA formal concepts are defined to be a pair (E, I), where E is a set of objects (called the *extent*) and I is a set of attributes (the *intent*). A *category* is a specific value assumed by a categorical variables, that so defines a group of units belonging to the same kind (such as birds of the same species). A *class* is a set of units, that are analyzed in the same context and once merged together form an unique dataset. Concepts are therefore the more complex structure in this theoretical thinking, and that's where the significant role of Symbolic Objects come from.
From a pure philosophical and ontological point of view, authors claim (Bock and Diday, 2000), that a great advantage of symbolic data analysis is that symbolic objects thus defined are able to make a synthesis of the following different theoretical tendencies that are cornerstones in the ontological tradition:

- Aristotelian Tradition
  The link is in the fact that symbolic concepts can have the explanatory power of logical descriptions of the concepts that they represent, given that concepts are characterised by logical conjunction of several properties.
- Adansonian Tradition
  Since the units of all the extension of a symbolic object are *similar* in the sense that they satisfy the same properties as much as possible, even if not necessarily Boolean ones. In that sense, the concepts that they represent are polythetic, so they cannot be defined by only a conjunction of properties, but members of same group will share most of the properties. This because in Adansonian Tradition a concept is characterised by a set of similar individual.
- Rosch prototypes
  Cause their membership function is able to provide prototypical instances characterized by the most representative attributes and individuals. So, prototypes will be the typical-type inside a given community, according to its features able to represent the category.
- Wille property (FCA)
  This property refers to the fact that an object is wholly described by means of a Galois lattice (Ganter and Wille, 1996) Given that SDA is derived directly from FCA, the so - called "complete symbolic objects" of SDA can be proved to be a Galois lattice, so this property is satisfied.

Symbolic data can be of different natures: intervals, histograms, distributions, lists of values, taxonomies and so on. According to the given nature of the data, different techniques and approaches are developed in order to analyze them, and some examples of well - known symbolic data are in (1.2), showing different kind of visual representation. A formal definition of symbolic variable in presented in Bock and Diday, 2000. A comprehensive ontology of the different nature of symbolic variables is proposed by Noirhomme-Fraiture and Brito, 2011, where variables are first of all divided between numerical and categorical, and then hierarchically partitioned due to their nature, as depicted in 1.1.

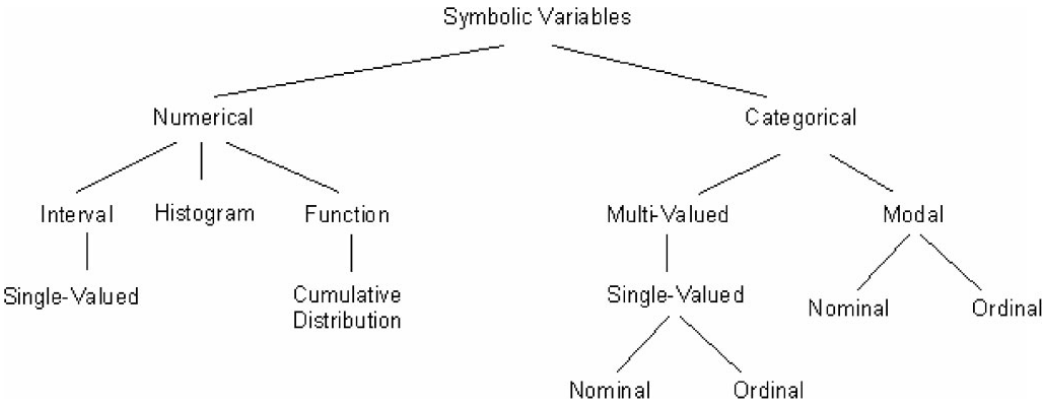
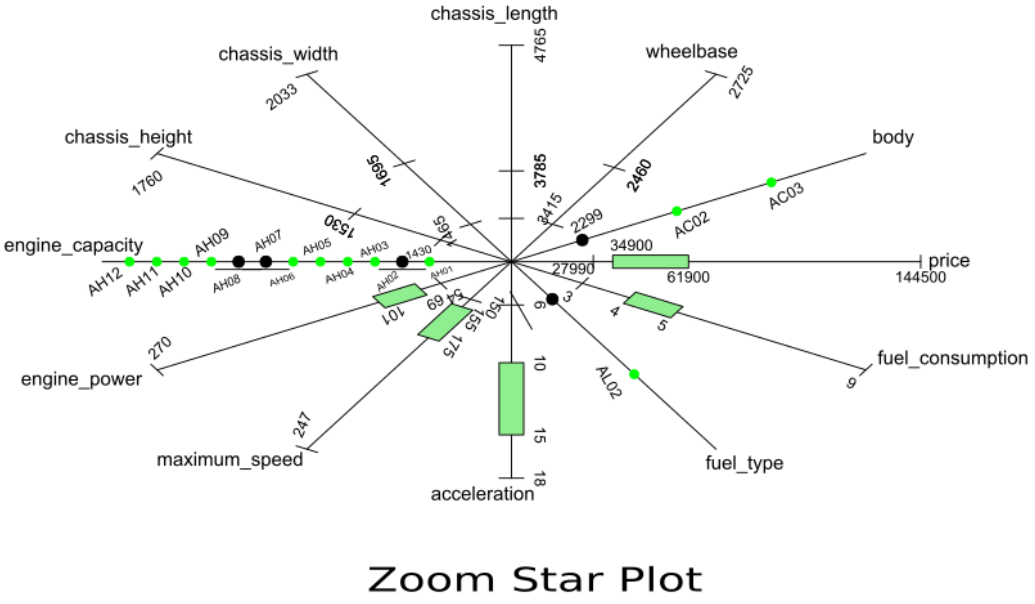
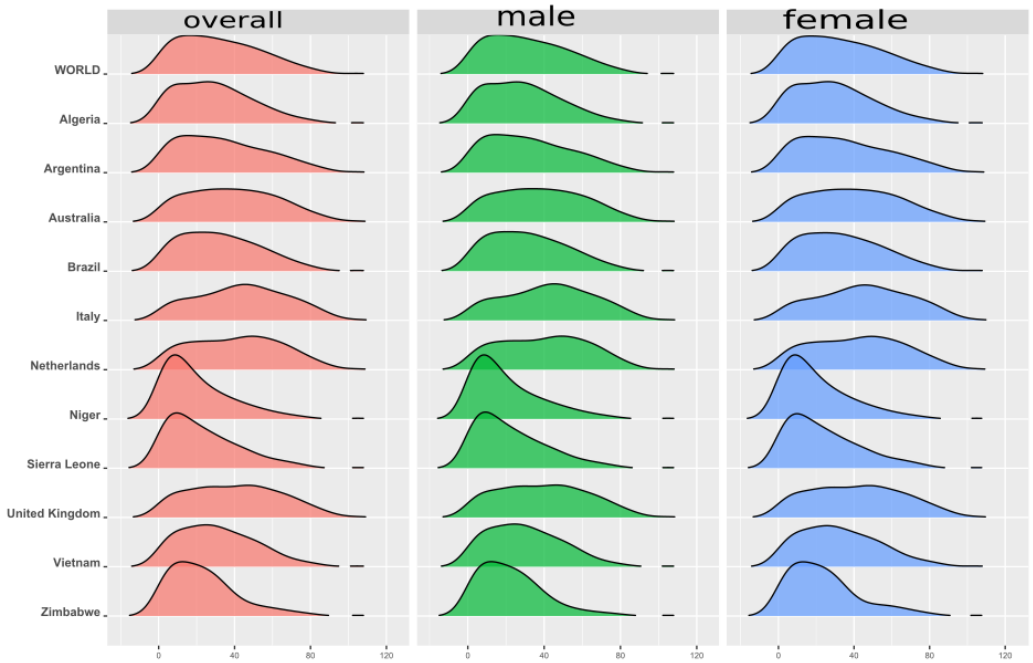


Figure 1.1: Ontology of symbolic variables, taken from Noirhomme-Fraiture and Brito, 2011.

In case of *interval data*, i.e. data expressed by interval of $\mathbb{R}$, a pair of numbers $[a, b]$ represents all the numbers $a \leq x \leq b$. One of the first proposed formalization to deal with such data was the interval arithmetic (Moore, Kearfott, and Cloud, 1979), that has introduced and defined algebraic properties and has proposed metric for such data. In particular, authors have started to conceive the *fuzzy set* theory (Moore and Lodwick, 2003) linked to the analysis of interval data. In case of categorical data, a remarkable approach is the one related to *compositional data* (Aitchison, 1982). A compositional data point is a representation of a part of a whole, like percentages, probabilities or proportions. Usually it is represented by a positive real vector with as many parts as considered. For example, if we look at the elements that compose the Planet Earth structure, we see that in first place there is iron (32.1%), followed by oxygen (30.1%), silicon (15.1%), magnesium (13.9%), while all the others elements account for 8.8% . Therefore, a compositional way to represent the data point "Earth" is vector of 5 elements: $[0.321, 0.301, 0.151, 0.139, 0.08]$. A dataset of compositional data,



(a) ZoomStar for different kind of Symbolic data (9 interval data, 3 multinomial or distributional data) in *cars* dataset. Credits to the R package **symbolicDA**



(b) Histogram-value data of some countries by age and gender, represented by means of Histogram. Credits to the R package **HistDAWass**

Figure 1.2: Various patterns for symbolic data and their representation

in this example, will be made up by a different in each row , resulting then in a vector of length 5 as observed value .
In the following section, the algebra and the features of distributional/histogram valued data will be deepen.

Chapter 2

Elements of Distributional and Histogram-valued Data

2.1 Distributional Data and Histogram-valued data in SDA

Distributional Data are a specific kind of data embedded in SDA. Each observation is defined by a *distribution*, in the wide sense of the term. It can be a frequency distribution, a density, a histogram-valued data or a quantile function. Such data are assumed to be a realization of a numeric modal symbolic variable. In particular, modal variables can model the description of an individual, of a group, or of a concept, by probability distributions, frequencies or, in general, by random variables (Irpino and Verde, 2015). For example, data retrieved from official statistics as macrodata, are usually described by means of basic statistics. When estimating a parameter, the estimation is presented under the distribution of such estimation in several samples. In these cases, even if we are observing a single variable, the proper expression of such variable is not a single value but instead a multi-valued quantity.
According to the literature (Bock and Diday, 2000, Noirhomme-Fraiture and Brito, 2011), for histogram symbolic data we consider the situation in which the support is continuous and finite, and each observed value is an histogram developed over such continuous finite support, as in 1.2. To deepen the structure of numerical nature of symbolic variables, each quantitative variable, from a general perspective, may then be *single-valued* (real or integer) as in the in classical framework, if it exspresses one single value per observation. Moving to the proper domain of SDA, a variable is *multi-valued* if its values are a finite vector of numbers belonging to a finite numerical support.

Further, *interval* variable occurs if its values are intervals (Moore, Kearfott, and Cloud, 1979). Usually, when dealing with an empirical distribution over a set of subintervals, the variable is called a *histogram-valued* variable.
Let's define X as the variable, D as its underlying domain and R as its range, i.e. where is theoretically possible to express its values. Given a set of statistical units of size n, $\mathbf{S} = (s_1, s_2, ..., s_n)$, in SDA framework, is possible to sum up the information contained in $\mathbf{S}$ by means of an application, leading to the symbolic variable X made up by p different variables, and with $i = 1, ..., n$:

$$X : S \to D \text{ such that } s_i \to X(s_i) = \alpha \tag{2.1}$$

where α is the single numeric result of the application and $D \subseteq \mathbb{R}$. It means that, with such application to S, all the values of the range R are still plausible, and are equal to the entire domain D. This is, given that only one α is the outcome, the simplest case of SDA, when it becomes a single standard numeric variable with only one realization for each observation.
When, on the other hand, values of $X(s_i)$ are finite sets of α_i, (2.1) becomes:

$$X : S \to R \text{ such that } s_i \to X(s_i) = (\alpha_{(1i)}, \alpha_{(2i)}, ..., \alpha_{(pi)}) \tag{2.2}$$

leading to a finite set of realization for each observation i. The defined variable deriving from (2.2) is a multi-valued ones, so. The application creates a finite set of values that describe the statistical unit.
In case of interval-valued data, the application leads to:

$$X : S \to R \text{ such that } s_i \to X(s_i) = [l_i, u_i] \tag{2.3}$$

I is in this case a $(n \times p)$ matrix containing the values of p interval variables on S. Therefore, each p-tuple of intervals $I_i = (I_{(i,1)}, I_{(i,2)}, ..., I_{(i,p)})$ defines a specific $s_i \in S$. Lastly, histogram-valued data are in this approach made up by aggregating microdata in several intervals (or bins) inside lower bound and upper bound $[l_i, u_i]$, providing more information than interval data about data distribution.

$$X : S \to R \text{ such that } s_i \to X(s_i) = [I_{i1}(p_1), I_{i2}(p_2), ..., I_{ik}(p_k)] \tag{2.4}$$

In this context, $[I_{i1}, I_{i2}, ..., I_{ik}]$ are the set of sub-intervals, associated with the observed frequencies $(p_1, p_2, ..., p_k)$. Therefore, it could be deepen the analysis of internal variation between the maximum and the minimum value of the distribution, while in interval data in (2.3) is not possible to make specific assumption about frequency distribution, but only about general distribution between boundaries. Further, the usually hypothesis made for in histogram-valued data, is that in each subinterval data are uniformly distributed. Of

course, if there is only one bin in the histogram structure, (2.4) simplifies in (2.3), and therefore interval-valued data are a special case of histogram-valued data.

2.1.1 Main statistics for histogram-valued data

The issue of histogram-valued data, and symbolic data in general, when trying to calculate basic descriptive statistics, is the nature itself of such data (Irpino and Verde, 2015). Each descriptive statistic has to take into account the degree of internal variation that exists inside observations, while in single-valued data only between observations variation is considered. So, questions arise about to the extent to which classical formalization and classical concepts of descriptive statistics can be adopted in case of distributional data. The central core of the different approach is, then, the dispersion evaluation inside each observation. As introduced in 2.1 and formalized in (2.4), inside each bin (sub-interval) of the histogram data are considered to be distributed as an uniform random variable. So data are equally spread from lower bound of the bin to the upper bound of such bin. Formally, for each h interval where $h : (I_1, I_2, ..., I_h, ..., I_k)$:

$$\phi_i X = \sum_{j<h} p_{ji} + p_{hi} \cdot \frac{x - l_{ji}}{u_{ji} - l_{ji}} \text{ where } (j = 1, ..., k) \tag{2.5}$$

where ϕ_i is the density distribution for the variable X calculated in i. We, thereafter, consider the definition proposed in (ibid.) for distributional symbolic variable:

Definition 2.1.1. A modal variable is called a distributional symbolic variable if for all i the measure ϕ_i has a given density ϕ_i, and so is possible to simplify the relationship as: $X_i = \phi_i$.

The debate about univariate and bivariate statistics for histogram-valued data is a consequence of the starting approach and the theorized paradigm that is behind the formalization of such distributional data. In SDA a common groundwork is that there are two different level of real data collection (Bock and Diday, 2000). First level, the lowest one, is related to elementary units. Aggregating together micro-data from basic units leads to obtain upper level data. This kind of histogram-data are so considered a generalization of observed values in a group of lower-level units. The analysis procedure that moves the computation from first level to second level has been deepen mainly by (Bertrand and Goupil, 2012; Billard and Diday, 2003). Most of the

assumptions are expressed about a generic set N formed by a number n of elementary units. It can be fully described by a distribution-valued variable X, with $X_i = \phi_i$. Going straightforward to the statistics, it implies that the mean, the variance and the standard deviation are the result of such statistics computed on a mixture of n density functions, one for each observation in the set N containing 1-level units. Usually, if all units are equally likely to be present in N, weights used to create the mixture are all equal to $\frac{1}{n}$. As presented in (Frühwirth-Schnatter, 2006), resulting mean of the mixture $\sum_{i=1}^{n} \frac{1}{n}\phi_i = \phi$ is:

$$E(X) = \mu = \sum_{i=1}^{n} \frac{1}{n}\mu_i \tag{2.6}$$

Further, variance is defined as:

$$E[(X-\mu)^2] = \sigma^2 = \sum_{i=1}^{n} \frac{1}{n}(\mu_i^2 + \sigma_i^2) - \mu^2 \tag{2.7}$$

In this approach, symbolic distributional data are suited to represent real situation in which group of individuals are the basis to form an upper level entity (employees nested in companies, pupils nested in schools and so on). It is, of course, a context that happen pretty often, therefore the 2-level paradigm has a wide range of possible application in real life, cause data that are naturally organized in hierarchical order are not uncommon. Further, if previous knowledge are available, is possible to change the weights to calculate the ϕ mixture giving more importance to groups that are known, for example, to be larger. Anyway, univariate statistics thus conceived are implicitly assuming that there is no significant difference between individuals inside the same group, or, at least, this kind of paradigm is not able to catch such heterogeneity. Indeed, switching values assigned to different individuals belonging to the same group, overall means and variance don't change. This framework does not allow to compare individuals aside from their groups.
In descriptive statistics context, mean (as expected value) can assume several definition and formalization to extend straightforward its properties and formalization to different kind of data. If we take into account recent developments and formalization of Frechet mean (Ginestet, Simmons, and Kolaczyk, 2012; Nielsen and Bhatia, 2013) such that:

Definition 2.1.2. with n elements described by the variable X, a *di* distance between two descriptions and a set of n real numbers $Z = (z_1, ..., z_n)$, a Frechet type mean (known as *barycenter*) M_{Fr} is the $argmin$ of the following function:

$$M_{Fr} = \underset{x}{arg\,min} \sum_{i=1}^{n} z_i di^2(x_i, X) \tag{2.8}$$

The minimization problem in 2.8, is a generalization problem of finding an entity of central tendency in a cluster of points (also called *centroid*). Other kind of means, such as harmonic or geometric means, are just the extension of the 2.1.2 using different kind of distance *di*.
Chisini mean (Graziani and Veronese, 2009) applies another approach to the definition of a mean. Often authors refers to the Chisini mean as representative or substitutive mean (Dodd, 1940), due to its definition that is considered to be useful in practical context. Formally:

Definition 2.1.3. a Chisini mean of single-valued variable X and a function F such that $F(x_1, ..., x_i, ...x_n)$ applied to a set of n object, the mean is defined as:

$$F(x_1, ..., x_i, ...x_n) = F(M_{chisini}, ..., M_{chisini}, ..., M_{chisini}) \tag{2.9}$$

where $(M_{chisini}, ..., M_{chisini}, ..., M_{chisini})$ is a vector that is the mere repetition of the Chisini mean n times.

Due to some analytical issues, the equation in (2.9), could not have a finite solution, and further the Chisini mean could be external to the interval of observations $[x_{min}, ..., x_{max}]$. Consequently, usually some constraints are imposed to the function F in order to obtain an unique final solution to the minimization problem.
It has to be pointed out that, to extend both Chisini and Frechet means to distributional variables, first step is to define a proper distance measure between distribution (as histogram-valued data). Several dissimilarity and distance measures between such symbolic data have been proposed (J. Kim, 2009) and compared to each other by several authors. The underlying idea about the overall comparison of two histogram-valued data in SDA is that the comparison has to take into account way more statistical aspects than in the comparison between two single-valued data. For example, given the different internal variation inside each distributional observation, two histograms could share even same mean and/or median, but nevertheless have a dissimilarity (or a distance) > 0. This can be due to a different degree of dispersion, in terms of variation, around a fixed central tendency index. Therefore, the basic idea behind distributional distances is that they should be able to compare much part of distributions as possible.
Some of such dissimilarity measures come from the extension of a given dissimilarity measure suited for interval-valued data, and then developed to deal with histogram-valued data. The underline assumption is that is possible moving from (2.3) and extend that formalization to (2.4), and so that if a measure is valid to compare two intervals, it is as well a proper choice to compare two groups of intervals (each of them being an histogram). To

formalize such kind of measures embedded in this approach, in the following we define *intersection* and *union* between histogram objects taking into account their relative frequencies p linked to the nb number of bins. For this formalization, histograms are assumed to be adequately transformed in order to have exactly same sub-intervals.

Definition 2.1.4. For two different histogram object $Hist_1$ and $Hist_2$, where for each bin there are relative frequencies such that $Hist_1 \rightarrow (p_{(1,1)}, p_{(1,2)}, ..., p_{(1,nb)})$ and $Hist_2 \rightarrow (p_{(2,1)}, p_{(2,2)}, ..., p_{(2,nb)})$, intersection between them is defined as:

$$Hist_1 \cap Hist_2 = p_{1,i} \cap p_{2,i}, \text{with } i = 1, ..., nb \tag{2.10}$$

$$\text{where each } p_{1,i} \cap p_{2,i} = \min_{1,...,i,...,nb}(p_{1,i}, p_{2,i}) \tag{2.11}$$

Therefore, from (2.11) we formalize the concept of intersection as a comparison, bin by bin, of their respective density, and considering only the amount of *shared density*. From a graphical point of view, such shared density is the only part in which histograms bins overlap 2.1.
Further, *union* between histogram objects is defined consequently as:

Definition 2.1.5. For two different histogram object $Hist_1$ and $Hist_2$, where for each bin there are relative frequencies such that $Hist_1 \rightarrow (p_{(1,1)}, p_{(1,2)}, ..., p_{(1,nb)})$ and $Hist_2 \rightarrow (p_{(2,1)}, p_{(2,2)}, ..., p_{(2,nb)})$, union between them is defined as:

$$Hist_1 \cup Hist_2 = p_{1,i} \cup p_{2,i}, \text{with } i = 1, ..., nb \tag{2.12}$$

$$\text{where each } p_{1,i} \cup p_{2,i} = \max_{1,...,i,...,nb}(p_{1,i}, p_{2,i}) \tag{2.13}$$

So, union thus defined leads to compare as well bins densities one by one of both histogram, but the result is the maximum value that such density shows. It is worthy to note that the sum of across bins densities in case of union is ≥ 1, while on the other hand sum of such densities in case of intersection is ≤ 1.

2.1.2 Main dissimilarity/distance measures for histogram-valued data

Some dissimilarity measures take into account, in their calculation, elements from both intersection and union between histograms, and that's why is useful to introduce them as part of histogram algebra. Further, some extensions of dissimilarity measures that developed from interval to continuous data, include concept of mean and standard deviation derived from formalization

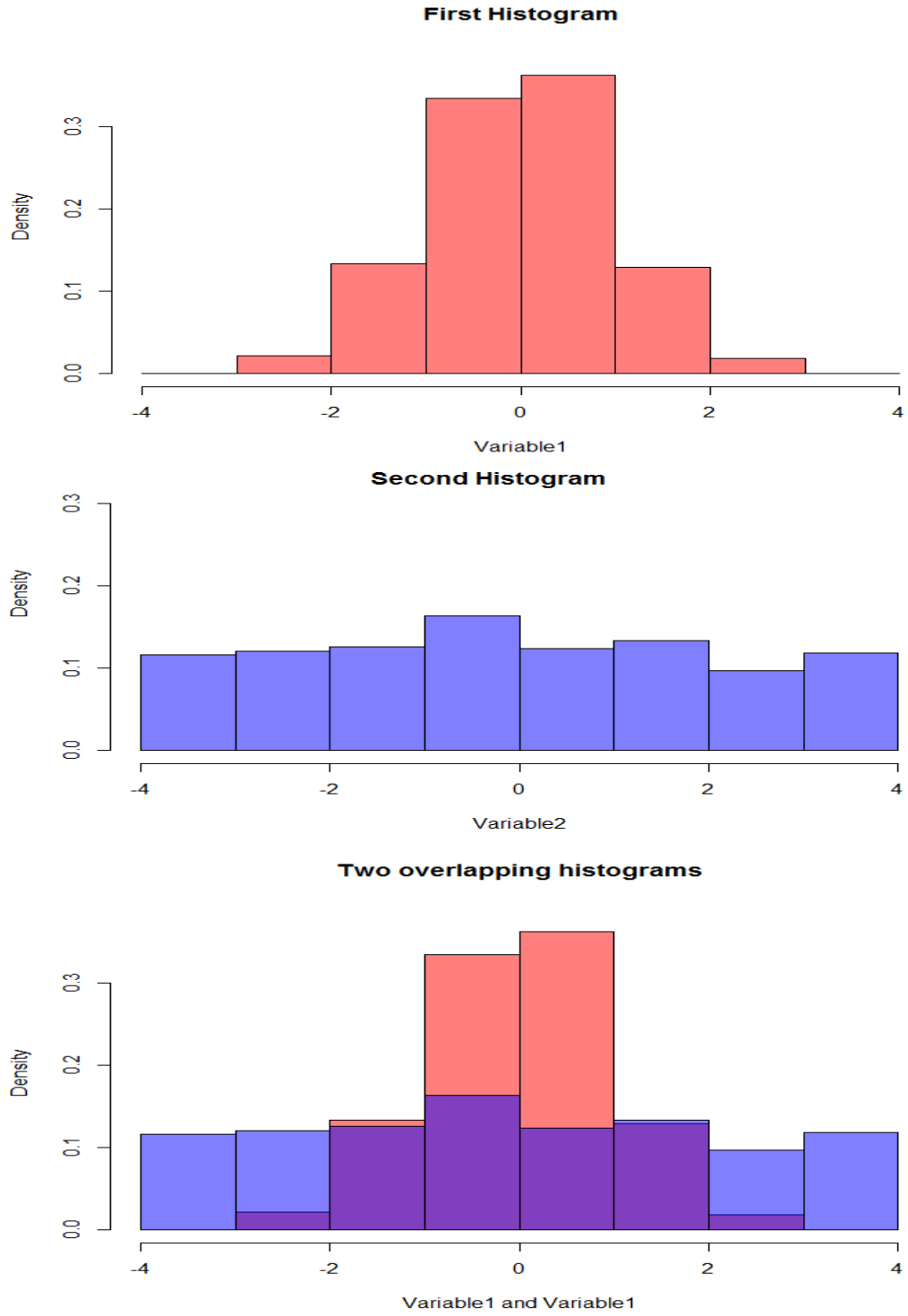


Figure 2.1: Two histograms and their overlapping densities. Intersection as defined in (2.1.4) is the purple part where they overlap, union as defined in (2.1.5) is the maximum height of each bin (red or blue)

of Billard and Diday (2003). It starts from the general assumption of uniform distribution in each sub-interval. So, to compute the mean inside each of them it has to be applied a simple arithmetic mean between boundaries l_i and u_i. Formally:

Definition 2.1.6. Given an histogram object $Hist$ that is made up by nb number of bins and as in 2.3 with l_i the lower bound of such bins and u_i the upper bound, mean according to Billard is defined as:

$$M_{Hist} = \sum_{i=1}^{nb} \frac{l_i + l_{i+1}}{2} p_i \tag{2.14}$$

Therefore, as mentioned, mean in (2.14) is a weighted mean (using observed density) of central values across bins. Standard deviation start from same assumption, and as formalized in (J. Kim, 2009):

Definition 2.1.7.

$$SD_{H.} = \sqrt{\sum_{i=1}^{nb} \frac{(l_i - M_{H.})^2 + (l_i - M_{H.})(u_i - M_{H.}) + (u_i - M_{H.})^2}{3} p_i} \tag{2.15}$$

These formalization are valid in case of natural histogram-valued data. When dealing with objects coming instead from a previous procedure of union or intersection of histogram objects, such definitions need a correction due to the fact that, as mentioned, (2.6) has to hold that $\sum_{i=1}^{nb} p_i = 1$. In case of union as defined in (2.1.5) such sum is $\sum_{i=1}^{nb} p_i > 1$ and in case of intersection from (2.1.4), $\sum_{i=1}^{nb} p_i < 1$. Therefore, we introduce the quantity p^* that is the normalization of such sum in order to obtain $\sum_{i=1}^{nb} p_i = 1$.

$$p^*_{(Hist_1 \cup Hist_2)i} = \frac{p_{(Hist_1 \cup Hist_2)i}}{\sum_{i=1}^{nb} p_{(Hist_1 \cup Hist_2)i}} \tag{2.16}$$

$$p^*_{(Hist_1 \cap Hist_2)i} = \frac{p_{(Hist_1 \cap Hist_2)i}}{\sum_{i=1}^{nb} p_{(Hist_1 \cap Hist_2)i}} \tag{2.17}$$

Therefore, mean and standard deviation of objects derived from union or intersection of histograms will take into account definitions in 2.16 and 2.17 to have estimation of such univariate statistics consistent with usual properties.

Definition 2.1.8. Mean for an histogram-valued object derived from union or intersection of simple histogram-valued objects are computed as:

$$M^*_{(Hist1 \cup Hist_2)} = \sum_{i=1}^{nb} \frac{l_i + l_{i+1}}{2} p^*_{(Hist_1 \cup Hist_2)i} \tag{2.18}$$

$$M^*_{(Hist1 \cap Hist_2)} = \sum_{i=1}^{nb} \frac{l_i + l_{i+1}}{2} p^*_{(Hist_1 \cap Hist_2)i} \tag{2.19}$$

Definition 2.1.9. Standard deviation for an histogram-valued object derived from union or intersection of simple histogram-valued objects are computed as:

$$SD^*_{(H.1 \cup H.2)} \tag{2.20}$$
$$= \sqrt{\sum_{i=1}^{nb} \frac{(l_i - M^*_{(H.1 \cup H.2)})^2 + (l_i - M^*_{(H.1 \cup H.2)})(u_i - M^*_{(H.1 \cup H.2)}) + (u_i - M^*_{(H.1 \cup H.2)})^2}{3} p_{(H.1 \cup H.2)i}}$$

$$SD^*_{(H.1 \cap H.2)} \tag{2.21}$$
$$= \sqrt{\sum_{i=1}^{nb} \frac{(l_i - M^*_{(H.1 \cap H.2)})^2 + (l_i - M^*_{(H.1 \cap H.2)})(u_i - M^*_{(H.1 \cap H.2)}) + (u_i - M^*_{(H.1 \cap H.2)})^2}{3} p_{(H.1 \cap H.2)i}}$$

From this moment on, for reason of brevity, $M^*_{(Hist1 \cap Hist_2)}$ will be $M^*_{\cap}$ and $M^*_{(Hist1 \cup Hist_2)}$ will be $M^*_{\cup}$. These measures will be embedded, in the following, in several distance/dissimilarity measures for distributional data that will take into account union and intersection between histograms. Some measures that will compute an outcome about the diversity between two symbolic objects are formalized as *distance measures*, while others are in the family of *dissimilarity measures*. The concepts of similarity, dissimilarity and distance extended to probability functions or histogram-valued data has been largely addressed in the literature (Cha, 2007; L. Lee, 1999). Similarity is formalized as:

Definition 2.1.10. Given two generic different objects a and b belonging to the same space Ω such that are comparable, a similarity measure S holds the following properties:

- $S(a; b) = S(b; a)$
- $S(a; a) = S(b; b) > S(a; b)$ for all $a \neq b$

First property is a formalization of symmetry that exists in any similarity measures, and second property stresses out as the maximum allowed value for such measure is the similarity between an object and itself. From the definition in 2.1.10, dissimilarity is derived as follows:

Definition 2.1.11. Given two generic different objects a and b belonging to the same space Ω such that are comparable, a dissimilarity measure D holds the following properties:

- $D(a;b) = D(b;a)$
- $D(a;a) = D(b;b) < D(a;b)$ for all $a \neq b$
- $D(a;a) = 0$ for all $a \in \Omega$

As for similarity, dissimilarity concept is formalized as symmetric one. A measure of dissimilarity such defined is always positive and is equal to zero any time that the measure is computed between an object and itself. Most of the time, is possible to transform a dissimilarity measure to a similarity measure (by means of simple mathematical tools) cause one is defined as *inverse functional* of the other. Usually, when dealing with traditional data point such single-valued data, the dissimilarity can be measured by distance measures. When dealing with symbolic data, on the other hand, further properties have to be satisfied by a proper *distance measure* to calculate a dissimilarity matrix between symbolic data (Nieddu and Rizzi, 2007) :

Definition 2.1.12. Given 3 generic different objects a, b and c, belonging to the same space Ω such that are comparable, a distance measure D holds all the properties of a dissimilarity measure as in 2.1.11, and further satisfies:

- $D(a;b) = 0 \rightarrow a = b$
- $D(a;b) \leq D(a;c) + D(b;c)$ for all $a, b, c \in \Omega$

Therefore, a distance measure in 2.1.12 satisfies the so-called triangle inequality (Khamsi and Kirk, 2011). As mentioned, several measures derived as an extension of similarity, dissimilarity and distance measures for interval data to histogram data. One of these is the *Gowda-Diday* similarity/dissimilarity measure (Diday and Esposito, 2003; Gowda and Diday, 1991a,b) for interval-valued data.
Authors defined 3 different quantities, normalized to be between 0 and 1, that are necessary to calculate such Gowda-Diday similarity (S_{GD}) and Gowda-Diday dissimilarity (D_{GD})measures.

Definition 2.1.13. First part of the S_{GD} measures the relative sizes of two interval-objects (X_1, X_2) in general. nb is the number of intervals, and $i = 1, \cdots, nb$. This part does not refer to the common parts between them:

$$S_1(X_1, X_2) = \frac{|u_{i1} - l_{i1}| + |u_{i2} - l_{i2}|}{2\big|\max(u_{i1}, u_{i2}) - \min(l_{i1}, l_{i2})\big|} \tag{2.22}$$

Definition 2.1.14. Given that:

$$\Gamma_{(i1,i2)} = \begin{aligned} &\left|\max(l_{i1}, l_{i2}) - \min(u_{i1}, u_{i2})\right| \text{ if } \max(l_{i1}, l_{i2}) \\ &< \min(u_{i1}, u_{i2}) \text{ or } = 0 \text{ otherwise} \end{aligned} \tag{2.23}$$

Second part of the S_{GD} measures the common parts between the two interval objects, similarly to 2.1.4:

$$S_2(X_1, X_2) = \frac{\Gamma_{(i1,i2)}}{\left|\max(u_{i1}, u_{i2}) - \min(l_{i1}, l_{i2})\right|} \tag{2.24}$$

Definition 2.1.15. Lastly, third part is related to the relative position of the two objects in the space, cause denominator is equal to the range, as the total length (from minimum value to maximum value) showed by the variable:

$$S_3(X_1, X_2) = 1 - \frac{|l_{i1} - l_{i2}|}{\max_i(u_i) - \min_i(l_i)} \tag{2.25}$$

Merging together the three different components, Gowda-Diday similarity S_{GD} is:

$$S_{GD}(X_{i1}, X_{i2}) = \sum_{i=1}^{nb}(S_1(X_{i1}, X_{i2}) + S_2(X_{1i}, X_{i2}) + S_3(X_{i1}, X_{i2})) \tag{2.26}$$

In case of multivariate contest in which there are present p interval-valued variables with $j = 1, \cdots, p$, and objects (X_1, X_2) are multi-interval objects, (2.26) extents to:

$$S_{GD}(X_{i1}, X_{i2}) = \sum_{i=1}^{nb}\sum_{j=1}^{p}(S_1(X_{i1j}, X_{i2j}) + S_2(X_{1ij}, X_{i2j}) + S_3(X_{i1j}, X_{i2j})) \tag{2.27}$$

The similarity measure (2.27) has been extended to accomplish the dissimilarity measure D_{GD} that is made up, in an analogue way to the similarity measure, in 3 different components:

Definition 2.1.16. D_{GD} is defined by 3 components: $Diss_1$ is due to position, $Diss_2$ due to spanning shared quota, $Diss_3$ is related to the content (relative position).

$$Diss_{GD}(X_{i1}, X_{i2}) = \sum_{i=1}^{nb}\sum_{j=1}^{p}(Diss_1(X_{i1j}, X_{i2j}) + Diss_2(X_{1ij}, X_{i2j}) + Diss_3(X_{i1j}, X_{i2j})) \tag{2.28}$$

The measure in (2.27) and its related dissimilarity measure in 2.1.16 has been deepen by authors in the following years after 1991, and it was argued that such measures considered together suffered from several shortcomings:

- If there is no overlapping parts among two interval-valued data, the dissimilarity measure is greater than the similarity measure.
- In case of identical length of the two intervals, the similarity measure is greater than the dissimilarity measure.
- Switching from (2.27) to the dissimilarity measure, the third component of the former is just another way to reproduce the latter.

According to (Gowda and Ravi, 1995), a new measure has been proposed as a sine-function of only 2 components to overcome such shortcomings:

Definition 2.1.17. S^*_{GD} is defined by 2 components, both sine functions:

$$S^*_1(X_{i1j}, X_{i2j}) = \text{sine}\left[90\left(\frac{|u_{i1j}-l_{i1j}|+|u_{i2j}-l_{i2j}|}{2|\max(u_{i1j},u_{i2j})-\min(l_{i1j},l_{i2j})|}\right)\right] \text{ and}$$

$$S^*_2(X_{i1j}, X_{i2j}) = \text{sine}\left[90\left(1-\frac{|a_{i1j}-a_{i1j}|}{\max_i(u_{ij})-\min_i(l_{ij})}\right)\right]$$

$$S_{GR}(X_{i1}, X_{i2}) = \sum_{i=1}^{nb}\sum_{j=1}^{p}(S^*_1(X_{i1j}, X_{i2j}) + S^*_2(X_{1ij}, X_{i2j})). \tag{2.29}$$

The simple development from 2.1.17 in order to create the dissimilarity measure led to switch from the sine-function to the cosine-function, due to their complementary nature.

Definition 2.1.18. $Diss^*_{GD}$ is defined by 2 components, both sine functions:

$$D^*_1(X_{i1j}, X_{i2j}) = \cos\left[90\left(\frac{|u_{i1j}-l_{i1j}|+|u_{i2j}-l_{i2j}|}{2|\max(u_{i1j},u_{i2j})-\min(l_{i1j},l_{i2j})|}\right)\right] \text{ and}$$

$$D^*_2(X_{i1j}, X_{i2j}) = \cos\left[90\left(1-\frac{|a_{i1j}-a_{i1j}|}{\max_i(u_{ij})-\min_i(l_{ij})}\right)\right]$$

$$Diss_{GR}(X_{i1}, X_{i2}) = \sum_{i=1}^{nb}\sum_{j=1}^{p}(D^*_1(X_{i1j}, X_{i2j}) + D^*_2(X_{1ij}, X_{i2j})). \tag{2.30}$$

While dealing with multi-valued histograms $\mathbf{X_{i1j}}$ and $\mathbf{X_{i2j}}$, the proposed extension of Gowda-Diday dissimilarity measure is:

Definition 2.1.19. $Diss^H_{GD}$ is defined by 3 components:
$D^H_{1j}(X_{i1j}, X_{i2j}) = \frac{\left|SD_{i1j} - SD_{i2j}\right|}{SD_{i1j} + SD_{i2j}}$
$D^H_{2j}(X_{i1j}, X_{i2j}) = \frac{SD_{i1j} + SD_{i2j} - 2SD_{(iHist1 \cap iHist2)j}}{SD_{i1j} + SD_{i2j}}$ and
$D^H_{3j}(X_{i1j}, X_{i2j}) = \frac{\left|M_{i1j} - M_{i2j}\right|}{u_{j,nb_j+1} - u_{j1}}$ and so at the end the final dissimilarity measure is:

$$Diss^H_{GD}(X_{i1}, X_{i2}) = \sum_{i=1}^{nb}\sum_{j=1}^{p}[D^H_{1j}(X_{i1j}, X_{i2j}) + (D^H_{2j}(X_{i1j}, X_{i2j}) + (D^H_{3j}(X_{i1j}, X_{i2j})] \tag{2.31}$$

Such components have similar meaning to the components defined for interval-valued data 2.1.16: $D^H_{1j}(X_{i1j}, X_{i2j})$ is a measure of relative size, $D^H_{2j}(X_{i1j}, X_{i2j})$ is related to the relative content, and the last part $D^H_{3j}(X_{i1j}, X_{i2j})$ indicates the relative position. But, obviously, these relative features are computed using means and standard deviations related to the nature of data, that are switched from intervals to histograms. Therefore, in this way more information about dispersion around the mean will be taken into account for each components. For each variable j, each component $0 < D^H_j < 1$.
Another largely addressed measure that was created originally for interval data in 1994 (Ichino and Yaguchi, 1994) is the *Ichino-Yaguchi dissimilarity*. The more recent formalization of this dissimilarity measure for histogram-valued data is proposed in (J. Kim and Billard, 2013). The original Ichino-Yaguchi dissimilarity proposed for interval-valued data is:

Definition 2.1.20. Given two interval-valued sets $\mathbf{X_{i1}}, \mathbf{X_{i2}}$ for a variable j, by means of cartesian operators *join* that is $\oplus$ and *meet* that is $\otimes$ such that their meaning is:

$$X_{i1} \oplus X_{i2} = [min(l_{i1j}, l_{i2j}), max(u_{i1j}, u_{i2j})] \tag{2.32}$$

$$\begin{aligned} X_{i1} \otimes X_{i2} = [\max(l_{i1j}, l_{i2j}), \min(u_{i1j}, u_{i2j})] \text{ if } \max(l_{i1j}, l_{i2j}) \\ < \min(u_{i1}, u_{i2}) \text{ or } = 0 \text{ otherwise} \end{aligned} \tag{2.33}$$

And defining γ_{IY} as a constant such that $0 < \gamma_{IY} < 0.5$, complete formula for Ichino-Yaguchi dissimilarity is:

$$\Phi(X_{i1j}, X_{i2j}) = \left|(X_{i1j} \oplus X_{i2j}\right| - \left|(X_{i1j} \otimes X_{i2j}\right| + \gamma_{IY}(2\left|X_{i1j} \otimes X_{i2j}\right| - \left|X_{i1j}\right| - \left|X_{i2j}\right|) \tag{2.34}$$

Also in this case, as for Gowda-Diday dissimilarity, the (2.34) is not a normalized measure, and needs to be divided by it's theoretical maximum value in order to have a dissimilarity measure between 0 and 1:

Definition 2.1.21. Normalized Ichino-Yaguchi dissimilarity measure is given by:

$$\Phi^*(X_{i1j}, X_{i2j}) = \frac{\Phi(X_{i1j}, X_{i2j})}{\max_i(u_{ij}) - \min_i(l_{ij})} \tag{2.35}$$

Starting from this approach, DeCarvalho (deCarvalho, 1994, 1998) extended Ichino-Yaguchi measure in (2.1.21) to each kinf of constrained Boolean objects, broadly speaking. Boolean symbolic objects (*BSO*), are objects that take into account into account simultaneously the variability , as range of values observed, and some kinds of logical dependencies between variables; these objects are therefore multi-valued. A *BSO* is properly defined by means of logical conjunction of properties. In DeCarvalho, a comparison function and an aggregation functions are the ground of the formalization to assess proximity level of *BSO*. A comparison function is defined in this approach as a proximity index based on a measure that is always ≥ 0, kwnown as description potential of a Boolean elementary event. It is cardinal of the disjunction of values on a variable of a *BSO*. Authors have proposed, to formalize comparison functions, indexes related to agreement and disagreement. Given two interval-valued variables (X_{i1j}, X_{i2j}), a quantity $c(X_{ij})$ that is the complementary part of the variable X_{ij} given its domain set, agreement-disagreement indexes are summarized in the following table:

Table 2.1 Disagreement and Agreement formalization in comparison of 2 Boolean interval objects, proposed by DeCarvalho

	Agreement	Disagreement	Marg. Total
Agreement	$AA = \lvert X_{i1j} \cup X_{i2j} \rvert$	$AD = \lvert X_{i1j} \cup c(X_{i2j}) \rvert$	$\lvert X_{i1j} \rvert$
Disagreement	$DA = \lvert c(X_{i1j} \cup X_{i2j}) \rvert$	$DD = \lvert c(X_{i1j}) \cup c(X_{i2j}) \rvert$	$\lvert c(X_{i1j}) \rvert$
Marg. Total	$\lvert X_{i2j} \rvert$	$\lvert c(X_{i2j}) \rvert$	Dom.(X_j)

The total of the right-below corner of the table 2.1, Dom.(X_j), is the global domain of the variable j.
From such table, DeCarvalho proposed five different similarity measure (known as comparison functions) to compare two *BSO*.

$$f_1 = \frac{AA}{AA + AD + DA} \tag{2.36}$$

$$f_2 = \frac{2AA}{2AA + AD + DA} \tag{2.37}$$

$$f_3 = \frac{AA}{AA + 2(AD + DA)} \tag{2.38}$$

$$f_4 = \frac{1}{2}\left[\frac{AA}{AA + AD} + \frac{AA}{AA + DA}\right] \tag{2.39}$$

$$f_5 = \frac{AA}{\sqrt{(AA + +AD)(AA + DA)}} \tag{2.40}$$

In these comparisons, the related dissimilarity function is $d_{DCl} = 1 - f_l \text{with} = (1, \cdots, 5)$. Both similarities and dissimilarities functions thus defined are between 0 and 1. Analogously, for an observed variable j, an aggregation function is a proximity index, and authors proposed to start from the Minkowski distance. From a general perspective, Minkowski distance between two interval-valued objects for variable j is defined as:

$$D_M^q(X_{i1j}, X_{i2j}) = \left[\sum_{i=1}^{nb} \phi(X_{i1j}, X_{i2j})^q\right]^{1/q} \tag{2.41}$$

This equation in (2.41) is defined using as ϕ the one proposed by 2.1.21 and as q a positive number. As q changes, different measures will make the comparison between the two *BSO*. While dealing with several variables, (2.41) becomes:

$$D_M^q(X_{i1}, X_{i2}) = \left[\sum_{i=1}^{nb}\sum_{j=1}^{p} \phi(X_{i1j}, X_{i2j})^q\right]^{1/q} \tag{2.42}$$

Therefore, final DeCarvalho dissimilarity measure takes into account both dissimilarity functions and aggregation functions. Given a set of weights W of length j such that each $w_j > 0$ and $\sum_j w_j = 1$, such proposed dissimilarity measure for multi-variable histogram data is:

$$D_{d_{DCl}}^q(X_{i1}, X_{i2}) = \left[\sum_{i=1}^{nb}\sum_{j=1}^{p} w_j d_{DCl}(X_{i1j}, X_{i2j})^q\right]^{1/q} \tag{2.43}$$

While ichino-Yaguchi, Gowda-Diday and DeCarvalho dissimilarities have been developed originally for interval-valued data and then extended to histogram-valued data, others measure have born directly as a tool to evaluate the degree of distance between probability density functions. It is likely

that the most exhaustive survey on such distances can be found in (Cha, 2007). Author has proposed a classification of such measures in several families, according to both theoretical assumptions behind the computation and according to a hierarchical clustering based on empirical correlation of a simulation study. Syntactic similarity, implementation aspects and semantics closeness are the most important features that are analysed to assess similarity between these measures. Particular emphasis is made on an entire family of distances that has come out from the concept, conceived in information theory field, of Shennon's metric for Entropy (Shannon, Weaver, and Burks, 1951). The entropy of a variable, in broad way, is the amount of information contained in a specific variable, in terms of innovative knowledge that its observed values are able to provide. Shannons entropy quantifies such amount of information, leading to theoretical and practical buildings around the information conception. Starting from definition of Shannons entropy derived from Boltzmann's -theorem as:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) log_b \mathrm{P}(x_i) \tag{2.44}$$

In 2.44, X is a variable where $\mathrm{P}(x_i)$ is the assumed probability to observe the symbol i and b is a generic value as base of the logarithm. Common values are $b = (2, e, 10)$. One well-known proposed measure is KullbackLeibler divergence (Kullback and Leibler, 1951), that is an asymmetric measure that takes into account relative entropy between one density function P_i and the reference distribution Q_i.

$$\mathrm{d}_{KL} = \sum_{i=1}^{nb} \mathrm{P}_i log_b \frac{\mathrm{P}_i}{\mathrm{Q}_i} \tag{2.45}$$

In the 2.45 the implicit assumption is that one distribution Q_i is considered as reference one, therefore the measure is able to catch how P_i is far, in terms of relative entropy, from the reference. That's where the asymmetric nature of 2.45 comes from. A symmetric version of such divergence is known as Jeffreys Divergence (Jeffreys, 1998).

$$\mathrm{d}_{KL} = \sum_{i=1}^{nb} (\mathrm{P}_i - \mathrm{Q}_i) log_b \frac{\mathrm{P}_i}{\mathrm{Q}_i} \tag{2.46}$$

In the case of 2.46, neither P_i nor Q_i are assumed to be a reference of any kind, so the comparison between them is calculated in a symmetric fashion. K-divergence is instead given by:

$$\mathrm{d}_{KL} = \sum_{i=1}^{nb} \mathrm{P}_i log_b \frac{2\mathrm{P}_i}{\mathrm{Q}_i + \mathrm{P}_i} \tag{2.47}$$

This divergence, as 2.45, follows an asymmetric rationale in the comparison. Its symmetric measure is defined as:

$$\mathrm{d}_T = \sum_{i=1}^{nb} \left[\mathrm{P}_i log_b \left(\frac{2\mathrm{P}_i}{\mathrm{Q}_i + \mathrm{P}_i} \right) + \mathrm{Q}_i log_b \left(\frac{2\mathrm{Q}_i}{\mathrm{Q}_i + \mathrm{P}_i} \right) \right] \tag{2.48}$$

Divergence in 2.48 is named Topsøe (M.-M. Deza and E. Deza, 2006), or information statistics (Gavin et al., 2003). Topsøe divergence could be halved, leading to a new distance measure called Jensen-Shannon divergence measure (Lin, 1991):

$$\mathrm{d}_{JS} = frac12 \sum_{i=1}^{nb} \left[\mathrm{P}_i log_b \left(\frac{2\mathrm{P}_i}{\mathrm{Q}_i + \mathrm{P}_i} \right) + \mathrm{Q}_i log_b \left(\frac{2\mathrm{Q}_i}{\mathrm{Q}_i + \mathrm{P}_i} \right) \right] \tag{2.49}$$

It is also known as information radius(IRad)(Manning and Schütze, 1999) or total divergence to the average (Dagan, L. Lee, and Pereira, 1997). As well as being symmetric, this measure is able to overcome another shortcoming. Kullback-Leiber measure 2.45 is theoretically allowed to not be a finite value, while 2.49 is always a finite value.
Jensen difference (Taneja, 2001) comes from the analysis about the relationship between idea of information radius and the concavity property of Shannon's entropy. Therefore, Jensen difference is defined as:

$$\mathrm{d}_{JD} = \sum_{i=1}^{nb} \left[\frac{\mathrm{P}_i log_b \mathrm{P}_i + \mathrm{Q}_i log_b \mathrm{Q}_i}{2} - \left(\frac{\mathrm{P}_i + \mathrm{Q}_i}{2} \right) log_b \left(\frac{\mathrm{P}_i + \mathrm{Q}_i}{2} \right) \right] \tag{2.50}$$

Another measure that is worthy to introduce is Jaccard Similarity (Jaccard, 1901), that is derived from the normalization of the inner product between two probability functions:

$$\mathrm{S}_{Jaccard} = \frac{\sum_{i=1}^{nb} \mathrm{P}_i \mathrm{Q}_i}{\sum_{i=1}^{nb} \mathrm{P}_i^2 + \sum_{i=1}^{nb} \mathrm{Q}_i^2 - \sum_{i=1}^{nb} \mathrm{P}_i \mathrm{Q}_i} \tag{2.51}$$

Irpino and Verde (Irpino and Verde, 2006) have proposed in 2006 a measure that starts from a different approach. Starting from a wide review about probability metrics that can be found mainly in (Gibbs and Su, 2002), authors adopt and extend a l_2 Wasserstein distance (Rüschendorf, 2001). A generic l_p Wasserstein distance (Givens, Shortt, et al., 1984) expresses the distance between two observed densities ϕ_1 and ϕ_2 using the inverse of such density function, namely the quantile functions ϕ_1^{-1} and ϕ_2^{-1}:

$$\mathrm{W}_p = \left(\int_0^1 \left[\phi_1^{-1}(t) - \phi_2^{-1}(t) \right]^p dt \right)^{\frac{1}{p}} \tag{2.52}$$

In case of $p = 2$, Wasserstein distance l_p becomes l_2 and its definition simplifies as:

$$\mathrm{W}_2 = \sqrt{\int_0^1 \left[\phi_1^{-1}(t) - \phi_2^{-1}(t)\right]^2 dt} \tag{2.53}$$

As pointed out by the authors, equation (2.52) is one of the possible extension of another distance for quantile functions, the classical L_p Minkowski distance, and therefore (2.53) is the extension of L_2 Minkowski distance. Quantile functions are known to have several statistical properties that can be useful in many contexts (Gilchrist, 2000). First of all, quantile functions are in univocal and unique relationship with their original density functions. Quantile functions are also always non-decreasing in the interval of their domain, that is between $[0 : 1]$. Further, it is left-continuos (Pfeiffer, 1990). In case of (2.53), so when $p = 2$, the Frchet mean defined in (2.1.2) and its objective function (2.8) of a generic distribution variable X, with respect to Wasserstein distance under the assumption of equal weights z_i, is the density function, corresponding to the quantile function that shows an average behaviour. This mean quantile is the solution of the optimization problem, analogously to (2.8):

$$M_W(X) = \underset{x}{arg\ min} \sum_{i=1}^{n} d_W^2(\phi_i, X) \tag{2.54}$$

The minimum value that solves (2.54) is equal to the probability density function having as quantile function x^{-1}:

$$x^{-1}(t) = \frac{1}{n}\sum_{i=1}^{n} \Phi_i^{-1}(t) \to \bar{\Phi}^{-1}(t), \quad \forall \quad t \in [0, 1] \tag{2.55}$$

Therefore, as final remark, is possible to compute the Frechet mean distribution as barycenter of the distributions; calculation start from the relationship one-to-one between distribution and its quantile, such that:

$$M_{Fr}(X) = \bar{\Phi} = \frac{d\left(\bar{\Phi}^{-1}\right)^{-1}}{dx} \to \frac{d\bar{\Phi}}{dx} \tag{2.56}$$

2.2 Clustering methods for histogram-valued data

In statistical context, cluster analysis (Hartigan, 1975; Jain and Dubes, 1988) is a family of procedure able to achieve the task to group a number of object

together, based on their characteristics. Therefore, these elements are merged together according to their level of similarity. Cluster analysis does not define one specific algorithm, but it is instead the general task to be solved. It can be pursued by means of several different algorithms and the choose of the proper distance measure (to assess the level of similarity between statistical units) plays a key role. Of course, analytical choices are derived, first of all, from the nature itself of the data. For what concerns histogram-valued data, we will introduce in the following three different kind of clustering: *hierchical clustering*, *k-means clustering* and *fuzzy k-means clustering*. First two methods, hierarchical clustering (Rokach and Maimon, 2005) and k-means clustering (Hartigan, 1975; Na, Xumin, and Yong, 2010) are defined as hard-clustering techniques. In hard clustering, each data point either belongs to a cluster completely or not. Formally (J. Kim, 2009):

Definition 2.2.1. If we have p random variables $(X_j, j = 1, \cdots, p)$, with symbolic objects, $(x_i, i = 1, \cdots, n)$ and with $x_i \in \Omega \rightarrow (x_1, \cdots, x_n)$, a *partition* P_r of Ω is a finite set of subsets such that $(C_u, u = 1, \cdots, r)$ that also satisfies:

- $C_u \cap C_v = \phi, \forall u \neq v = 1, \cdots, r$
- $\bigcup_{u=1}^{r} C_u = \Omega$

It means that all the subsets $(C_1, \cdots, C_r)$ of a given r partition P_r are disjoint, and exhaustive of the entire set Ω. Further, these subsets are considered to be non-empty in such a way that every element is included in one and only one of the subsets (Halmos, 2017).

2.2.1 Hierarchical clustering for histogram-valued data

In data mining and statistics in general, hierarchical clustering (also known as hierarchical cluster analysis or HCA) is a consolidated method of clustering objects starting from the idea that is possible to build a hierarchy of clusters. Formally (J. Kim, 2009):

Definition 2.2.2. A hierarchical structure on the space Ω is a finite set of subsets $H \rightarrow (C_u, \cdots, C_r)$ such that:

- $\Omega \in H$
- $\forall \quad x_i$ in Ω, single objects $x_i \in \Omega$

- $\forall \quad C_u,\ C_v \in H \quad \forall \quad v \neq u \rightarrow (v,u) = 1, \cdots, n$ and $C_u \cap C_u \in (\Phi, C_u, C_v)$.
  This property points out that each pair of clusters is disjoint, or one subset is contained into the other.

Several different strategies have been developed in order to obtain optimal hierarchical clustering, and it is possible to split them into two major big families. First one is *agglomerative*. The rationale is that in the beginning each elements x_i is considered as a cluster by itself, and then, step by step, units are merged in the same cluster according to their similarity. At the end, all units belonging to Ω are grouped in the same cluster. On the other hand, the *divisive* approach starts where the agglomerative approach ends. All units, as starting points, belong to only one cluster, and then are recursively assigned to several groups. The former is a *bottom-up* approach, while the latter is a *top down*. These two different approach have different properties and shortcomings. As claimed by several authors (Wilks, 2011), the agglomerative clustering method is more widely used than the divisive clustering method. This is mainly due to some computational issues related to the number of possible bi-partitions that are theoretically present when performing a divisive algorithm ($2^{n-1} - 1$). On the other hand, divisive algorithm is considered to be a better choice if is possible to overcome such complexity, cause is able to face in a better way first steps of the procedure. It also guarantees a better representation of the main structure that is behind the data
Further distinction is made between *polythetic* and *monothetic* algorithms (Wiggerts, 1997). A distinguishing feature of a polythetic algorithm is that the criteria that is consistently used to perform the cluster is based on all variables at the same time, in a simultaneous fashion. On the other hand, monothetic algorithms work independently on one variable per time. Therefore, the goal behind the analysis is different: monothetic clustering leads to have groups which elements share some properties, while polythetic clustering merges together units that are close (in terms of similarity) but not necessarily show same values of some variables. Furthermore, in polythetic analysis, observations are not ordered according to some specific features, but the algorithm identifies a splitting observation, or splinter cluster (that is an entity, an unit or a group, that is clearly different and separated from the rest of the data) considering all variables. Then the remaining observations, one at a time, are moved into such group if enough similar according to the given dissimilarity distance, and not moved if such similarity is not achieved. In the following, two different approach for hierarchical clustering will be presented and briefly discussed:

1 A divisive polythetic algorithm using both Euclidean extended Ichino Yaguchi dissimilarity and cumulative density functions extensions (Billard and J. Kim, 2017; J. Kim and Billard, 2011).

2 An agglomerative hierarchical clustering using a Wasserstein based metric proposed in (2.52, 2.53) based on the Ward criterion (Irpino and Verde, 2006).

For what concerns first proposal, author were interested in the clustering of histogram data considering the construction of hierarchical trees by using a polythetic clustering algorithm. The algorithm they have proposed is based on dissimilarity matrices that contain dissimilarity measures between observations. Authors, after a brief review about distance-dissimilarity measures and univariate statistics about histogram-valued data, introduce the analytical aim of their analysis. Main task is to obtain a divisive clustering of the complete set of observations Ω, where each group is internally as homogeneous as possible, while comparing clustering, they have to be externally as heterogeneous as possible. Given a pre-assigned stopping rule R, in each iteration, the algorithm determines which C_u has to be divided into (C_u^1, C_u^2). It means that in each iteration, until the end, clusters are recursively divided and units moved into a new group just created. Given that $w = (1, 2)$, that $\Omega = P_r = (C_1, \cdots, C_r)$, that m_v is the size in terms of number of elements of the cluster C_v and that $\lambda = \sum_{u_1}^{m_v} w_u$ authors have proposed the average weighted dissimilarity such that:

$$\overline{D}_v(X_u^v) = \frac{1}{\lambda - w_u} \sum_{u_1 \neq u = 1}^{m_v} w_u w_{u_1} d\left(X_u^v, X_{u_1}^v\right), (u = 1, \cdots, m_v) \tag{2.57}$$

In this equation, $d\left(X_u^v, X_{u_1}^v\right)$ is the dissimilarity between the two observations inside the brackets, $(X_u^v, X_{u_1}^v)$. Index u_1 shifts in the summation, therefore such dissimilarity is between X_u^v and all the other elements belonging to the same cluster. From this, authors propose the maximum average weighted dissimilarity (MAWD):

$$MAWD = \max_{u,v}(\overline{D}_v(X_u^v)) \quad (u = 1, \cdots, m_v; \quad v = 1, \cdots, r) \tag{2.58}$$

If an observation maximizes (2.58), so that its dissimilarity is the biggest calculated in the data, such observation will be moved in the cluster in the next iteration, and so its original cluster C will be split into two clusters. The open question is the calculation of the new units rearrangement across the two new clusters from the original one. Let's say that the two sub-clusters are now

(C_1^*, C_2^*). Then, each observation is moved from C_1^* to C_2^* and the within-cluster variations between the two sub-clusters is calculated. When such variation is negative, the units keep staying in the original cluster C_1^*, and when such variation is positive units move to C_2^* (J. Kim and Billard, 2011). Further, to perform divisive hierarchical cluster, has also been proposed the within- cluster variance as a criterion partitioning a cluster (Chavent, 1998, 2000):

$$I(C_u) = \frac{1}{2\tau} \sum_{i_1=1}^{n_u} \sum_{i_2=1}^{n_u} w_{i_1} w_{i_2} D^2(X_{i_1}, X_{i_2}) \tag{2.59}$$

Authors have shown main results of this technique using both simulated data and real data about diabetes disease (downloaded from `stanford.edu/hastie/Papers/LARS/`). Main findings are that, as expected, running the same algorithm while changing the measure that is used to compute the dissimilarity matrix will slightly change the final result of the hierarchical clustering; further, as widely pointed out, using histogram-valued data instead of classical data will lead to have a better understanding of the underlying variation of the phenomena while with classical data-points, taking into account only between data variation, a portion of information is lost.
The other procedure to perform hierarchical-clustering has been proposed by (Irpino and Verde, 2006). . The distance they introduce, discussed in (2.52, 2.53), holds interesting properties with respect to hierarchical clustering. It indeed allows to define a measure of inertia of data related to a barycenter that satisfies the Huygens theorem of decomposition of inertia (Haas, 1925). The core idea is that, according to (Billard and Diday, 2003) histogram data can be considered as a special case of compositional data. Compositional data (Aitchison, 1982), as introduced in 1.4, are made up by vectors of non-negative real components having a constant sum. Usually, when components are conceived as part or percentage of the whole, their sum is equal to one. In this case, histogram data (which components sum, as any probability function, is equal to 1) can be considered as a special case of compositional data. As introduced by (Mallows, 1972a), this metric, that evolved from the L^2 Kantorovich metric, can be considered as the expected value of the distance between homologous points of the supports of the two distributions, considering as measure the squared Euclidean one. As presented in 2.4, let's say that an histogram description of i by means of n_i is made up by intervals with density π_{ui} (while in 2.4 it was denoted by p) as follows:

$$X_i \to [(I_{1i}, \pi_{1i}), \cdots, (I_{2i}, \pi_{2i}), \cdots, (I_{ui}, \pi_{u,i}), \cdots, (I_{n_i i}, \pi_{n_i i})] \tag{2.60}$$

The following function in defined as the cumulative weights associated with

the elementary intervals of X_i in case of $l = 1, \cdots, n_i$:

$$w_i = \sum_{h=1}^{l} \pi_{hi} \tag{2.61}$$

On the other hand, for $l = 0$, all the elements $w_i = 0$. Under the assumption of uniformity distribution in each interval, from density function ϕ_i is possible to define distribution function Φi:

$$\Phi_i(z) = w_i + (z - \underline{z}_{li}) \frac{w_{li} - w_{l-1i}}{\underline{z}_{li} - \bar{z}_{li}} \tag{2.62}$$

In this case, $\underline{z}$ is lower bound of interval and $\bar{z}$ is upper bound, with $\underline{z} < z < \bar{z}$. Its inverse function is a step function that assumes different values such that:

$$\Phi_i^{-1}(t) = \begin{cases} \underline{z}_{1i} + \dfrac{t}{w_{1i}}(\bar{z}_{1i} - \underline{z}_{li}), & \text{for } 0 \leq t < w_{1i} \\ \underline{z}_{1i} + \dfrac{t - w_{l-1i}}{w_{li} - w_{l-1i}}(\bar{z}_{li} - \underline{z}_{li}), & \text{for } w_{l-1i} \leq t \leq w_{1i} \\ \underline{z}_{n_i i} + \dfrac{t - w_{n_i-1i}}{1 - w_{n_i-1i}}(\bar{z}_{n_i i} - \underline{z}_{n_i i}), & \text{for } w_{n_i-1i} \leq t < 1 \end{cases}$$

From this, authors evaluate that each couple (w_{l1}, w_l), starting from distance in 2.53, permits to identify two uniformly dense intervals, one for i and one for j, such that:

$$I_{li} = [Phi_i^{-1}(w_{l-1}; Phi_i^{-1}(w_l)] \quad I_{lj} = [Phi_j^{-1}(w_{l-1}; Phi_j^{-1}(w_l)] \tag{2.63}$$

Assumption independence still holds as in SDA framework, so is possible to express each interval in "radius and center" form; therefore, these intervals are defined as follows:

$$I = [a, b] \longleftrightarrow I(t) \longleftrightarrow c + r(2t - 1) \quad \text{if} \quad 0 \leq t \leq 1$$
$$\text{where} \quad c = \frac{a+b}{2} \quad \text{and} \quad r = \frac{b-a}{2} \tag{2.64}$$

By means of a vector of m weights $p = (\pi_1, \cdots, \pi_l, \cdots, \pi_m)$, and solving the minimization problem of the barycentric histogram, Given p histogram variables for the description of i and j, it is possible to express a multivariate version of 2.53:

$$d_w^2(X_i, X_j) \longleftrightarrow \sum_{k=1}^{p} \sum_{l=1}^{m_k} \pi_1^{(k)} \left[\left(c_{li}^{(k)} - c_{lj}^{(k)} \right)^2 + \frac{1}{3} \left((r_{li}^{(k)} - r_{lj}^{(k)} \right)^2 \right] \tag{2.65}$$

From this, is possible to figure out a second property of this distance embedded in this approach. Starting from a barycenter histogram X_b able to describe n histogram data, is possible to express a measure of inertia of data by means of the measure d_w^2. The total inertia, let's say (TI), with respect a barycentric description X_b of a set of n histogram data, is given by :

$$TI = \sum_{i=1}^{n} d_w^2(X_i, X_b) \tag{2.66}$$

Total Inertia, as a kind of deviation from the barycenter histogram X_b, may be assumed to be a measure of variation of histogram around their "center". In case of clusters, it can be decomposed as a sum of within cluster Inertia, WI, and between cluster inertia, BI:

$$TI = WI + BI \leftrightarrow \sum_{i=1}^{k} \sum_{i \in C_h 1} d_w^2(X_i, X_{b_h}) + \sum_{h=1}^{k} |C_h| d_w^2(X_{b_h}, X_b) \tag{2.67}$$

In this formulation, $|C_h|$ is the cardinality of cluster C_h. As concluding remark, while performing a hierarchical clustering agglomerative procedure, in order to pass from n to $n1$ clusters, the two clusters corresponding to the minimum d_{Ward} (Ward, 1963) are joined:

$$TI(C_s \cup C_t) = TI(C_s) + TI(C_s) + \frac{|C_s||C_t|}{|C_s| + |C_t|} d_w^2(X_{b_s}, X_{b_t}) \tag{2.68}$$

$$d_{Ward}(C_s, C_t) = \frac{|C_s||C_t|}{|C_s| + |C_t|} d_w^2(X_{b_s}, X_{b_t}) \tag{2.69}$$

The implementation of such procedure, including the d_w^2 up to the visualization of the dendrogram, that is a consolidated way to visualize the result of a hirarchical clustering analysis (Langfelder, B. Zhang, and Horvath, 2007), is present in the R package **HistDAWass** (Irpino, 2018). The following is an example of the application of such analysis on simulated data. After generating 1000 random observations from 10 different random variables, they have been transformed in 10 histogram-valued data using their empirical bins and p_i, as well as their empirical mean and standard deviation. First and second histograms are generated from a random variable with $mean = 2$, X_3 and X_4 with $mean = 4$, and histograms from X_6 up to X_{10} are created from a random variable with $mean = 10$. Even with slightly different variation around such mean, we expect that histograms generated from random variables with same mean are more likely to be clustered together in the early steps of the procedure. Empirical descriptions are summed up 2.2. In the

Table 2.2 10 different histogram-valued data described by their empirical means and their empirical standard deviations. Simulated data.

Histograms	Mean	S.D.
X˙1	1.890	2.021
X˙2	2.172	3.113
X˙3	5.566	2.035
X˙4	4.048	2.023
X˙5	3.979	2.971
X˙6	6.981	3.032
X˙7	7.045	3.869
X˙8	7.003	4.105
X˙9	7.165	4.097
X˙10	7.279	4.082

beginning, each histogram is considered as a cluster by itself, and then, in agglomerative fashion, they are merged together at different height, according to their similarity measured by means of d_w^2. Such eight is the calculated distance between clusters, and histograms X_7 and X_9 are merged in the 1st step cause are the closest ones, with a distance (height) of 0.24.

2.2.2 K-means clustering for histogram-valued data

The aim of clustering is basically to find structure in data and is therefore exploratory in nature. One of the most popular and simple clustering algorithms, K-means, was first conceived in 1957, and then published later on (S. Lloyd, 1982). *K-means* is still widely used in many scientific fields, even if from its creation thousands of clustering algorithms have been developed as well. It is build as a method of vector quantization, originally from signal processing. K-means clustering aims to perform a partition of n observations into k clusters. In each of the k cluster, n_k observations belong to the cluster with the nearest mean k, that is a centroid of such cluster. Data space is so divided into k "cells", analogously to what happens in a Vonoroi diagram representation (Reddy, Jana, and Member, 2012) of a partitioned region. The standard implementation of K-means clustering is an iterative process named Lloyds algorithm (Drake and Hamerly, 2012). In the following, it will be introduced a procedure to perform a dynamic clustering of histogram-valued data introduced by (Irpino, Verde, and DeCarvalho, 2014) and implemented in the R package **HistDAWass** (Irpino, 2018).
This method is build on the concept of Dynamic Clustering (DC) (Diday, 1971; Diday and Simon, 1976). DC, as other clustering methods, needs first

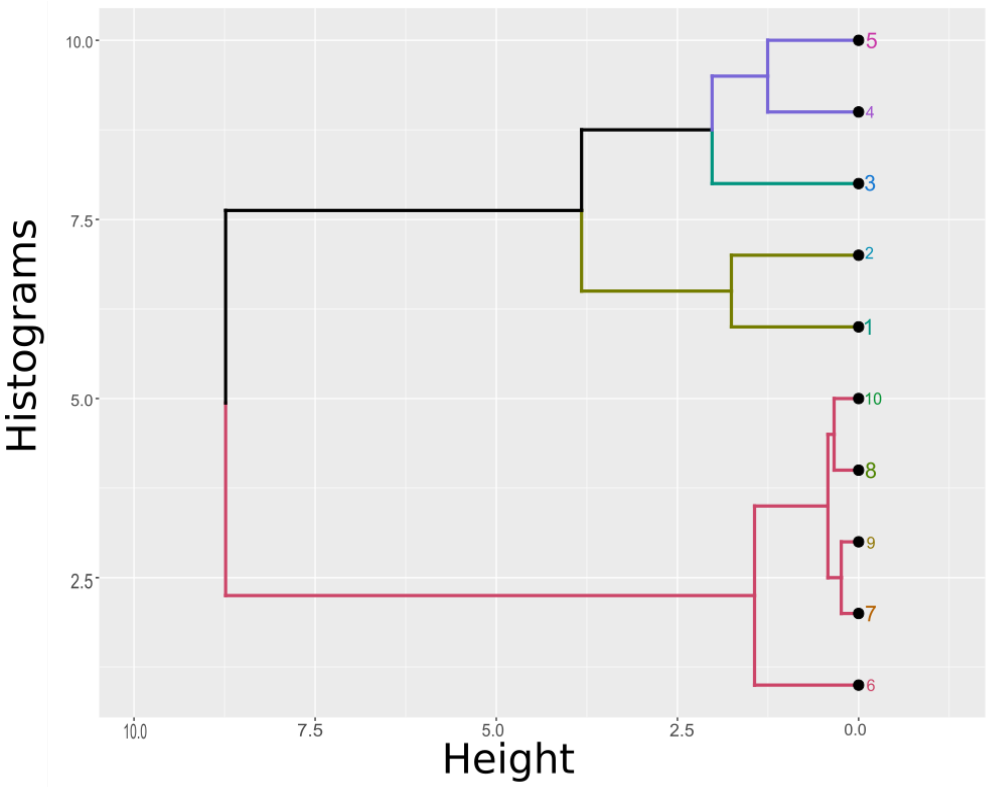


Figure 2.2: Dendrogram of hierarchical agglomerative cluster on 10 histograms. Different colors indicate different level of agglomeration at different height.

of all to define a proximity-similarity function, in order to assign the units to the clusters. Then, it has to be choosen a proper way to sum up the information contained in the units belonging to the same group; in other words, to identify each cluster with an individual inside it that optimizes a given criterion function. It is common to call that unit a *prototype*. For what concern the issue related to the choose of a proximity function, the use of standard distances allows to find spherical groups that share same size of variability, but one great advantage of clustering using adaptive distances is the possibility of identifying clusters that have a different level of variation and does not constrain them to have the exactly same orientation in the space (as directions of variables). The algorithm is suited to deal, at the same time, with both the best partition into k clusters and their best representation in terms of prototypes.
Authors in (Irpino, Verde, and DeCarvalho, 2014) start in their procedure defining, as distance between histograms and clusters, the standard d_w^2 (squared)

Wasserstein distance between the histogram x_i and the prototype g_k:

$$d(X_i, g_k) = \sum_{j=1}^{p} d_w^2(x_{ij}, g_{kj}) \tag{2.70}$$

In this case, no system of weights is defined and the general criterion to be satisfied is the minimization problem:

$$\Delta(\mathbf{G}, P) = \sum_{k=1}^{K} \sum_{i \in C_k} d_w^2(X_i, G_k) \tag{2.71}$$

where all the prototypes of the cluster $C_k : (k = 1, \cdots, k)$ are contained in the vector $G_k = (g_{k1}, \cdots, g_{kp})$. Further , $P = (C_{1,\cdots,C_k})$ is the partition to be found by means of the DC, and its corresponding set of prototypes is $\mathbf{G} = (G_1, \cdots, G_k)$.
Further, authors propose a modification of (2.71) implementing two vectors of coefficients that assign weights for each component of each variables (the mean and the dispersion). Such vectors are: $\Lambda_{mean_x} = \left(\lambda\frac{1}{x}, \cdots, \lambda\frac{p}{x}\right)$ and $\Lambda_{Disp} = \left(\lambda^1_{Disp}, \cdots, \lambda^p_{Disp}\right)$. This leads to obtain a 2^{nd} distance measure, the so-called *Globally Component-wise Adaptive Wassertein Distance* (GC-AWD):

$$\Delta(\mathbf{G}, P, \Lambda) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \lambda^j_{mean_x} (\bar{x}_{ij} - \bar{x}_{gkj})^2 + \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \lambda^j_{Disp} d_w^2(X^c_{ij}, g^c_{kj}) \tag{2.72}$$

The last distance that has been proposed, the 3^{rd} one, is given by the *Cluster Dependent Component-wise Adaptive Wassertein Distance*(CDC-AWD). It is:

$$\Delta(\mathbf{G}, P, \Lambda) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \lambda^j_{(k,mean_x)} (\bar{x}_{ij} - \bar{x}_{gkj})^2 + \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \lambda^j_{(k,Disp)} d_w^2(X^c_{ij}, g^c_{kj}) \tag{2.73}$$

where the two new vectors to be applied to mean value and dispersion are $\Lambda_{(k,mean_x)} = \left(\lambda\frac{1}{(k,mean_x)}, \cdots, \lambda\frac{p}{(k,mean_x)}\right)$ and $\Lambda_{Disp} = \left(\lambda^1_{(k,Disp)}, \cdots, \lambda^p_{(k,Disp)}\right)$. The starting solution, at step zero, is $(\mathbf{G}^0, \Lambda^0, P)$, and then the dynamic clustering algorithm, based on one of the 3 different schemas of adaptive distances (2.71, 2.72, 2.73), alternates

the solution to 3 different criterion. In the first two steps, the algorithms give the solution G for the best prototype (to best represent each cluster) of all the cluster, as well as the solution for the best adaptive distance (locally for each cluster) identified by Λ. In the last step, the algorithm gives the solution for the best partition P. When a step does not change in a significant way the solution already found in the previous step, the algorithm stops finding a stationary point that represents a local minimum in terms of within cluster sum of squares. To assess the quality of partition and clustering in terms of within homogeneity of clusters and external heterogeneity, holding the decomposition of inertia for the Wasserstein distance (2.66) a measure called QPI (Quality Partition Index):

$$QPI = 1 - \frac{WSS}{TSS} \leftrightarrow \frac{BSS}{TSS} \tag{2.74}$$

Given the simulated histograms presented earlier and their statistics in 2.2, the k-means algorithm was performed with different number of cluster ($k = 1, k = 2, \cdots, k = 9$). Then QPI was calculated for each clustering partition. Given that the starting solution $(\mathbf{G}^0, \Lambda^0, P)$ is going to influence the final result (that is still theoretically allowed to be only a local minimum and not a global optimum), 100 repetition of the algorithm are performed for each k and the best solution in terms of QPI is kept. A good number of clusters according to 2.3 is 4. For a review about how to compare two partitions, the remind is to **??**.

2.2.3 Fuzzy k-means clustering for histogram-valued data

A Fuzzy k-means clustering analysis (Dunn, 1973) is part of the clustering family called *soft partition*. It allows an object to be part of a cluster but not in a deterministic way. So it does not strictly belong to it, but every object belongs to a cluster in a determined degree, that represents an uncertainty level in the units allocation. On the other hand, *hard partition* (like k-means) is a crisp clustering, in the sense that objects belong or not to a cluster without uncertainty (or probabilistic evaluation) of such results. More specifically, in soft clustering, divisions allow statistical units to belong to multiple clusters, and does not force an object to participate in only one cluster or even construct hierarchical trees on group relationships. The original algorithm was developed in Bezdek, Ehrlich, and Full, 1984, using the Euclidean, Diagonal, or Mahalonobis distance, and then implemented for other distances as well (Hathaway and Bezdek, 1994). The clustering criterion used to aggregate subsets is a generalized least-squares objective

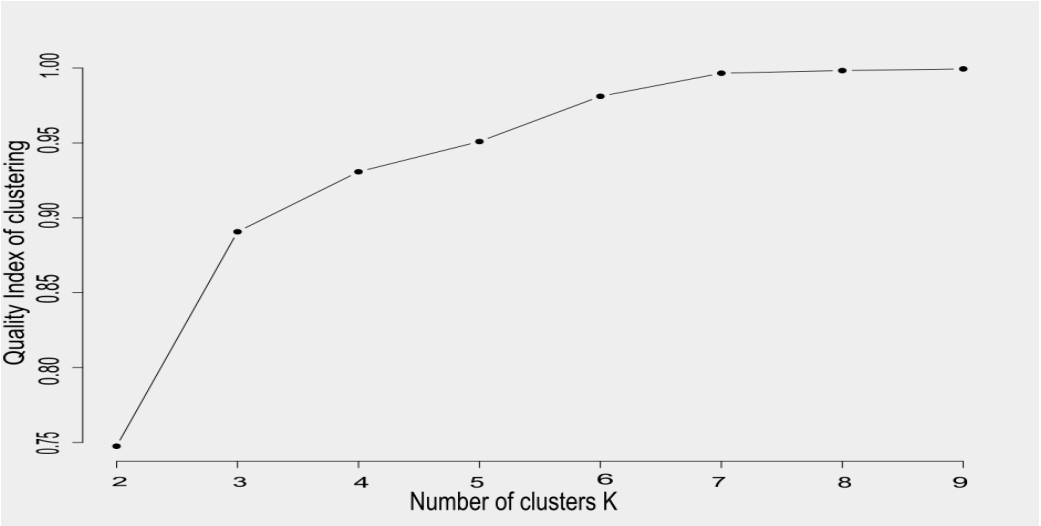


Figure 2.3: Quality Partition Index (QPI) for different number of cluster k, from 2 to 9. According to elbow rule, best number of cluster seems to be either 3 or 4.

function. The degree to which each units belongs to the different k clusters that has been identified, is usually called (*membership*) (Zadeh, 1968). This concept measures to what extent a unit is likely to belong a given cluster, and it is expressed with a number in $[0, 1]$. Although this, given that such number does not express a likelihood of belonging to a set, it is not a probability from a statistical point of view. Additionally, the sum of the memberships for each sample point has to be 1. From this, it is known that a crisp allocation is just a special case of fuzzy membership function. The fuzzy algorithm proposed by (Hathaway and Bezdek, 1994) follows the following steps:

1 For data matrix $\mathbf{X}$, initial values for k (number of clusters), m (weighting positive exponent), A (positive defined matrix of weights) are fixed and the adequate norm chosen. Then, algorithms starts with an initial matrix of fuzzy k-partition $U^{(0)}$.

2 Means of k-clusters are contained in a vector $v^{(t)}$ in each step with $t(1, \cdots, T_{max})$ with formula:

$$v_i = \frac{\sum_{T=1}^{T_{max}} (u_{it})^m x_t}{\sum_{T=1}^{T_{max}} (u_{it})^m} \tag{2.75}$$

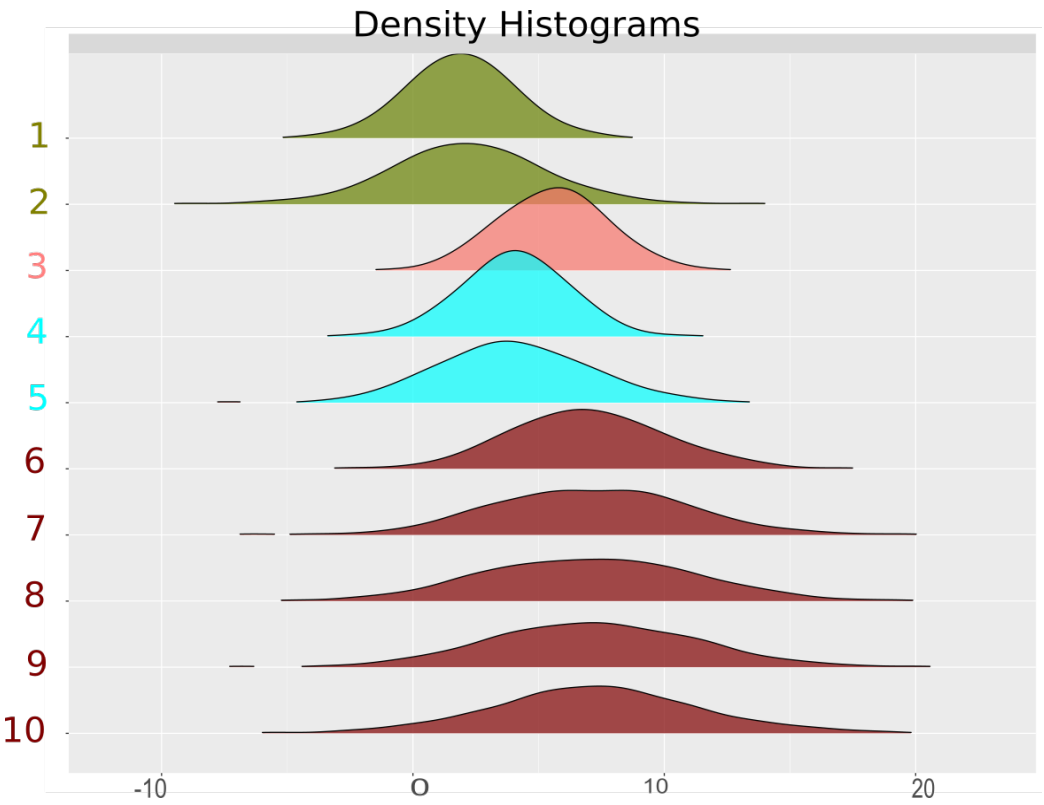


Figure 2.4: Histograms densities. According to the best partition with $k = 4$ clusters, different colors highlight different groups.

3 Analogously, it uses (2.75) to update U:

$$u_{it} = \left(\sum_{j=1}^{k} \left(\frac{d_{it}}{d_{jt}} \right)^{\frac{2}{m-1}} \right)^{-1} \tag{2.76}$$

4 If the distance between $u^{t+1} - u^t < \epsilon$, algorithm stops. If not, it keep updating solution starting from 2^{nd} step.

When dealing with distributional data, as histogram-valued data, a large part of the application that can be found in literature refers to image segmentation and classification and color image retrieval field (Küçüktunç, Güdükbay, and Ulusoy, 2010; Qing, Hua, Qiang, et al., 1992; Vertan and Boujemaa, 2000. This is mainly due to a known shortcoming of fuzzy analysis, that is an high computational cost (S. Chen and D. Zhang, 2004; Krinidis and Chatzis,

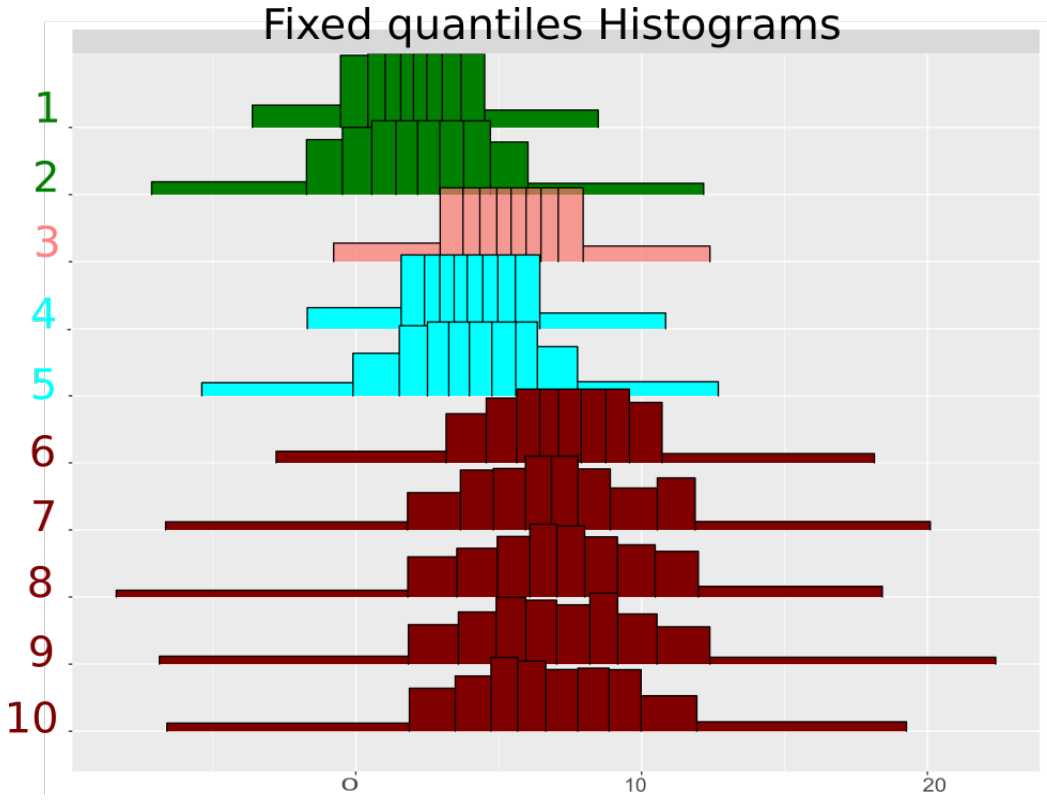


Figure 2.5: Histograms described by equal quantiles, in this case with 10 quantile (deciles). According to the best partition with $k = 4$ clusters, different colors highlight different groups. Same as in 2.4

2010). Therefore, in practical application (for example about colors spectrum), most time is efficient to reduce data-size grouping single observations in histogram-valued data, in an advantageous trade-off between computational cost and lost of information. Formally, steps performed in 2.2.3 still holds, but formalization about distances between histogram objects need to be added. It leads, so, to satisfy the three criterion $\Delta(\mathbf{G}, P, \Lambda)$, so best partition, best cluster and best weights, but with a partition P that is not binary, but allowed to present values $p_i[0 : 1]$. In the example of simulated data proposed in 2.2, membership values are presented in table 2.3.In each row is present one value that is very high (higher than 0.9), and it is consistent with the crisp k-means performed in previous section, leading to similar results but with a different approach.

Prototypes, as centroids of each cluster, are slightly different between the two different approaches, even if same "schema" of distances is adopted. It has to be taken under consideration that histograms are generated to

Table 2.3 Membership values, 4 clusters and 10 histograms. Each row sum to 1.

	Cluster1	Cluster2	Cluster3	Cluster4
1	0.000606	0.015945	*0.980850*	0.002600
2	0.000774	0.016783	*0.980077*	0.002366
3	0.000000	0.000001	0.000000	*0.999999*
4	0.001619	*0.957179*	0.011886	0.029316
5	0.002338	*0.970522*	0.011762	0.015378
6	*0.907619*	0.013766	0.002976	0.075639
7	*0.999525*	0.000120	0.000029	0.000326
8	*0.999205*	0.000208	0.000055	0.000532
9	*0.998396*	0.000427	0.000117	0.001060
10	*0.999378*	0.000158	0.000038	0.000426

Table 2.4 Prototypes (centroids) of the 4 clusters for the 10 histograms. They are just slightly different for crisp and fuzzy k-means

Prototypes	Crisp k-means	Fuzzy k-means
P1	[m= 2.0814 ,sd= 2.5716]	[m= 2.0796 ,sd= 2.572]
P2	[m= 6.9742 ,sd= 3.877]	[m= 6.9777 ,sd= 3.8517]
P3	[m= 3.9559 ,sd= 2.4928]	[m= 3.9568 ,sd= 2.4865]
P4	[m= 5.4486 ,sd= 2.0274]	[m= 5.4299 ,sd= 2.0109]

be rather easy to allocate in clusters. On real data, is pretty common to have objects that, for 2 or more clusters, present very close values of the membership function.

Chapter 3

Archetypes, prototypes and archetypoids in statistical learning

The term *archetype* has been widely adopted as a common word with the conceived meaning of "original pattern from which copies are made". It derives from the Latin noun "archetypum", latinisation of the Greek noun archetupon (with adjective form archtupos), which means "first-molded", in sense of "who was molded as first" (George et al., 1996). The term, nowadays, is used in a lot of different fields, and its meaning slightly change due to referring context. However, the basic concept is that an archetype is a pure-type: therefore, the term can be used to indicate a standard example, an ideal type, a symbol of perfection and so on. Indeed, the word is an union between the two terms arch, that means "beginning, origin, start" and tupos, which can mean, amongst other things, "pattern, model, type". Archetypes have been develop, likely, first in philosophy, where Platonic philosophical ideas referring to pure forms, archetypes, which embody the fundamental and essential characteristics of a concept or an element. Archetypes were then adopted in literary analysis, psychology and anthropology.
For what concerns statistics and categorization through statistical learning (Robert, 2005), human brain tends to build, in a process that is somehow instinctive but still based on the knowledge, complex relationships between complex items. These objects that can be described by several features, but humans tend to categorize them in no more then 4-5 categories (Cowan, 2010), and categories are stored in our long-term memory, and it has been proofed that we refer to the categories and recall them in the working memory, developing connections and bridges among them that improve our overall knowledge (Towse et al., 2008). The intimate relationship between

archetypes and categorization while learning, by the way, lies on the geometrical structure of concepts in a conceptual space framework (Gärdenfors, 2004) and the mathematical property and characteristics of archetypes.
From the point of view of the comparison betweem AA and other well-established techniques for dimensionality reduction and/or clustering, as pointed out by (Bauckhage and Thurau, 2009), one great advantage is the definition of archetypes as sparse mixture of data-points (and then each point is defined as convex combination of archetypes). This leads to overcome the shortcomings of some techniques such PCA and kernel PCA (Jolliffe, 2011; Schölkopf, Smola, and Müller, 1998) where the final components, as basis elements, lack of mathematical-physical meaning. Further, tecniques based on non-negative matrix factorization, (Finesso and Spreij, 2004; D. D. Lee and Seung, 1999) and its alternating least squares solution, leads to obtain characteristics parts, AA is able to produce archetypal that are composites. Given that the coefficient vectors of a data-points convex combination lays in a simple, AA is a technique that can provide subsequent probabilistic ranking, clustering (especially soft), and latent class models n order to classify units.
Application of AA on real data, and so on real problems, have of course already been promoted. It has been demonstrated as it can be an useful tool in benchmarking (Mittas, Karpenisi, and Angelis, 2014; Porzio, Ragozini, and Vistocco, 2006, 2008), how it can be used to analyse market segmentation (Elder and Pinnel, 2003; S. Li et al., 2003). Results were found using AA in spatio-temporal dynamics and cellular flames (Stone, 2002; Stone and Cutler, 1996). Studies were performed on cystic fibrosis airways (Thøgersen et al., 2013). In astronomy, it was used to explore and clusterize galaxy spectra (B. H. Chan, Mitchell, and Cram, 2003) and it was promoted as a tool for what concerns sensory analysis (D'Esposito, Palumbo, and Ragozini, 2011).
Following sections are organized as follows. In 3.1 the main formulation and definition of Archetypal Analysis will be introduced and discussed. In 3.2 elements of Prototypical analysis, derived from AA, will be explored. Archetypoids will be discussed in 3.3, while approaches of AA for Symbolic Data will be deepen in 3.4.1, especially for interval-valued data 3.4.1.

3.1 Archetypal Analysis (AA)

Archetypal analysis (Cutler and Breiman, 1994) is a method of unsupervised learning that aims to represent each object in a data set as a mixture of *individuals of pure type*, known as archetypes. These archetypes are build and defined to be a linear combination of data points (a "mixture of individ-

uals" James et al., 2013). Archetypes are selected in a minimization fashion, reaching optimum when the squared error is minimum in representing each individual as a mixture of archetypes.
Archetypes are defined as the p points contained in archetype matrix $\mathbf{Z}$ that satisfy:

$$x_i' = \alpha_i' \mathbf{Z} \tag{3.1}$$

The calculation and computation of such archetypes is a non-linear least squares problem, which is solved using an alternating minimizing algorithm. In the original formulation, given a $\mathbf{X}_{(n \times m)}$ data matrix, with so n individuals and m variables, let's define a group of p vectors $z_k = \sum_j \beta_{kj} x_j$ as a linear combination of original data. Further, constraints are introduced: $\beta_{kj} \geq 0$ and $\sum_j \beta_{ki} = 1$. Therefore, α_{ik} solutions of minimization problem, satisfies:

$$\sum_i \left\| x_i - \sum_{k=1}^{p} \alpha_{ik} z_k \right\|^2 \tag{3.2}$$

The minimum value reached by (3.2) is called Residual Sum of Squares RSS_p. Archetypes are useful in unsupervised learning also due to their location properties. Given a Convex Hull (CH) of original data points, in case of $p = 1$, so only one archetypes is identified, the sample mean is the solution to minimize RSS. If $1 < p < N$, all the z_k vectors of archetypes lie on the boundary of CH to minimize RSS. if number of archetypes is equal to N, $RSS = 0$. For all the proofs behind these results, the reference is (Cutler and Breiman, 1994). To introduce the algorithm to accomplish alternating least square, let's first discuss the constraints that define archetype problem.

- $\alpha_{ik} \geq 0$ and $\beta_{kj} \geq 0$
- $\sum_{i=1}^{p} \alpha_{ik} = 1$ and $\sum_{j=1}^{n} \beta_{kj} = 1$

The coefficients of the archetypes are *alpha*'s and the coefficients of the data set are *beta*'s. From this, minimization problem in (3.2), to find best *alpha*'s and best *beta*'s, can be written as:

$$RSS = \sum_{i=1}^{n} \left\| x_i - \sum_{k=1}^{p} \alpha_{ik} \sum_{j=1}^{n} \beta_{kj} x_j \right\|^2 \tag{3.3}$$

This could be solved using a general-purpose constrained non-linear least squares algorithm, but to overcome most computational issues, authors proposed an alternating constrained least squares algorithm. It solves first for

α's given the mixture $\beta_{kj}x_j$, that is a combination of original data-points. Then, it solves the mixture given the α's. In each step, algorithm solves a convex least squares problem.

1 Solve for α's $\rightarrow \left\|x_i - \sum_{k=1}^{p} \alpha_{ik} z_k\right\|^2$

2 Let's define $v_i = \dfrac{\left(x_i - \sum_{k \neq l}^{p} \alpha_{ik} z_k\right)}{\alpha_{il}}$ and the quantity $\bar{v} = \dfrac{\sum_{1=l}^{n} \alpha_{il}^2 v_i}{\sum_{1=l}^{n} \alpha_{il}^2}$. Minimization problem becomes now $RSS = \sum_{1=l}^{n} \alpha_{il}^2 \left\|v_i - z_l\right\|^2 \leftrightarrow \sum_{1=l}^{n} \alpha_{il}^2 \left\|v_i - \bar{v}\right\|^2 + \sum_{1=l}^{n} \alpha_{il}^2 \left\|v_i - z_l\right\|^2$.

3 In each step, α's and β's are updated, and algorithm keep solving iteratively for both. If improving in RSS is negligible, algorithm stops.

The proposal computational procedure to solve this minimization problem can be found also in (Lawson and Hanson, 1995), and it is a Constraint Non-Negative Least Squares. Other developments in the convergence criterion and in the problem formulation can be found in (Eugster and Leisch, 2009a), that developed the R package **archetypes**. Authors start from the concept of "approximation", and state the data are best approximated by convex combinations of the archetypes. From this, definition of RSS and the relative minimization problem is the following:

$$RSS = \left\|\mathbf{X} - \mathbf{A}\mathbf{Z}^T\right\|_2 \tag{3.4}$$

In this matrix notation, $\mathbf{X}$ is matrix with data points of dimension $(n \times m)$, $\mathbf{A}$ is the matrix with coefficients of archetypes of dimensions $(n \times k)$ and $\mathbf{Z}$ is the matrix containing the k dimensional archetypes for m variables. Analogously to Cutler and Breiman, 1994, constraints 3.1 apply here to $\mathbf{A}$ and $\mathbf{B}$, that is the matrix with data coefficients $(n \times k)$. Archetypes are defined as convex combinations of data point $\mathbf{Z} = \mathbf{X}^T\mathbf{B}$, and algorithm solves iteratively for $\mathbf{A}$ for given archetypes $\mathbf{Z}$, and vice-versa it finds best archetypes $\mathbf{Z}$ given coefficients $\mathbf{A}$. The algorithm they propose is build in the following steps:

1 Data initialization and preparation. Constraints 3.1 are applied and data are scaled.

2 Find best α for given $\mathbf{Z}$, solving n convex least squares problems: $\min_{\alpha_i} \dfrac{1}{2} \left\|\mathbf{X}_i - \mathbf{Z}\alpha_i\right\|_2$.

3 Recalculate archetypes $\mathbf{Z}^*$, solving linear equations $\mathbf{X} = \mathbf{A}\mathbf{Z}^{*T}$

4 Find best $\mathbf{B}$ for given $\mathbf{Z}^*$, that is equal to solve k convex least squares problems $\min_{\beta_j} \frac{1}{2} \left\| \mathbf{Z}_j^* - \mathbf{X}\beta_j \right\|_2$

5 Recalculate archetypes $\mathbf{Z} = \mathbf{XB}$ and RSS

6 Post processing phase, rescaling results (archetypes)

Authors, instead of a (Cutler and Breiman, 1994; Lawson and Hanson, 1995) Constraint Non-Negative Least Squares solution for $\mathbf{Z}^* = \mathbf{A}^{-1}\mathbf{X}$, have proposed the application of Moore-Penrose pseudoinverse, able to provide an approximation of the unique solution by a least square approach (Courrieu, 2008), and a QR decomposition with $\mathbf{A} = QR$ where R is an upper triangular matrix and Q is orthogonal. Lastly, calculation of RSS is made by means of spectral norm (Golub and Van Loan, 1996): $\|\mathbf{X}\|_{2spect.} = \sqrt{\lambda_{max}(\mathbf{X} * \mathbf{X})}$. As pointed out by (James et al., 2013, Page 554), while each part of the minimization function is solved as a convex problem, the problem overall is not convex, and this leads to obtain a local minimum as result of the convergence. As last remark, it is worth to note that is possible to write down in a more synthetic way in matrix notation the overall minimization problem to find the k archetypes having as objective function (in Frobenius norm):

$$\min_{\mathbf{A}_k, \mathbf{B}_k} RSS_k = \min_{\mathbf{A}_k, \mathbf{B}_k} \left\| \mathbf{X} - \mathbf{A}_k \mathbf{B}_k^T \mathbf{X} \right\|_F \tag{3.5}$$

3.2 Prototypes in statistical learning

The term *Prototype* was first defined, in a scientific fashion, by Rosch (Rosch, 1973) in the field of cognitive sciences. It was defined as exemplars being made to be *ideal*, in the sense that a prototype contains the most representative features and characteristics inside a certain category in a space. As general meaning, apart from cognitive sciences, it is used to indicate an early sample, model, that stands out for its representativeness and, therefore, other tend to imitate. It is a term used in a variety of contexts and fields, including semantics, design, electronics, programming, informatics and philosophy. In some design work-flow models, e.g., creating a prototype (this process is often known as materialization) is an intermediate step between the theoretical formalization and the evaluation of an idea. Other authors (Medin and Schaffer, 1978; Rocha, 1999) have pointed out as prototypes can be observed or unobserved (i.e. abstract, not physical) entities: not necessarily as real elements of the category. However, the degree or representativeness of a data-point for a given category can be measured using a proper distance measure from the prototype, taking into account the data nature (Timm

et al., 2004). From a statistical point of view, prototypes are considered useful in supervised and unsupervised learning framework to perform classification and clustering (Borgelt, 2006). Prototypes are considered so crucial cause they are able to captures peculiar characteristics of the data distribution (like location, size, and shape). Specifically, in recent years, their role have increased for what concerns cluster analysis (Ragozini, Palumbo, and D'Esposito, 2017) leading to a strand of prototype-based clustering method, given that the identification of a prototype allows to represent a cluster by a single data-point. A candidate, to be a prototype, has to satisfy an adequacy criterion to be chosen as most representative of its group. Further, it has been claimed that there is an inherent value of having a set of prototypical elements in data-points (Bien and Tibshirani, 2011a,b). Several procedure (and, so, several criteria) have been proposed to find a consistent approach to identify prototypes. According to (John Lu, 2010; Tan et al., 2007), it is common to accomplish such identification using a constant radius method, g.e. the k-means algorithm (2.70, 2.71, 2.72, 2.73), and the related moving center methods. In the following section 3.2.1 it will be introduced how prototypes will be, on the other hand, identified starting from Archetypal Analysis.

3.2.1 Prototype identification from Archetypal Analysis

The basic idea, that is the cornerstone of the prototypes identification from archetypal analysis, is that a prototype has to satisfy, in terms of maximination, the so-called *typicality-prototypicality degree* (Rifqi, 1996). This concept derives from prototypes definition in cognitive science (Rosch and B. B. Lloyd, 1978) for which *resemblance family* is the ground that is the origin of the construction of categories (Tversky, 1977). Further, is has been shown how the typicality-prototypicality degree degree combines two different complementary components (Lesot and Kruse, 2007): *internal resemblance* and *external dissimilarity*. The former starts from the assumption that each object shares different common features with other members of the category, but no globally shared feature can be identified (Lesot, Rifqi, and Bouchon-Meunier, 2008). The latter measures the total dissimilarity to objects of other classes. Therefore, the steps in order to perform the identification of prototypes in this approach, are 1 Compute the internal resemblance degree and the external dissimilarity. 2 Aggregate together both (internal resemblance and the external dissimilarity) to obtain an overall typicality degree. 3 Choose prototypes according to overall typicality degree and a predefined

threshold.
Analogously to these steps formalized as a general procedure that lies between clustering and categorization in cognitive sciences framework, the three steps procedure proposed in (Ragozini, Palumbo, and D'Esposito, 2017) is as follows:

Step 1

- Performing AA, first step is to maximize a proper external dissimilarity criterion. For this, archetypes act as first stage well-separated prototypes.

Step 2

- Data can be now clustered around each archetype, leading to a maximization of an internal resemblance criterion. From this, prototypes are identified. In such a step, prototypes are figured out in the space spanned by the archetypes, cause in such space compositional distances properties can be exploited (Aitchison, 1982; Aitchison et al., 2000).

Step 3

- From the space spanned by archetypes, prototypes are reverted to the original space of data-points, to get their final versions.

When aggregation operator is applied, it yields a compromise between these to instances of Step 1 and Step 2. In this case the partition is not known in advance, so AA is also able to identify the proper number of clusters (and, therefore, of prototypes). This means that the approach is data-driven, and in each step information from data are extracted incrementally, to obtain at the end of Step 3 a class of prototypes that maximize the interpretation and comprehension of data-points patterns and structure.
Formally, let's assume a set of n objects Ω and the possible partitions $\mathbf{C} = (C_1, \cdots, C_K)$ of Ω in K groups. If measures (ρ and δ) are defined for resemblance and dissimilarity (R and D), the prototipicality index T can be written as follows:

$$T(x, C_k) = f(R(\mathbf{x}, C_k), D(\mathbf{x}, C_k)) \tag{3.6}$$

where function f is a function able to combine together resemblance and dissimilarity. Resemblance $R(\mathbf{x}, C_k) = P(\rho(\mathbf{x}, \mathbf{x}_i))$ is a function that measures similarity between the object $\mathbf{x}$ with $\mathbf{x}_i \in C_k$. The second part of the equation (3.6) refers to the dissimilarity $D(\mathbf{x}, C_k) = \Delta(\delta(\mathbf{x}, \mathbf{x}_i))$. It is a way

to evaluate the dissimilarity between $\mathbf{x}$ and all the $\mathbf{x}_i \notin C_k$. From this, the set of prototypes $\mathcal{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_K)$ is then defined as:

$$\mathcal{P} = \{\mathbf{p}_k \in \Re^p | \mathbf{p}_k = \arg\max_{\mathbf{x}_i} T(x, C_k), k = 1, \ldots, K\}. \tag{3.7}$$

Let's denote prototypes in Step 1 $\mathcal{P}^1$, in Step 2 $\mathcal{P}^2$ and in Step 3 $\mathcal{P}^3$. As mentioned, the K prototypes in first Step $\mathcal{P}^1_K$ are equal to the K archetypes, let's say, $\mathcal{A}_K$. In this stage of the procedure prototypes are extreme points with respect to data cloud, they lie far from each other, characterize the data structure and are able to recover the global variability (Ragozini, Palumbo, and D'Esposito, 2017). Let's recall that the number K of prototypes/archetypesis still unknown but, in this data-driven approach, are assumed to be determined by the data cloud structure. Moving from 1st Step to 2nd Step, archetypes $\mathcal{A}_K$ are used as basis vectors to create a space that is spanned by them, as formulated in (3.1). In that space, each original point is represented by a combination of archetypes, using α's coefficients to reconstruct original data matrix. Here the space is a *K-dimensional simplex* where archetypes are the vertices. In this space, is possible to obtain a partition by means of a clustering procedure around such archetypes. As classifiers, is possible to adopt both a crisp or a fuzzy rule. The former is:

$$C_k = \{\mathbf{x}_i : \arg\max_j \alpha_{ij} = k\}, \qquad k = 1, \ldots, K. \tag{3.8}$$

While the latter is:

$$C_k^\tau = \{\mathbf{x}_i : \alpha_{ik} > \tau\}, \qquad 0 < \tau < 1, \;\; k = 1, \ldots, K. \tag{3.9}$$

In both cases, to maximize internal resemblance in each group of the partition, *centroids* of such partition are selected as prototypes in 2nd Step:

$$\min_{(\mathbf{c}_1, \ldots, \mathbf{c}_K)} \sum_{\mathbf{x}'_i \in C_k} d(\boldsymbol{\alpha}_i, \mathbf{c}_k) \forall k \tag{3.10}$$

with $d(\cdot, \cdot)$ an appropriate dissimilarity measure in the space $\mathcal{S}^K$. These centroids are $\mathcal{P}^2$. Last step is to revert $\mathcal{P}^2$ to the original space from the space spanned by archetypes, to obtain $\mathcal{P}^3$:

$$\mathcal{P}^3 = \mathbf{c}_k \mathbf{A}(k) \tag{3.11}$$

After this last step, the procedure is accomplished and $\mathcal{P}^3$ final version of prototypes are identified.
Further, starting from identification of $\mathcal{P}^3$, categorization is possible by means

of Voronoi Tessellation (Azrieli and Lehrer, 2007; Palumbo and Ragozini, n.d.). The approach is embedded in the Conceptual Space Theory (Gärdenfors, 2004). If is possible to assume that the metric inside conceptual space is the Euclidean one, the categories $c(\mathbf{p}_k)$, related to the partition, obtained through the 3 steps, correspond to cells derived from a Voronoi tessellation (Edelsbrunner and Seidel, 1986) based on the prototypes $\mathcal{P}^3$. Thus, the prototypes are able to identify categories in a thorough way, and categories are convex regions of the conceptual space as defined by Gardenfors. Is possible then to classify all the other points belonging to such the conceptual space.

3.3 Archetypoids

The concept of *Archetypoid* has been introduced in the statistical literature by (Guillermo Vinué, Epifanio, and Alemany, 2015). One of the biggest shortcoming of archetypes and prototypes as defined and calculated in previous sections, is that, even though they are able to hold many useful mathematical-geometrical properties (for example archetypes from AA are a convex combination of the sampled individuals), they are not necessarily observed individuals. And, in many real cases and in real data analysis, this can be a strong "contra". Indeed, in certain problems, it is crucial that the data are summarized by means of real subjects, that is, observations of the sample. From an interpretative point of view, if archetypes or prototypes are "artificial", no one individual can fit it 100%. Thus, to counter-face this problem, archetypoid are developed as real (observed) archetypes. Real applications where only real units are interesting and useful have been proposed in several academic publications, as well as in company and business field. For example, a robust version of archetypoids has been used to analyze hyperspectral imagery (Sun et al., 2017) and to evaluate and interpret sports performances (Guillermo Vinué and Epifanio, 2017). Several extensions of archetypoids have been developed for more complex data, such functional data archetypoids (Epifanio, 2016) and applied to financial time series (Moliner and Epifanio, 2018).
Formally, starting from definition of archetypes given in (3.1), (3.1) and (3.2), archetypoids are derived from such formulation just by adding a further constraint. It is that the results of AA z_k have to be a real data-points. This is a mixed-integer optimization problem to be solved. Authors in (Guillermo Vinué, Epifanio, and Alemany, 2015) developed an algorithm ad hoc to overcome the computational costs of two well-known algorithms to face this optimization problem: branch and bound algorithms and genetic algorithms. Another problem is that the results provided by the genetic algorithm did

not hold the constraints of the archetypoids. The true solution, however, can be found only with a combinatory approach, trying one by one all the data-points and searching for the one able to minimize the objective function. Still, with a real data large datasets, the computational cost is too high to consider this an efficient procedure. Authors decided so to develop an algorithm based on the Partitioning Around Medoids (PAM) clustering algorithm (Kaufman and Rousseeuw, 2009; Van der Laan, Pollard, and Bryan, 2003). Let's recall that the medoid is that real object of the cluster that is able to minimize the average dissimilarity to all the units of the cluster. The archetypoid algorithm is presented in 4 steps as follows:

1 BUILD phase. In this step is important to choose good initial units from the set of n data-points that act as first-step k archetypoids.

2 SWAP phase. Calculate RSS for both archetypoids and points that are not archetypoids, switching them iteratively. It is based on α's coefficients.

3 SELECT Select the configuration with lower RSS

4 REPEAT Repeat from step 2 to 4 until there is no change at all in archetypoids

5 END.

How to choose the initial archetypoids of 1^{st} step is, at this stage, the open question still on the table to start the algorithm. In the following, the most used (according to the literature) 5 criteria will be introduced:

- It is possible to initialise randomly with a simple sample random procedure (with no replacements allowed) of n units belonging to the original dataset to be the k starting archetypoids.

- Another approach is to compute the Euclidean distance (or a proper distance for such space) between the k archetypes from AA and the n individuals and choosing the nearest ones (as proposed in Epifanio, Vinué, and Alemany, 2013).

- Further, procedure can start identifying the individuals with the maximum α value for each archetype from AA, i.e. the data-point with the largest relative share for the respective archetype. It has been used in Eugster, 2012 and Seiler and Wohlrabe, 2013, given that α's represent how much each archetype contributes to the approximation of each individual.

- The fourth choice works, on the other hand, on β's instead of α's. It identifies archetypoids with maximum value of β, so it chooses units that are used the most in contributing in the generation of the archetypes.

- Last possibility consists of using *FURTHESTSUM* initialization (Mørup and Hansen, 2010, 2012), that is a way to select archetypes in a stepwise fashion.

The algorithm above presented is intended to act in a similar way as a PAM, with a BUILD and a SWAP phases. The idea behind the SWAP phase of the presented algorithm is similar to PAM analysis, and it yields to computational costs much higher than BUILD phase. What changes, from PAM to archetypoids identification, is the objective function that here is RSS. PAM, indeed, is suited to clustering around k central points (medoids) and archetypoids identification is aimed at finding k units able to describe, characterize and represent the extreme types in the data cloud. The SWAP phase, in this case, aims to improve the original chosen set of archetypoids at step 1 by exchanging iteratively selected and unselected individuals, checking if such replacements are able to reduce the RSS, so if the original data representation by means of combination of such extreme points is improved. In the inner loop, new α's are calculated in order to evalute new RSS, and if there is an improvement algorithm restarts at the end of step 1 using new archetypoids as initial values. Then, 2^{nd} phase is repeated until no changes occur in any of the archetypoids. Given that all the theoretically allowed swaps are considered, the final results is just a function of the new recalculation of RSS and α's, no matter the objects order in data. It is worth to note that the coefficients used to construct archetype from original data, β's, are not update in the same way as in archetypes 3.1. In this algorithm, β's are just binary: 1 for the individual chosen to be archetypoid, 0 for the others. The RSS is calculated using a spectral 2-norm Eugster and Leisch, 2009b, or with the Frobenius Norm. Archetypoids, given that are real observation and not artificial points, have different locations with respect to archetypes and prototypes. If the number of archetypoids is $K = 1$, it is the medoid of data cloud with one cluster (according to the Euclidean distance as dissimilarity). For what concerns archetypes, if $K = 1$ the solution is the sample mean. In case of $K = N$, the archetypoids are equal to the set of vertices of the convex hull of $\mathbf{X}$ and $RSS = 0$. If $1 < K < N$, it's not possible to state that the archetypoids lie on the boundary of the convex hull of $\mathbf{X}$. On the other hand, it is possible to state that for archetypes. Usually archetypoids lie on such boundary if they have a normal distribution, but overall it depends on the distribution of the observations.

3.4 Archetypes, Prototypes and Archetypoids for Complex data

Archetypal analysis was developed, in its original form, to deal with simple data-points. And, then, even prototypes and archetypoids have been originally formalized in the simplest case of a data matrix containing observations that are multivariate points. But, as discussed above, in SDA as well as in others approach, the nature of data could present a more complex structure than simple points. One example is about relational data embedded in Social Network Analysis approach (Scott, 2017). This approach is a statistical extension of mathematical graph theory (D. B. West et al., 2001) and it aims to describe patterns and structure behind relationships (usually called links or edges) between units (usually called vertices). The data nature is so more complex, cause it does not express a simple value but is a *relational data*, representing the degree and the direction of such relationship. In Ragozini and D'Esposito, 2015 authors use the archetypal analysis to analyze a group of networks, with the aim of classifying them (and to summarise them) by using a small number of networks from the original 36. The aim is to find a small number of representative networks that can be used as a benchmark for the other networks, as well as an useful tool to condense the most important information and features of the data set. For what concerns Social Networks Analysis, in (Ragozini, De Stefano, and D'Esposito, 2017) a 3 steps procedure has been developed to figure out prototypes of networks. In first step authors describe a network through a mixture of features referring to different scale network structures, then they find prototypes in the space spanned by such features, and lastly by reverting to the original networks space they figure out the final prototypes. For the case of functional data (Ramsay, 2005), where functional data are such data for which each observation is a whole function, archetipal analysis have been first proposed in (Costantini et al., 2012). Functions, in this work, were expressed in a functional basis, and the standard multivariate procedure to find archetypes was applied to the coefficients in this orthonormal basis. Analogously to what happens in Functional PCA (Manteiga and Vieu, 2007), this method holds only in case of orthonormal vectors as basis. On the other hand, in (Epifanio, 2016) the proposed methodology is developed to figure out functional archetypes and archetypoids without the orthonormal constraints, and it is valid whatever the basis used for approximating the functions.
In this section, a brief literature review about archetypes, prototypes and archetypoids for complex data have been presented. In the following, it will be deepen how these statistical learning techniques apply to interval data,

due to the clear link between interval-data and histogram-data.

3.4.1 Archetypes and Prototypes for interval-valued data

The formalization of interval data has already been presented in 2.4, defining briefly what interval data represent in a Symbolic Data Analysis approach. The first comprehensive wide formalization of Interval Algebra is considered Moore, 1962, and another crucial review about Interval arithmetic is for sure Kearfott, 1996, while Miller and Yang, 1997 focuses specifically on association rules for interval data. Let's first of all introduce, for the purpose of describe different proposal of statistical learning suited for interval data, the four elementary operations rationale when dealing with a set of intervals. Let's assume two interval objects, x and y, both defined between a lower and an upper bound: $x \to [x_l; x_u]$ and $y \to [y_l; y_u]$. The four elementary operations work in the following way:

$$x + y = [x_l + y_l; x_u + y_u] \tag{3.12}$$

$$x - y = [x_l - y_l; x_u - y_u] \tag{3.13}$$

$$x \times y = \left[\min(x_l, y_l); \max(x_u, y_u)\right] \tag{3.14}$$

$$x \div y = x \times \frac{1}{y} \leftrightarrow [x_l, x_u] \cdot \left[1/, y_u; 1/, y_l\right] \tag{3.15}$$

Further, given a number n of intervals $[x_{il}; x_{iu}]$, with $i = 1, \cdots, n$, the mean between x_i intervals is computed as follows:

$$\left[\frac{1}{n}\sum_{i=1}^{n} x_{il}; \frac{1}{n}\sum_{i=1}^{n} x_{iu}\right] \tag{3.16}$$

From these concepts, a proper distance between two intervals x and y:

$$q : (x, y) \in \mathbb{R} \times \mathbb{R} \leftrightarrow q(x, y) \in \mathbb{R}_0^+ \tag{3.17}$$

as defined in Corsaro and Marino, 2010 is:

$$q(x, y) = \sup\left[|x_l - y_l|, |x_u - y_u|\right] = |x_c - y_c| + |\Delta x - \Delta y| \tag{3.18}$$

where each interval valued data x is described as an interval of real numbers such that:

$$x = [x_l, x_u] = [x_c - \Delta x, x_c + \Delta x] \tag{3.19}$$

Let $\mathbf{A}$ be a matrix such that:

$$\mathbf{A} = [A_l, A_u] = [A \in \mathbb{R}^{m \times n}] \quad with \quad A_l \leq A \leq A_u \tag{3.20}$$

with A_l and A_u two rectangular matrices $\in \mathbb{R}^{m \times n}$. The defined $\mathbf{A}$, that is a set of matrices, is called *interval matrix* of dimension $\mathbb{R}^{m \times n}$. Further, two crucial elements of interval matrix need to be introduced in order to describe, likewise in (3.19), interval objects. Let's define *center* as:

$$A_c = \frac{1}{2}(A_l + A_u) \tag{3.21}$$

and the *radius* of an interval matrix as:

$$\Delta A = \frac{1}{2}(A_u - A_l) \tag{3.22}$$

Therefore, matrix $\mathbf{A}$ can be expressed as a function of both center and radius as follows:

$$\mathbf{A} = [A_l, A_u] = [A_c - \Delta A, A_c + \Delta A] \tag{3.23}$$

In D'Esposito, Palumbo, and Ragozini, 2006, a single interval matrix $\mathbf{X}$ is formally proposed as composed by two different matrices: $\mathbf{X}^c$ that contains all the *midpoints* (center matrix) and $\mathbf{X}^r$ including all the *ranges* (range matrix). Further, authors point out in this work the geometrical properties of such interval matrices, according to the number of columns (variables) included in the matrix. In one dimension cartesian space, each interval observations is a segment; in two-dimensions, it is a rectangle; in three-dimensions, it is a parallelepiped. When number of variables for each interval unit is higher than 3, each unit is configurable as a parallelotope.
However, algebraic operations between interval matrices such $\mathbf{A}$, are formalized analogously to operations with matrices containing simple single-valued data. The pointwise algebraic operations follow the rules introduced in ((3.12), (3.13), (3.14) and (3.15)). A *distance matrix* between two interval matrices, let's say $\mathbf{X}$ and $\mathbf{Y}$, is the non-negative matrix representing the pointwise distance as in (3.17) between all the elements (i, j) in $\mathbf{X}$ and $\mathbf{Y}$:

$$q(\mathbf{X}, \mathbf{Y}) \rightarrow q(\mathbf{X_{ij}}, \mathbf{Y_{ij}}) \tag{3.24}$$

The extension of archetypal analysis to interval-valued data starts from a similar non-convex minimization problem as in (3.5), but when dealing with interval matrices such as $\mathbf{X}$ and $\mathbf{Y}$, is possible to formalize them using center as in (3.21) and radius as in (3.22). According to (3.24), indeed, is possible to obtain a metric on the set $\mathbb{R}^{m \times n}$ using center and radius:

$$||q(\mathbf{X}, \mathbf{Y})||_F = || \, |X_c - Y_c| + |\Delta X - \Delta Y| \, ||_F \tag{3.25}$$

From this Frobenius Norm is possible to formalize the objective function in Interval Archetypal Analysis (IAA). The advantage of this metric is that is able to handle both the distance between the centers, for the aim of localization, and the radius of intervals, that are a way to summarize the accuracy. The IAA problem, for the interval matrices $\mathbf{X}$ and $\mathbf{Y}$, and with A and B the matrices including the α's and β's, is therefore as follows:

$$\min_{A,B} ||q(\mathbf{X}, A \cdot B \cdot \mathbf{X})||_F = || \, |X_c - (A \cdot B \cdot \mathbf{X})_c| + |\Delta X - \Delta(A \cdot B \cdot \mathbf{X})| \, ||_F \quad (3.26)$$

Under the usual constraints about α's and β's in (3.1), archetypes for IIA can be written as:

$$\mathbf{Z} = B \cdot \mathbf{X} \quad (3.27)$$

Further, as proposed in ibid., another approach is to relax the formula in (3.27), to split the matrix $\mathbf{Z}$ into $\mathbf{Z}^c$ and $\mathbf{Z}^r$. The former is an archetypal midpoints matrix, the latter is an archetypal ranges matrix:

$$Z^c = B^c \cdot X^c \qquad \Delta Z = B^r \cdot \Delta X \quad (3.28)$$

In this approach, thus, centers of the final interval archetypes lie on the convex hull of the centers, and radii of the final interval archetypes belong to convex hull of the radii. So, B^c and B^r satisfy independently 3.1. Overall archetypes don't lie, considering at the same time center and range, on the convex hull of interval data units. Let's underline that α's are the same for both matrices of centers and radii. To solve the minimization problem, authors in this work have proposed a modification of the Hausdorff distance (Rockafellar and Wets, 2009, Page 117), that is a norm between pairs of closed sets, suited for interval data, consistent with results found in (Neumaier, 1990; Palumbo and Irpino, 2005).

3.5 On the use of AA as benchmarking tool

Benchmarking plays a relevant role in performance analysis, and it a common practice that is adopted in several domains. Important reviews and discussion about benchmarking practices in general, with a wide overview on real data applications, can be found in Camp, 1989; Spendolini, 1992; J. Zhu, 2014. Most of the time benchmarking analysis make use of several statistical techniques and methods. This happens, of course, when quantitative benchmarking is the framework in which performance analysis is carried out. There are, on the other hand, contexts in which only a qualitative benchmarking is performed, without considering the additional contribution

of quantitative analysis. However, the standard definition of benchmarking is that it is a measurement, in a qualitative and/or quantitative perspective, of the quality of an organization's policies, products, programs, strategies. Special emphasis is on the comparison with standard measurements (best or worst performances), or similar measurements of its peers (similar companies). Further, is important to assess how close are the units to such good/bad standards, in order to evaluate the state of art of the overall performances, and to figure out if some specific groups are under-performing. In the quantitative framework, a simple method to benchmarking is gap analysis suited for single measures (single-valued data points), that is comparing performances through both analytical and graphical tools. In management and business literature, gap analysis is about the comparison of real estimated performance with potential or desired performance, usually linking the outcome to several input resources in a regression-wise fashion (S. W. Brown and Swartz, 1989). Therefore, if an organization does not make the best use of current resources in terms of efficiency, it may produce or perform below an idealized potential. Identifying gaps between the optimized allocation, from the theoretical point of view, of the inputs, and the real observed allocation-level of such resources, can reveal specific areas of underperformances. If indicators are needed starting from multivariate data, more sophisticated techniques are needed, in order to assess in a meaningful way the performance level of each unit, without leave out important features, but instead creating complex set of indicators highlighting several aspects of the performances as in Camp, 1995. In literature, it has been proposed to use multi-factor gap analysis (Eyrich, 1991) and the analytic hierarchy process (Saaty, 1990). Let recall that the former hypothesis the complete absence of relationships between different indicators (independent indicators assumption), and the latter performs the analysis based on subjective, and hardly measurable, opinions. In Smith, 1990 more sophisticated, complete and exhaustive and statistical techniques were used and discussed, mainly clustering analysis, multivariate regression and frontier analysis. When a clustering procedure is performed, homogeneous groups are determined based on several features, mainly in multivariate context, and so units belonging to the same cluster can be assumed to be "similar", and groups are heterogeneous if compared to each others. These concepts and properties of clusters are exploited in benchmarking, so that attainable targets can be defined (Binder, Clegg, and Egel-Hess, 2006; Koh, Gunasekaran, and Saad, 2005; Talluri, 2000). Further, most clustering analysis are able to identify abstract entities (centroids, mediods and so on) that are able to stress out a resume of the information contained in each group, allowing for simpler comparisons. Regression analysis has instead an other aim, that is trying to

explain in a causal fashion the link between inputs and the outcome, that is the performance; an application of multilevel regression model to evaluate educational performances can be found in Goldstein, Bonnet, and Rocher, 2007. In the recent years, also sentiment analysis has been exploited in order to make assessment about different domains. An application of several techniques of sentiment analysis applied to twitter data has been performed in Abbasi, Hassan, and Dhar, 2014

These are just some of the statistical techniques that have been proposed as tools for quantitative benchmarking, while a pretty recent debate has turned out about the role of archetypes in benchmarking area. First works focused to point out and discuss AA in such direction and for this use, have been Porzio, Ragozini, and Vistocco, 2006 and Porzio, Ragozini, and Vistocco, 2008. The aim of using AA in benchmarking is to adopt an exploratory and graphical approach, in order to consider AA as a basis for a data driven benchmarking procedure. Identified archetypes will play the role of reference performers, so that it will be possible to analyse their features, and real units' performances will be compared to the archetypes' ones. As discussed earlier, archetypes are extreme points with external location, and for this reasons they are suited to be reference abstract units in a benchmarking perspective, using as extreme reference the external part of data-cloud that lies on the convex hull.

The proposal, for the next section, is to apply AA in case of histogram data to evaluate and discuss some features of Italian School System performances.

Chapter 4

Archetypes for Histogram-valued data

When the data are histogram-valued data, we are taking under consideration a dataset made up by several observations, each of them being a univariate histogram for each variable. In this section the aim is, in the first instance, to develop the archetypal analysis for such kind of data starting from the Residual Sum of Squares as objective function, exploiting a proper formalization of distances based on histogram descriptions based on radii and centers, considering also previous formalized constraints about coefficients.

4.1 Formal definition of histogram-valued data archetypes

Let's denote with $\mathscr{X}$ a symbolic data table, with n observation for p histogram variables. In this context, let recall the general archetypes problem (as formalized in matrix notation in (3.5)) to be solved to find K archetypes:

$$\min_{\mathbf{A}_K, \mathbf{B}_K} RSS_K = \min_{\mathbf{A}_K, \mathbf{B}_K} \left\| \mathscr{X} - \mathbf{A}_K \mathbf{B}_K^T \mathscr{X} \right\|^2 \tag{4.1}$$

under the usual constraints about elements of matrices $\mathbf{A}_K$ and $\mathbf{B}_K$ analogously to 3.1:

- $\alpha_{ik} \geq 0$ and $\sum_{i=1}^{p} \alpha_{ik} = 1$
- $\beta_{kj} \geq 0$ and $\sum_{j=1}^{n} \beta_{kj} = 1$

where α_{ik} is the generic element of the matrix $\mathbf{A}$ and β_{kj} is the generic element of the matrix $\mathbf{B}$. They have the same role as in previous defined formulation: $\mathbf{A}$ contains all the elements that express the contribution of each

archetypes to represent each observation from original data, and $\mathbf{B}$ includes all the coefficients to represent individual contribution in the generation of the K archetypes from original data $\mathscr{X}$. At this stage, the issue is to find a proper function to measure the "distance" (in terms of dissimilarity or divergence) between the original matrix of histograms $\mathscr{X}$ and the "reconstructed" data matrix of histograms, let's say $\tilde{\mathscr{X}} = \mathbf{A}_K \mathbf{B}_K^T \mathscr{X}$, given so the estimation of $\mathbf{A}$ and $\mathbf{B}$.
One approach, to better deal with archetypal problem, is to give an operational description of histogram objects and their distances starting from centers and radii perspective. Let recall the definition of Wasserstein distance (Irpino and Verde, 2006; Rüschendorf, 2001), derived mainly from Gibbs' work (Gibbs and Su, 2002), as formalized in 2.52 and 2.53. Given a generic l_p Wasserstein distance (Givens, Shortt, et al., 1984), it expresses the distance between two observed densities ϕ_1 and ϕ_2 using the inverse of such density function, namely the quantile functions ϕ_1^{-1} and ϕ_2^{-1}:

$$\mathrm{W}_p = \left(\int_0^1 \left[\phi_1^{-1}(t) - \phi_2^{-1}(t) \right]^p dt \right)^{\frac{1}{p}} \tag{4.2}$$

Let recall that if $p = 2$, Wasserstein distance l_p becomes l_2 and its definition simplifies as:

$$\mathrm{W}_2 = \sqrt{\int_0^1 \left[\phi_1^{-1}(t) - \phi_2^{-1}(t) \right]^2 dt} \tag{4.3}$$

This equation (2.53) is a way to express the extension of L_2 Minkowski distance. Similar to this concept of distance between quantile functions, let define the Mallows Distance d_M (Mallows, 1972b) between two histogram objects $\mathbf{X}_1$ and $\mathbf{X}_2$, given a generic set of weights w, as follows:

$$d_M(\mathbf{X}_1, \mathbf{X}_2) = \sqrt{\int_0^1 \left[\psi_1^{-1}(w) - \psi_2^{-1}(w) \right]^2 dw} \tag{4.4}$$

In this context, ψ_1 is the distribution function of the first histogram object $\mathbf{X}_1$ and ψ_2 is the distribution function of the second histogram object $\mathbf{X}_2$. So, ψ^{-1} is the inverse of such distribution function, and as pointed out by (Irpino, Verde, and DeCarvalho, 2014) it can be a drawback for its high computational cost. As defined in 2.4, deriving from 2.3, histogram-valued data can be written as an extension of interval-valued data, cause each bin is assumed to be an interval variable, and the union of such bins made up an histogram itself. Let consider an histogram description of an histogram

variable $\mathbf{X}$, with H number of sub-intevals (bins) with observed empirical frequencies π_u such that:

$$\mathbf{X} = [I_1(\pi_1), I_2(\pi_2), \cdots, I_u(\pi_u), \cdots, I_H(\pi_H)] \tag{4.5}$$

From this, the weights w_u are defined as cumulative weights associated with w_u: $\sum_{u=1}^{H} \pi_u$. Each π_u is also defined as difference between weights, so each $\pi_u = w_u - w_{u-1}$. Under the assumption of uniform distribution in each interval I_u between lower and upper bounds, empirical distribution $\psi_u(\mathbf{X})$ can be written as:

$$\psi_u(\mathbf{X}) = w_u + (x - x_{Lu})\frac{w_u - w_{u-1}}{x_{Uu} - x_{Lu}} \tag{4.6}$$

where capital letters in the subscript of x indicates lower bound of u interval (x_L) and upper bound of u interval (x_U). Therefore, the inverse distribution function (quantile) can be expressed as a piecewise function:

$$\psi_t^{-1}(\mathbf{X}) = x_{Lu} + \frac{t - w_{u-1}}{w_u - w_{u-1}}(x_{Uu} - x_{Lu}) \tag{4.7}$$

After ordering ascendently weights, without repetitions, to identify a set of uniformly dense intervals (as proposed by ibid.):

$$\mathbf{w} = (w_0, \cdots, w_u, \cdots, w_m) \tag{4.8}$$

where weights are such that:

$$w_0 = 0, w_m = 1 \quad max(H_1, H_2) \leq m \leq (H_1 + H_2 - 1) \tag{4.9}$$

given that the objects to compare are $\mathbf{X}_1$ and $\mathbf{X}_2$, and their weights have been merged together in only one vector $\mathbf{w}$. From (4.8) and (4.9), the squared distance from 4.4 between histogram objects $\mathbf{X}_1$ and $\mathbf{X}_2$ becomes as follows:

$$d_M^2(\mathbf{X}_1, \mathbf{X}_2) = \sum_{u=1}^{m} \int_{w_{u-1}}^{w_u} \left[\psi_1^{-1}(t) - \psi_2^{-1}(t)\right]^2 dt \tag{4.10}$$

For each couple (w_{u-1}, w_u), it allows to identify two uniformly dense intervals, with respect to $\mathbf{X}_1$ and $\mathbf{X}_2$. These are defined inside the bounds $I_{u1} = [\psi_1^{-1}(w_{u-1}); \psi_1^{-1}(w_u)]$ for $\mathbf{X}_1$ and $I_{u2} = [\psi_2^{-1}(w_{u-1}); \psi_2^{-1}(w_u)]$ for $\mathbf{X}_2$. For each interval, the proposal is to obtain the computation of center and radius from the inverse of the distribution function. Centers, for an histogram object $\mathbf{X}_1$ and its sub-intervals, are defined as:

$$c_{u1} = \frac{(\psi_1^{-1}(w_{u-1})) + \psi_1^{-1}(w_u)}{2} \tag{4.11}$$

Radii, for an histogram object $\mathbf{X}_1$ and its sub-intervals, are defined as:

$$r_{u1} = \frac{(\psi_1^{-1}(w_{u-1})) - \psi_1^{-1}(w_u)}{2} \tag{4.12}$$

Again, under the assumption of uniformly distribution within each interval, it is possible to write intervals as a function of center and radius, exploiting the relationships between quantile function and centers-radii elements: $I_u = c_u + r_u(2t-1) \quad \text{for} \quad 0 \leq t \leq 1$. From this, Mallows' distance in 4.10 between $\mathbf{X}_1$ and $\mathbf{X}_2$ can be rewritten as:

$$d_M^2(\mathbf{X}_1, \mathbf{X}_2) = \sum_{u=1}^{m} \pi_u \int_0^1 \left[(c_{u1} + r_{u1}(2t-1)) - (c_{u2} + r_{u2}(2t-1))\right]^2 dt \tag{4.13}$$

it simplifies in a function of differences between centers and radii as follows:

$$d_M^2(\mathbf{X}_1, \mathbf{X}_2) = \sum_{u=1}^{m} \pi_u \left[(c_{u1} - c_{u2})^2 + \frac{1}{3}(r_{u1} - r_{u2})^2\right] \tag{4.14}$$

Equation in 4.14 is the univariate case of the multivariate general case with p variables, as in 2.65, already presented when discussing dynamic clustering:

$$d_M^2(\mathbf{X}_1, \mathbf{X}_2) = \sum_{j=1}^{p} \sum_{u=1}^{m} \pi_u^{(j)} \left[(c_{u1}^{(j)} - c_{u2}^{(j)})^2 + \frac{1}{3}(r_{u1}^{(j)} - r_{u2}^{(j)})^2\right] \tag{4.15}$$

It means that the overall distance (in terms of Mallows distance defined by centers and radii) between two multivariate histogram objects, is the sum of such Mallows difference computed for each variable, under the assumption of independence between variables. After this crucial assumption, it is possible to rewrite the general archetypes problem for the histogram symbolic data table $\mathscr{X}$, in a similar way as proposed in D'Esposito, Palumbo, and Ragozini, 2006, (Page 351). Let consider the matrix including the centers of $\mathscr{X}$, let say $\check{\mathscr{X}}$, and the matrix containing the radii of the symbolic data table $\mathscr{X}$, let say $\Delta\mathscr{X}$. Let assume that as mentioned, K is the total number of archetypes to be found, with $k \,\text{in}\, (1, \cdots, k, \cdots, K)$, the RSS_K problem as proposed in 4.1, using centers and radii notation and exploiting interval arithmetic properties from Mallows and Wasserstein perspective, becomes:

$$\sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{u=1}^{m} \pi_u \left[\left(c_{iju} - \sum_{k=1}^{K} \alpha_{c,ik} \beta_{c,ki}^{\intercal} \check{\mathscr{X}} \right)^2 + \frac{1}{3} \left(r_{iju} - \sum_{k=1}^{K} \alpha_{r,ik} \beta_{r,ki}^{\intercal} \Delta\mathscr{X} \right)^2 \right] \tag{4.16}$$

So, the overall minimization problem is to find best α's and best β's to minimize the RSS_K given the identified number K of archetypes, taking into account both the distance between centers and the distance between radii and the weights. General weights are related to the frequencies π_u, and specific weights for radii are the constant $\frac{1}{3}$. In both part of equation 4.16, coefficients α's and β's belonging to the matrices $\mathbf{A}_K$ and $\mathbf{B}_K$ are considered to be the same. This constraint implies the algebraic linkage between centers space and radii space. The overall minimization problem, given the vector $\mathbf{P}$ including all the π weights, and recalling that $\tilde{\mathscr{X}}$ is the symbolic data table containing the reconstruction of original histograms in the symbolic data table $\mathscr{X}$, is therefore defined as follows:

$$\begin{aligned}\min_{\mathbf{A}_K,\mathbf{B}_K} RSS_K &= \min_{\mathbf{A}_K,\mathbf{B}_K} \left\| \mathscr{X} - \mathbf{A}_K\mathbf{B}_K^T\mathscr{X} \right\|^2 \to \min_{\mathbf{A}_K,\mathbf{B}_K} \left\| d\left(\mathscr{X}, \tilde{\mathscr{X}}\right) \right\|^2 \\ &= \min_{\mathbf{A}_K,\mathbf{B}_K} \mathbf{P}\left[\left(\check{\mathscr{X}} - \mathbf{A}_K\mathbf{B}_K^T\check{\mathscr{X}} \right)^2 + \frac{1}{3}\left(\Delta\mathscr{X} - \mathbf{A}_K\mathbf{B}_K^T\Delta\mathscr{X} \right)^2 \right] \end{aligned} \quad (4.17)$$

As for simple data points as well for interval-value data, archetypes identification for histogram-valued data produces K archetypes that share the same nature of the original data, therefore in this case archetypes are proper histograms, with their bins and their quantile associated to each sub-interval, holding the same properties of classical histograms. Plus, other matrices to be found, $\mathbf{A}_K$ and $\mathbf{B}_K$, include coefficients that are single-valued, so these matrices are single-valued matrices. As already have been discussed in 3.1, archetypes are useful in finding interesting patterns in data structure above all for their location. For what concerns histogram archetypes, this point will discussed in the following section.

4.2 On the location of archetypes for Histogram-valued data

Let consider the simplest case in which AA is performed on single-valued data points. In this context, it is known that if only one archetypes is identified, it coincides with the sample mean of data. This result has been showed in Cutler and Breiman, 1994 and has been discussed in Eugster and Leisch, 2009b. On the other hand, the most common and useful case is when the number K of archetypes is $1 < K < N$, with N being the overall number of data points that fully define the boundary of the convex hull. In this case, archetypes lie on such convex hull, as in fig 4.1. When the number N

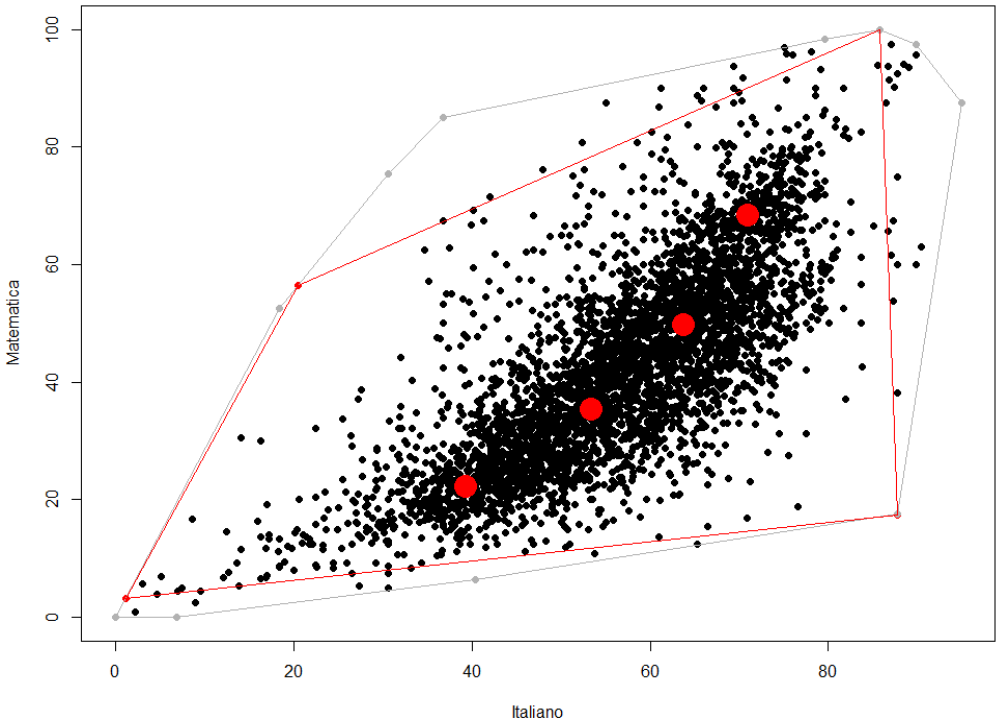


Figure 4.1: Location of 4 archetypes, extreme points lying on the convex hull, compared to 4 centroids of fuzzy k-means, that represent barycenter of their own group and so are located in more central positions of data cloud. Example Data from Invalsi Test 2015-2016 for Italian and Math domains.

of elements of the boundary of the convex hull are the same number of K archetypes, so when $K = N$, the objective function RSS decreases to its minimum, and at the end $RSS = 0$. The archetypes thus identified coincide with all the points that define the convex hull. So, usually a meaninful upper bound in the number of allowed archetypes is N. These theoretical results, as pointed out in Eugster and Leisch, 2009b and Bauckhage and Thurau, 2009 among others, are not always reached, according to some convergences issues and computational difficulties.
In case of histogram-valued data, a discussion about location can start from a similar problem of identifying archetypes in case of interval-valued data. In that case, in Corsaro and Marino, 2010 is recalled a result that has already been found in D'Esposito, Palumbo, and Ragozini, 2006 about relationships between convex hull in a centers-radii perspective. At the end, is has been

proofed that centers and radii of the archetypes respectively belong to convex hull of the centers and the radii of the data.
To get into the mechanism of the histogram-valued data archetypes identification in order to better discuss their location, let's introduce a real data example.
Data are retrieved from **HistDAWass**, and we consider 19 France regions with the distributions of marginal costs of farms. It contains two histogram variables: "Y˙TSC" (Total costs of a farm), "X˙Wheat" (Costs for Wheat). The matrix of distribution is, so, made up by 19 observations (rows) and 2 variables (columns), for a total of 38 histogram-valued objects. Further, regions can be divided into South, Center and North according to their position in France. Most important descriptive statistics for such histograms are presented in 4.1.

Table 4.1 Main descriptive statistics for histogram-valued objects with respect to 19 different France regions, about Farm and Wheat costs

	Mean	—	Median	—	Min	—	Max	—	St.Dev.	—	Skeweness	—	Kurtosis	
Region	Farm	Wheat	Farm	Wheat	Farm	Wheat	Farm	Wheat	Farm	Wheat	Farm	Wheat	Farm	Wheat
Ile-de-France	39206.90	40708.11	38579.09	40936.70	15101.78	0.00	66067.46	85712.46	13320.35	21232.91	0.15	0.10	2.02	2.20
Champagne-Ardenne	27809.90	16264.45	24375.91	13234.38	2298.01	0.00	71493.21	52334.79	19200.88	15807.18	0.52	0.58	2.20	2.09
Picardie	51505.53	36590.05	49513.34	32382.43	17590.67	8249.86	101260.37	78105.15	21225.37	19628.74	0.42	0.47	2.34	2.07
Haute-Normandie	34783.63	20430.56	31633.47	16595.58	11500.32	1993.09	72321.28	63881.53	16582.96	15094.71	0.45	0.97	2.11	3.24
Centre	32800.69	24312.89	29638.70	21555.85	9021.08	0.00	69930.23	65976.32	15824.75	17743.12	0.54	0.49	2.32	2.24
Basse-Normandie	30563.84	6205.59	27009.09	3407.05	7557.68	0.00	70557.57	28597.22	16707.12	7010.48	0.62	1.49	2.37	4.41
Bourgogne	29227.60	11502.43	25095.33	5581.26	6526.54	0.00	67128.45	47199.18	16045.79	13342.92	0.62	1.08	2.37	2.99
Nord-Pas-de-Calais	39700.86	19037.06	37133.65	16845.54	15989.00	3803.65	81271.21	45862.94	16879.63	10473.71	0.64	0.73	2.46	2.68
Lorraine	42035.00	18691.11	39705.14	16140.41	15372.60	1018.54	83318.66	52682.15	18629.02	13191.63	0.43	0.74	2.12	2.70
Alsace	21441.74	3804.42	17231.75	3340.37	3949.95	0.00	56199.29	13271.32	13490.40	3680.48	0.83	0.77	2.75	2.62
Franche-Comte	28852.49	3883.17	25616.55	2062.21	11294.17	0.00	64517.21	17692.13	12789.81	4458.26	0.95	1.38	3.17	4.07
Pays de la Loire	31911.27	5361.55	26009.39	4581.86	9794.71	0.00	86776.96	16225.09	19111.77	4453.67	1.14	0.63	3.47	2.44
Bretagne	40957.75	4463.18	28831.50	3689.93	10033.73	0.00	158739.07	14344.58	32832.51	3965.28	1.88	0.69	5.90	2.47
Poitou-Charentes	25139.94	7281.20	22925.15	5241.36	8397.97	0.00	53483.60	25965.88	11509.42	6850.35	0.65	0.95	2.52	3.00
Aquitaine	18764.01	319.01	15318.80	0.00	3959.19	0.00	48919.34	3474.00	11364.52	757.92	0.87	2.62	2.87	8.93
Midi-Pyrenees	16663.50	2362.43	14952.22	923.36	5897.12	0.00	37476.51	12583.55	7869.31	3163.08	0.74	1.52	2.75	4.45
Limousin	15286.13	543.76	14179.89	0.00	5833.37	0.00	31710.67	3319.51	6325.19	824.62	0.71	1.68	2.76	4.98
Rhone-Alpes	16573.55	1350.32	14006.74	0.00	4329.28	0.00	44556.14	7913.45	10040.80	2032.99	1.00	1.59	3.17	4.56
Auvergne	16803.60	1640.82	14814.14	471.58	5689.51	0.00	40160.15	11047.90	8512.10	2567.20	0.92	1.98	3.06	6.21

Density approximations of such distribution are presented in 4.2. From such visual displays it is clear that some of the regions present heavy tails; i.e. Bretagne for Farm costs and Champagne-Ardenne for Wheat costs. One of the most important aim of the AA is to find a way to sum up these "extreme behaviour" by means of few salient abstract entities, that are precisely the

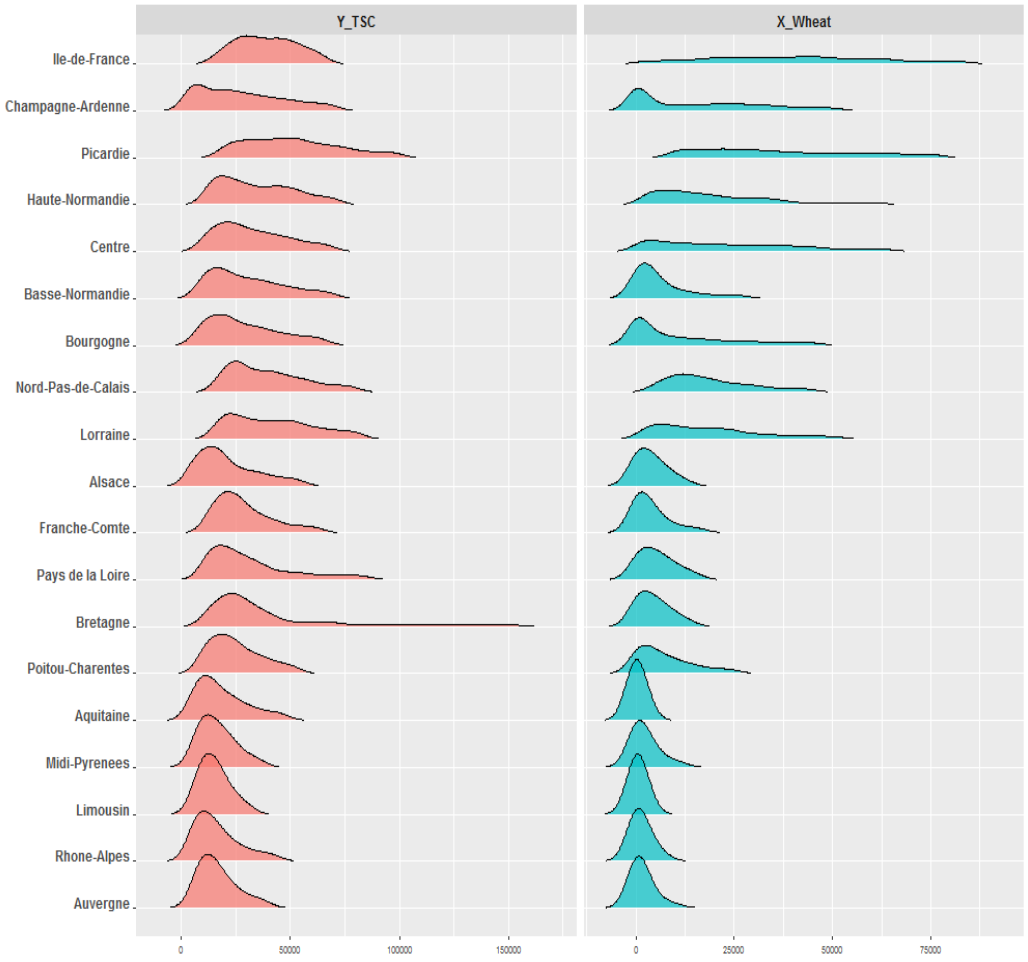


Figure 4.2: Density approximation of Farm and Wheat cost for 19 France Regions.

archetypes. First step, in performing AA, is to decide how many archetypes have to be found, to make AA advantageous in terms of statistical learning. As said, RSS is the objective function of the AA problem, so it is usually used, in relationship with the number K of archetypes, as a benchmarking function to take the decision about the proper number of archetypes to include in the analysis. One of the most common rule is, simply, the elbow method rule, used often in many clustering procedures to have a good proxy of the best choice of number of clusters. The idea is that, when from the graphic displays there is a negligible gain in terms of objective function RSS adding one more archetypes to the previous solution, it is not worthy to add such additional archetype to the final solution. So, K will be decided with

respect to the last significant gain in RSS as visualized in 4.3; it seems that when $K = 3$ the best trade-off between number of entities (archetypes) and quality of representation (RSS) is achieved. Archetypes are identified us-

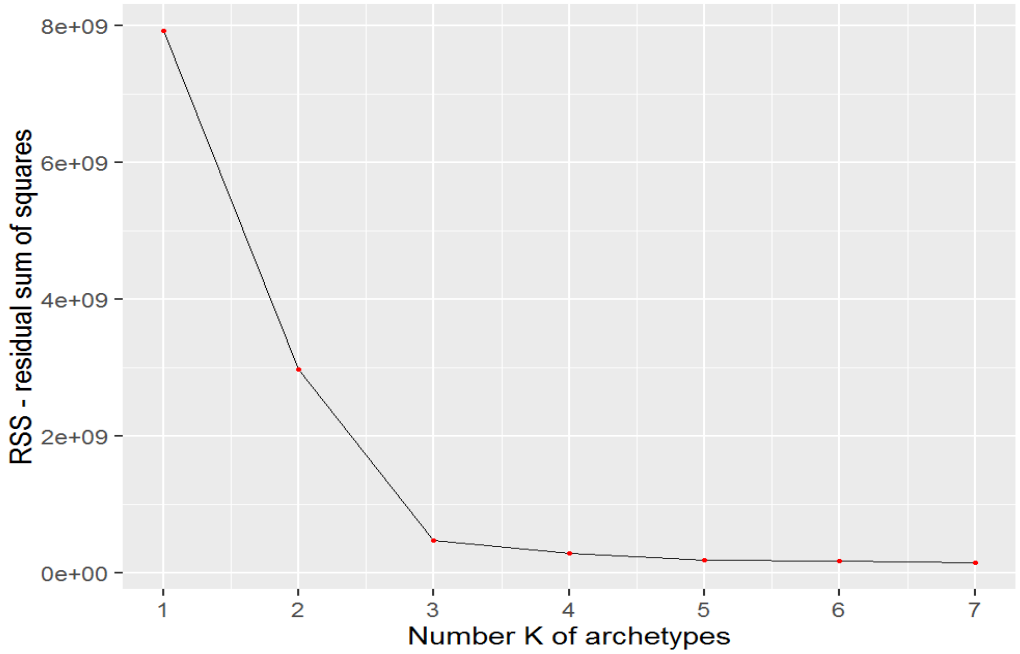


Figure 4.3: Residual sum of squares for different number of Archetypes (from 1 to 7). According to elbow method rule, best choice is K=3

ing mathematical optimization in MATLAB, by means of `fmincon` non-linear programming solver, starting from elaboration developed for interval-valued archetypes as in Corsaro and Marino, 2010. In order to perform the estimation of the archetypes, the first stage of the data preparation phase is to express each histogram as a function of centers and radii. To improve the algorithm speed, as proposed for other clustering techniques for histogram-valued data (like k-means, fuzzy k-means and so on), histograms are registered to share the same number of bins, in this case 8, and so to have the same proportion of the distribution in each sub-interval (each π_u is equal to 0.125 and the 8 bins have the same area). From this, these histograms objects are expressed in the symbolic data-table as follows:

$$\begin{bmatrix} C_{1,1.1} & \cdots & C_{1,8.1} & C_{1,2.1} & \cdots & C_{1,8.1} & R_{1,1.1} & \cdots & R_{1,8.1} & R_{1,2.1} & \cdots & R_{1,8.1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ C_{19,1.1} & \cdots & C_{19,8.1} & C_{19,2.1} & \cdots & C_{19,8.1} & R_{19,1.1} & \cdots & R_{19,8.1} & R_{19,2.1} & \cdots & R_{19,8.1} \end{bmatrix}$$

Where the rows are 19, one for each France region, divided into centers (16

centers each row, 8 for each variable) and radii (16 radii each row, 8 for each variable). Therefore, matrix data-table has dimensions $(19, 32)$. Archetypes are identified after 444 iterations, when a local minimum that satisfies the constraints is found. It means that the optimization is completed because the objective function is non-decreasing in feasible directions, given both the selected values of the optimality and constraints tolerances.

```
Optimization completed: The relative first-order optimality measure, 3.680873e-07,
is less than options.OptimalityTolerance = 1.000000e-06, and the relative
maximum constraint violation, 2.381247e-17, is less than options.
ConstraintTolerance = 1.000000e-06.

Optimization Metric                                          Options
relative first-order optimality =   3.68e-07     OptimalityTolerance = 1e-06 (selected)
relative max(constraint violation) =   2.38e-17  ConstraintTolerance = 1e-06 (selected)
```

Given that archetypes estimation has a similar flavour to the identification of centroids in k-means analysis (both analysis figure out abstract entities useful to point out some patterns in data), results thus obtained in AA will be compared to centroids obtained after performing k-means algorithm for histogram-valued data. The basic idea is that it is likely that, as for single-valued AA, archetypes are "more extreme" compared to the centroids, leading to represent some different aspects with respect to overall data structure.

As can be seen in 4.5, centroids seem to follow some general trends inside the 19 regions, acting like an abstract barycenter of a group of regions. Let recall that both analysis (k-means and AA) are performed considering at the same time both variables (farming cost and cost of wheat), so centroids and archetypes have to be considered as bivariate abstract entities, and their location have to be discussed taking also into account this. As well as from distributions presented in 4.5, median values showed in 4.2 highlights how the highest median is found for first archetype (extreme behaviour in terms of both high cost, farming and wheat) while lowest median for farming cost is found for the second archetypes and so on. Overall, according to both histograms visual displays and median comparison, archetypes such identified look more "extreme" with respect to centroids.

Furthermore, an other element to compare relative position between archetypes and centroids, is to perform a Principal Component Analysis (PCA) between all the variables involved and project archetypes and centroids as supplementary individuals. In this context, PCA takes into account 32 numerical variables (2 histogram variables made up by 8 bins, therefore 16 centers and 16 radii).

Table 4.2 Median values for histogram-valued objects with respect to 19 different France regions, about Farm and Wheat costs, and about centroids and archetypes.

	Y_TSC	X_Wheat
Ile-de-France	38579.09	40936.70
Champagne-Ardenne	24375.91	13234.38
Picardie	49513.34	32382.43
Haute-Normandie	31633.47	16595.58
Centre	29638.70	21555.85
Basse-Normandie	27009.09	3407.05
Bourgogne	25095.33	5581.26
Nord-Pas-de-Calais	37133.65	16845.54
Lorraine	39705.14	16140.41
Alsace	17231.75	3340.37
Franche-Comte	25616.55	2062.21
Pays de la Loire	26009.39	4581.86
Bretagne	28831.50	3689.93
Poitou-Charentes	22925.15	5241.36
Aquitaine	15318.80	0.00
Midi-Pyrenees	14952.22	923.36
Limousin	14179.89	0.00
Rhone-Alpes	14006.74	0.00
Auvergne	14814.14	471.58
Cl.1	37700.57	24076.08
Cl.2	25694.70	5399.72
Cl.3	15083.92	789.22
Archetipo1	42313.85	28348.93
Archetipo2	1567.90	5402.29
Archetipo3	29108.18	13117.25

From 4.4, it comes out that there is some connection between geographical location (South, Center and North) and position in factorial map; it means that regions located in same area share some patterns in their costs about farming and wheat. Most of the South regions (light blue) are in the below-right part of the factorial map, most of the North regions (green) are in the right-top, while on the left part (negative coordinates for first axis) there is a cluster of regions from different area. First two axes, combined, explain a high amount of variation of the original data, about 81%, so it is worthy to analyse where archetypes and centroids are located in the factorial map, assuming that their position can be considered a good proxy of their

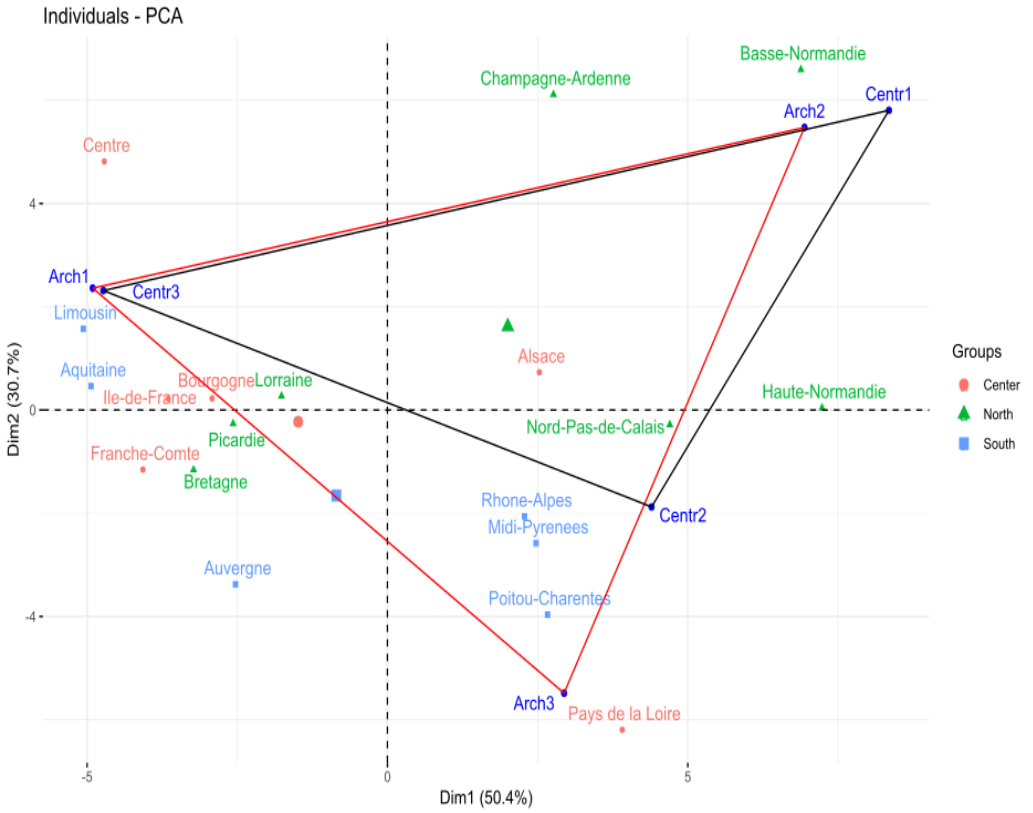


Figure 4.4: PCA with 19 regions, coloured according to geographical position. Three Centroids from k-means and 3 archetypes act as 6 supplementary rows.

real position with respect to each others and to the 19 regions in the original multidimensional space. First archetype and third centroid are very close, almost overlapped, and pretty close are as well second archetype and first centroid. It is understandable so that these 4 abstract entities, while figured out by means of different functions, represent a similar way to describe some patterns in data. Though, a significant difference is found for what concerns third archetypes and second centroid, with the former being way more extreme with respect to the factorial map than the latter. Looking at the two different lines connecting centroids (black line) and archetypes (red line), the red triangle is able to cover a bigger surface than the black triangle; it is about 35% larger. It means that, on the average, archetypes are more extreme with respect to centroids, covering a bigger part of the factorial map. It is interesting to analyze the trend in terms of archetypes location when the K number of archetypes increases from the allowed minimum, $K = 1$, to a bigger number, let say $K = 7$. Let first look at the factorial map in 4.6.

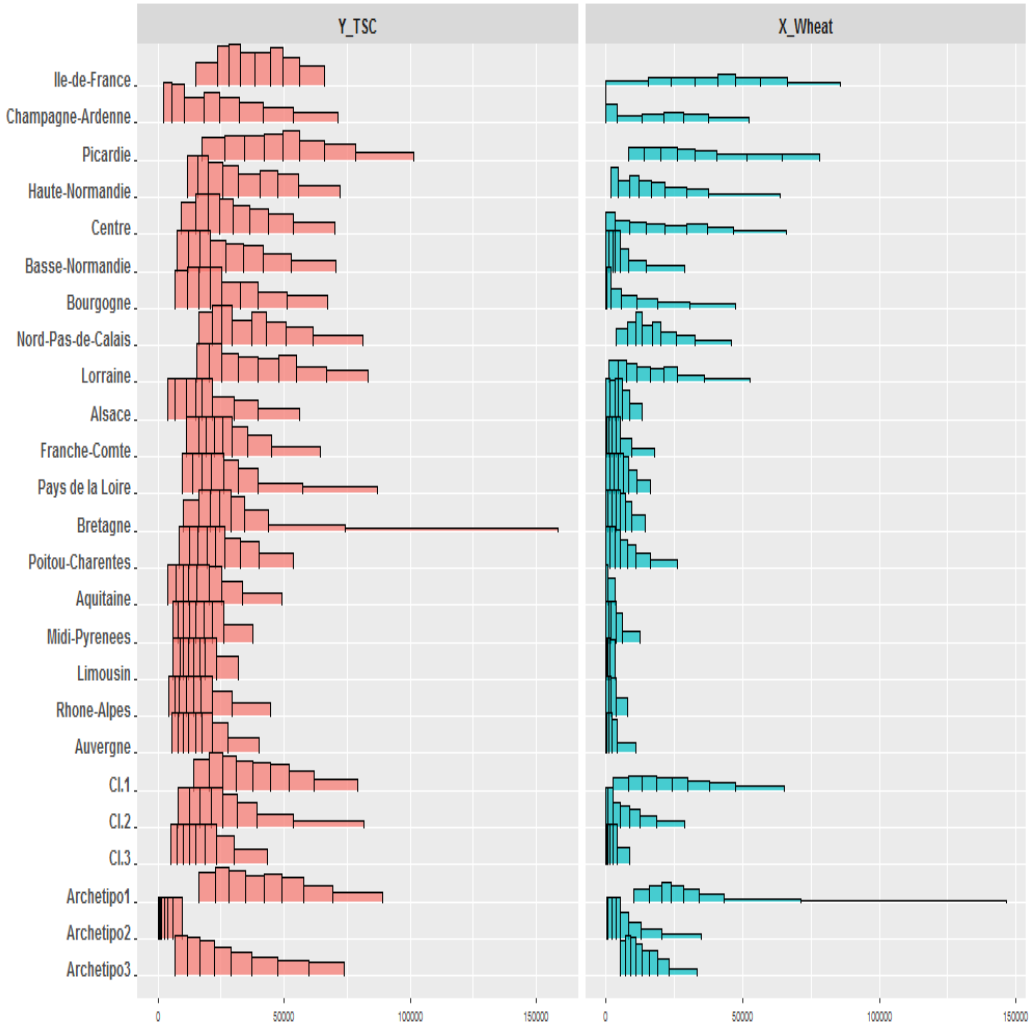


Figure 4.5: Density approximation of Farm and Wheat cost for 19 France Regions in 8 equi-frequent bins. Centroids from k-means and 3 archetypes highlight different patterns, with archetypes that seems to represent more extreme behaviour.

When only one archetype is identified, so when $K = 1$, it is located in the middle of the factorial map, very close to the origin of axes. It is a similar concept to what happens in archetypal analysis for simple data points, when the quantity that minimizes RSS in case of only one archetype is the sample mean. When 2 archetypes are identified, one is located to the right-upper part, the other one is in the left-bottom in an intermediate position. With $K = 3$, the archetypes in the right-upper part of the map is very close the

one archetype identified with $K = 2$. Combining the 3 archetypes locations, we have an interpretation of the salient patterns in data already discussed in 4.4. If an additional archetypes is identified, the 4 salient entities are divided one for each quadrant, describing so, each of them, a very specific pattern in data structure. Furthermore, there is an increase of the number of units contained in an hypothetical line joining 4 archetypes.

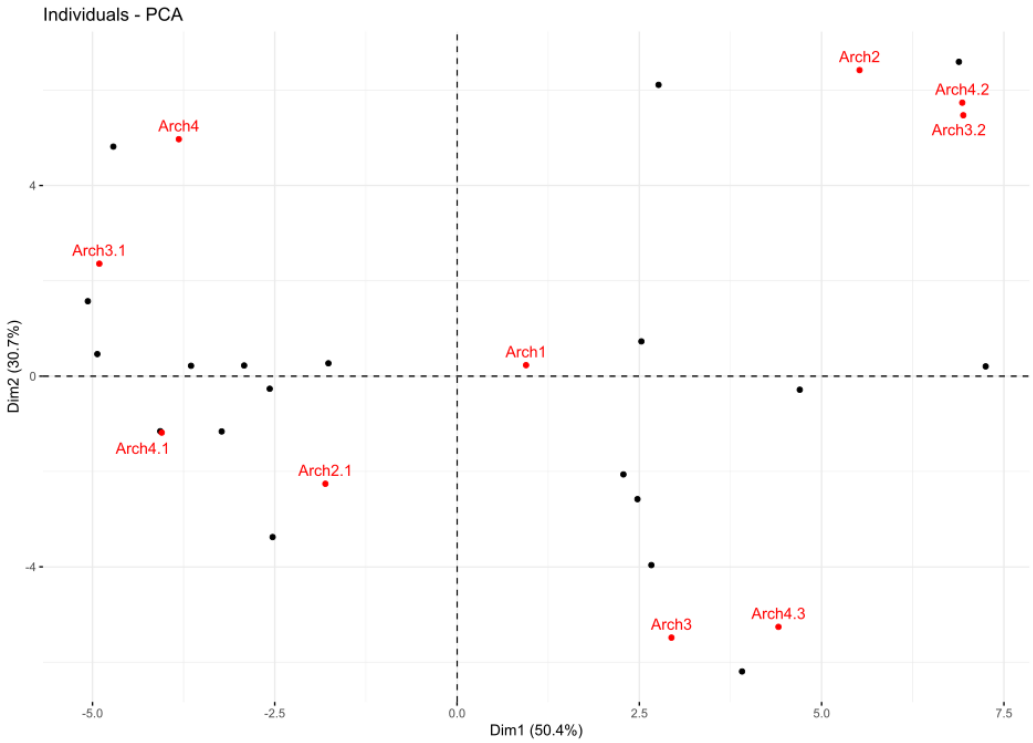


Figure 4.6: PCA of centers and radii with 19 regions in black and archetypes as supplementary rows in red. Different number of archetypes are identified and visualized in this plot: $1 \leq K \leq 4$.

Even if a solution including more than 4 archetypes is likely redundant to describe 19 units, in 4.7 it is proposed the identified location on the factorial map in case of 5, 6 and 7 archetypes, always as supplementary rows. It is made just with the purpose to discuss such location. In case of $K = 5$, in clockwise order, archetypes are located extreme left-up, extreme right-up, extreme right, extreme bottom, extreme left. Is seems that their location is becoming more extreme every time that an additional archetype is added to the analysis, and this is consistent with the general idea and purpose of the traditional archetypal analysis. In case of $K = 6$, 4 out of 5 previous discovered archetypes are basically still there with the same coordi-

nates, with only the left-extreme archetype "Arch5.2" that is now split into "Arch.6.5", left-bottom, and "Arch.6.5", extreme-left. Some small changes happen moving from 6 archetypes to 7, but the general principle holds. The first archetype, "Arch1" in 4.6, is a sort of centroid, in the wide sense of the term, of the matrix including histogram-valued objects in centers and radii notation. As more archetypes are identified, more extreme behaviour are caught, to the point that identifying additional archetypes (for example, moving from 6 to 7 in 4.7), is not adding real interpretative power in terms of patterns. Archetypes as presented in 4.5, are built using centers and radii

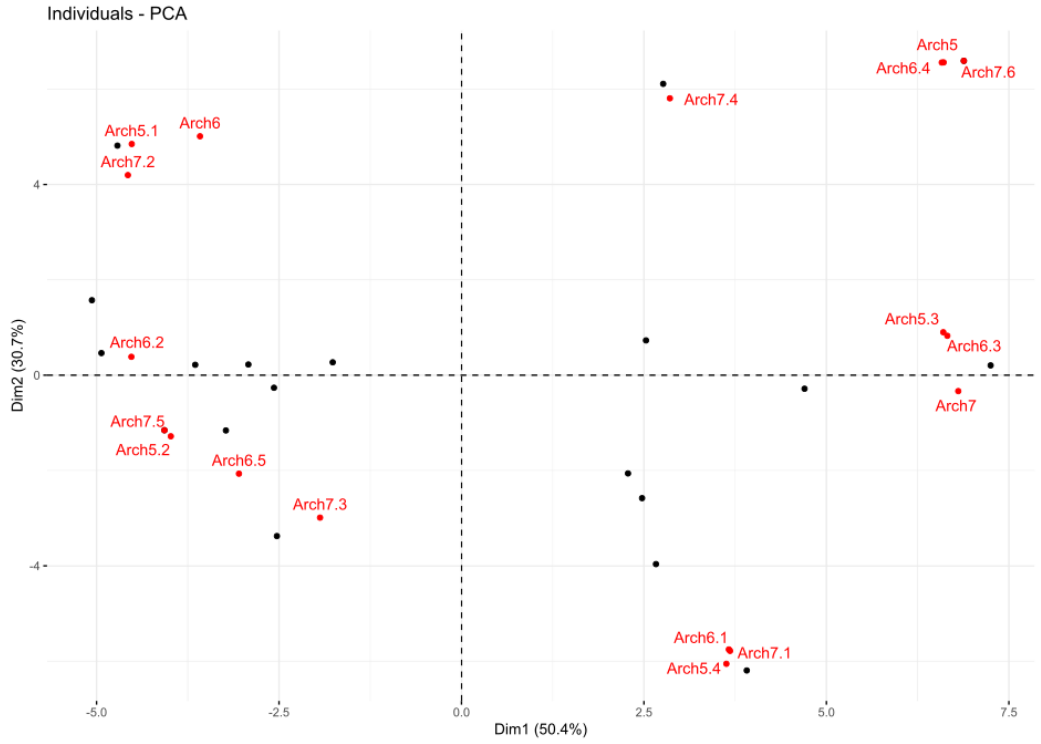


Figure 4.7: PCA of centers and radii with 19 regions in black and archetypes as supplementary rows in red. Different number of archetypes are identified and visualized in this plot: $5 \leq K \leq 7$.

notation, given that the algorithm estimates a 16 centers and 16 radii for each archetype, finding archetypes that are 32-dimensional, treating each center and each radius as a variable itself. Further, it is possible to "reconstruct" original observations using archetypes and α's coefficients, expressing each histogram as a linear combination of archetypes, using elements in the matrix of $\mathbf{A}_K$. As already discussed, α's are share same properties, given the constraints, of compositional data, and it is possible to exploit such com-

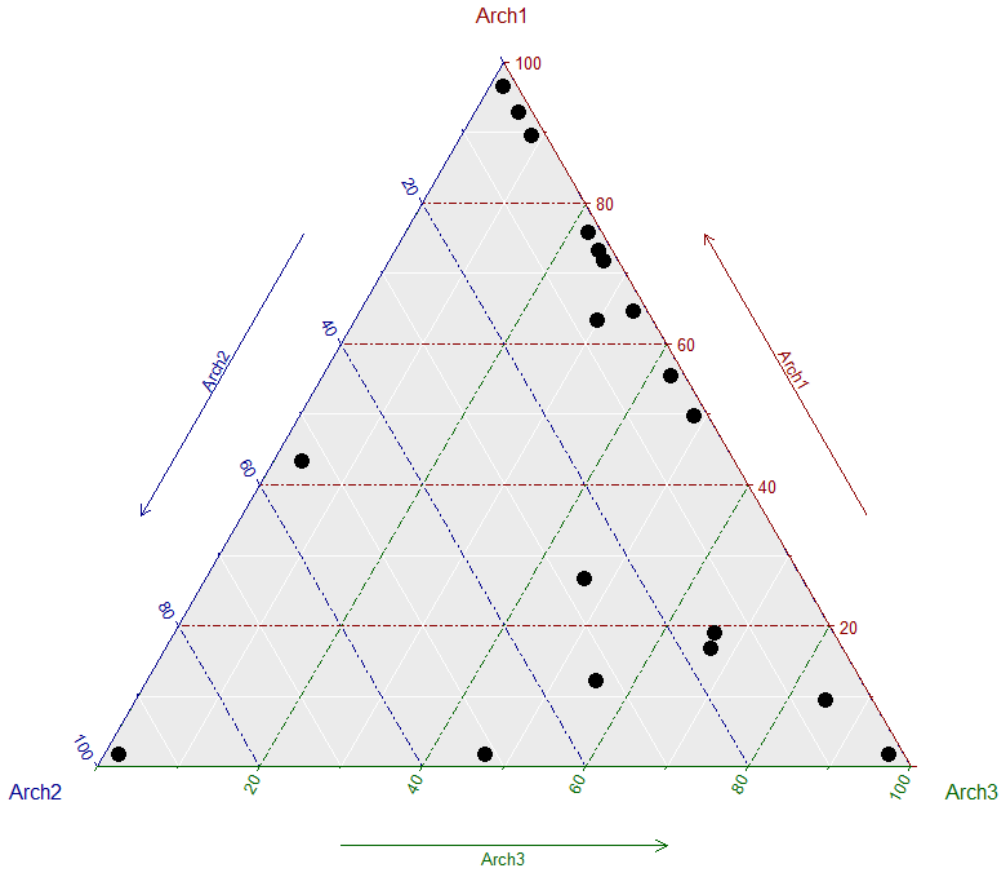


Figure 4.8: A ternary plot, to visualize space spanned by 3 archetypes. It highlighting alpha's coefficients with respect to corners (archetypes).

positional data properties to make further assessments about relationships between observations in the space spanned by archetypes. Ternary plot is a common way to visualize such space, in which coordinates of points (19 France regions) are given by α's coefficients and the vertices of the triangle are the 3 archetypes.

In the first ternary representation in 4.8, points are displayed highlighting the amount of each of the 3 archetypes in each units, that reflect the coordinates assumed by each point. Therefore, points very close to one corner of such triangle has coordinates, with respect to that archetypes, close to 1, and very close to 0 for the other two archetypes; while points in the middle have likely coordinates similar for all the three archetypes, and so they are represented as a weighted mixture of all three archetypes with weights pretty similar to each others. In this sense, this representation recall the idea of a fuzzy-clustering using vertices of ternary plots, the archetypes, are

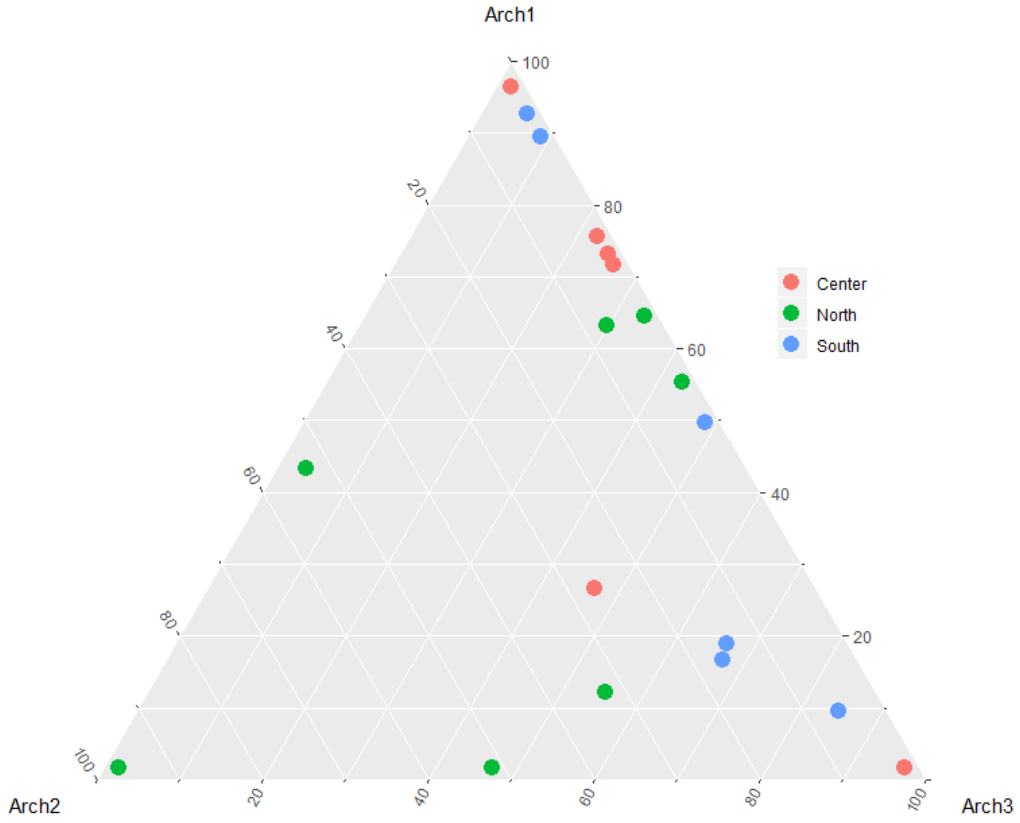


Figure 4.9: A ternary plot, to visualize space spanned by 3 archetypes. it uses different colors for different regions area, as in PCA in 4.4

barycenters. In the second ternary plot, in 4.9, points are coloured with respect to their geographical area, and it is a way to evaluate if archetypes space produces similar results compared to PCA in 4.4. Starting from the second archetype (bottom-left in 4.9), it can be pointed out that the closest points are green points (North regions), as in the PCA. Third archetypes is, on the other hand, closer to one Center region (red) and 3 Southern regions (blue), as well as in PCA. Let recall that PCA is unable to explain a certain amount of variation, so in the ternary plots some patterns are slightly different from the factorial map, but overall they seem to highlight similar data structure, and , most important, archetypes play an analogue role in both representation. Last ternary plot is made to stress out the results of a clustering analysis (k-means) performed in the compositional space spanned by archetypes, using the Aitchinson distance and figuring out 4 well-separated clusters. Points have different colors according to the the cluster they are assigned to. Green points are closer to 2^{nd} archetype, blue points are closer

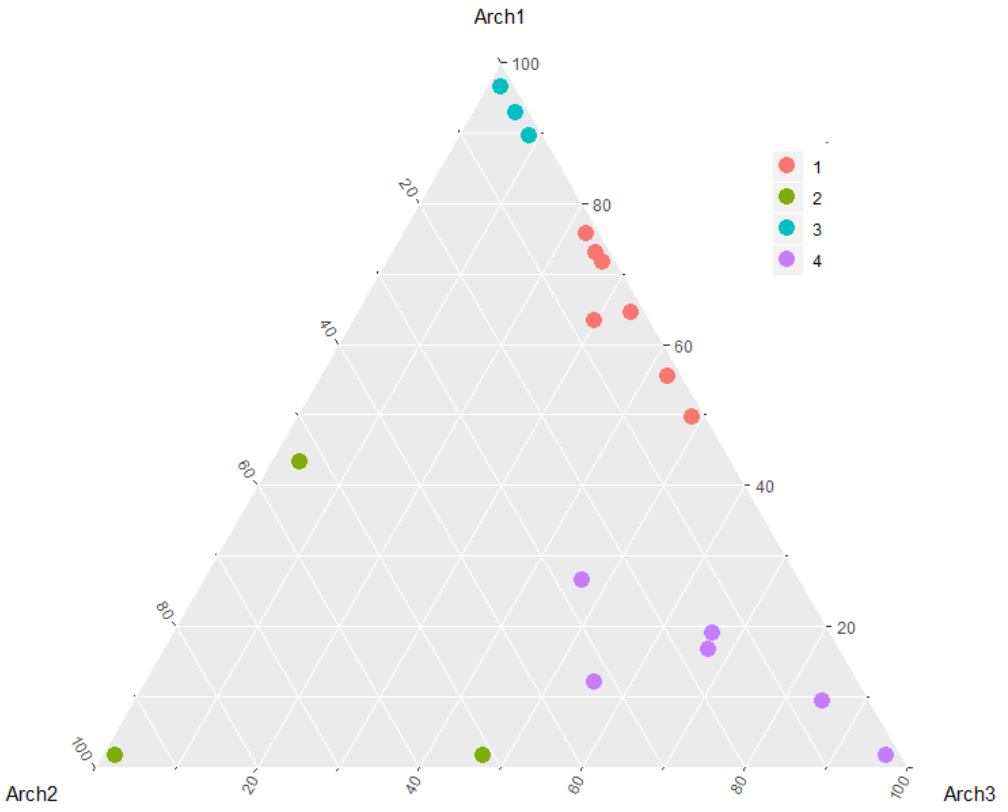


Figure 4.10: A ternary plot, to visualize space spanned by 3 archetypes. It uses different colors according to different groups identified by clustering in compositional space using Aitchinson distance, with 4 clusters identified

to 1st archetype and purple points are closer to 3rd archetype. There is a further group, the red one, that is made up by points lying somehow between top part, 1st archetype, and right part, 3rd archetypes. It explains how is possible to exploit compositional properties in the space spanned by archetypes, leading also to the definition of clusters inside this space. In the following sections it will be deepen how AA for histogram-valued data are an useful tool for benchmarking analysis, and, in particular, an application to Italian school system will be proposed

4.3 On the algorithm for the histogram archetypes identification

The algorithm for histogram-valued data archetypes identification is build in 3 different components:

1 Constraints for the optimization.

2 Function to be optimized.

3 Archetypes function optimization procedure given constraints and the function.

The whole algorithmic procedure is developed, in MATLAB environment, using the routine `fmincon`, that is a non-linear programming solver of MATLAB Optimization Toolbox. The optimization routine is made by means of an implemented method that is a sequential quadratic programming method. General idea behind this kind of implementation can be found in Fletcher, 2013 and in Gill, Murray, and Wright, 1981. The application is similar to the one developed for interval archetypes in Corsaro and Marino, 2010. It is an iterative method, and in each iteration a quadratic programming problem is solved, using a quadratic Lagrangian function approximation with respect to the optimization problem. The algorithm is based on a line search strategy (Wächter and Biegler, 2006). In the line search strategy, the algorithm chooses a search direction, and tries to solve a one-dimensional minimization problem, and then it calculates the gain. At each iteration the algorithm follows a criteria to choose such direction and searches along this direction for a new best solution in the new iteration. For this implementation, the algorithm chooses direction based on Quasi-Newton approach (D. Kim, Sra, and Dhillon, 2010). Quasi-Newton Methods (QNMs) are a wide class of optimization methods that are used in Non-Linear Programming context when the traditional full Newtons Methods are: (i) too time consuming (ii) too difficult or complex to use. Methods belonging to this group allow to find the global minimum of a generic function that is at least twice-differentiable. The advantage is that an approximation of the Hessian is used. Some possible shortcoming are related as well to the fact that an approximation is included and not the analytically computed Hessian, but most of the time for this kind of problem the trade-off in terms of time and complexity is positive in using a QNM method. Such Hessian is updated adding the gained information of each iteration. The following pseudocode, in two parts (first part in 1 and second part in 2), describes the initialisation of constraints, the

Algorithm 1 Archetypes for histogram data - preparation

procedure CONSTRAINTS INITIALISATION(X, m, K, v) ▷ Inputs of the constraints
  $v_{(2*K*m,1)} \leftarrow$ matrix of random values between 0 and 1
  $X \leftarrow$ data matrix
  $K \leftarrow$ number of archetypes
  $m \leftarrow$ number of variables
  $ceq\alpha \leftarrow \underline{0}_{(1,m)}$ ▷ Matrix of zeros
  $ceq\beta \leftarrow \underline{0}_{(1,K)}$ ▷ Matrix of zeros
  for i in $1 : m$ **do**
    $ceq\alpha(i) = ceq\alpha(i) + v(i + j * m)$
  for j in $0 : K - 1$ **do**
    $ceq\beta(i) = ceq\beta(i) + v(K * m + i + j * K))$
  ceq=merge$[(ceq\alpha)^\intercal; (ceq\beta)^\intercal]$
  return ceq ▷ output, starting values for constraints
procedure ARCHETYPES FUNCTION(v, X, m, n, K) ▷ Inputs of the archetypes function
  $n \leftarrow$ number of observation
  for i in $1 : m$ **do**
    for j in $1 : K$ **do**
      $\alpha(i, j) = v(i + (j - 1) * m)$ ▷ Update α's
  for i in $1 : K$ **do**
    for j in $1 : m$ **do**
      $\beta(i, j) = v(K * m + i + (j - 1) * K)$ ▷ Update β's
  $A = \alpha$
  $B = \beta$
  $XBA_r = abs(A * B) * X(:, n + 1 : 2 * n)$ ▷ Archetypes radii
  $XBA_c = A * B * X(:, 1 : n)$ ▷ Archetypes centers
  $XBA = merge[XBAc; XBAr]$ ▷ Archetypes radii and centers
  $funct = (X(:, 1 : n) - XBA_c).^2 + (1/3) * (X(:, n+1 : 2*n) - XBA_r).^2$
  RSS: definition of function to minimize according to 4.1
  $f =$ Frobenius Norm(funct) ▷ Final function to optimize
procedure OPTIMIZATION OPTIONS(Maxiter, Maxfuneval,Tolerances) ▷ Options for constrained optimization
  maximum number of function evaluations $\leftarrow$ 1000000
  maximum number of iteration $\leftarrow$ 100000
  constraint level tolerance $\leftarrow e^{-6}$
  function level tolerance $\leftarrow e^{-6}$ ▷ Options fixed

Algorithm 2 Archetypes for histogram data - Optimization

procedure OPTIMIZATION THROUGH `FMINCON`(Optimization Options, Function, Constraints)
$lb = \underline{0}_{(2*K*m,1)}$ ▷ Vector of zeros
$ub = \underline{1}_{(2*K*m,1)}$ ▷ Vector of ones
optimization result =`fmincon`(v)Archetypes function(v,X,m,n,K),v0,...lb,ub,...
under constraints *ceq* and optimization options
for $i = 1 : m$ **do**
for $j = 1 : K$ **do**
$\alpha(i, j) = v(i + (j - 1) * m)$ ▷ Update α's
for $i = 1 : K$ **do**
for $j = 1 : m$ **do**
$\beta(i, j) = v(K * m + i + (j - 1) * K)$ ▷ Update β's
if both: (i) "TolCon" constraints tolerance satisfied (ii) "TolFun" tolerance about function satisfied **then**
Archetypes centers $\leftarrow \beta * X(:, 1 : n)$ ▷ Building archetypes centers using β's
Archetypes radii $\leftarrow \beta * X(:, n + 1 : 2 * n)$ ▷ Building archetypes radii using β's
Data centers rec. $\leftarrow \alpha *$ Archetypes centers ▷ Reconstructing original data cent. using α's
Data radii rec. $\leftarrow \alpha *$ Archetypes radii ▷ Reconstructing original data radii using α's

Figure 4.11: Pseucocode for histogram archetypes identification. Constraints initialisation, archetypes function and Optimization Options are in Algorithm 1 (in 1). Optimization through `fmincon` is in Algorithm 2 (in 2)

definition of the function to optimize, the choice of the optimization options and lastly the output.

The pseudocode in 2 describe the algorithm procedure using the `fmincon` optimization. It follows a reasoning to search the new step to do in each iteration; it is explained in the ongoing flowchart for Line Search Strategy:

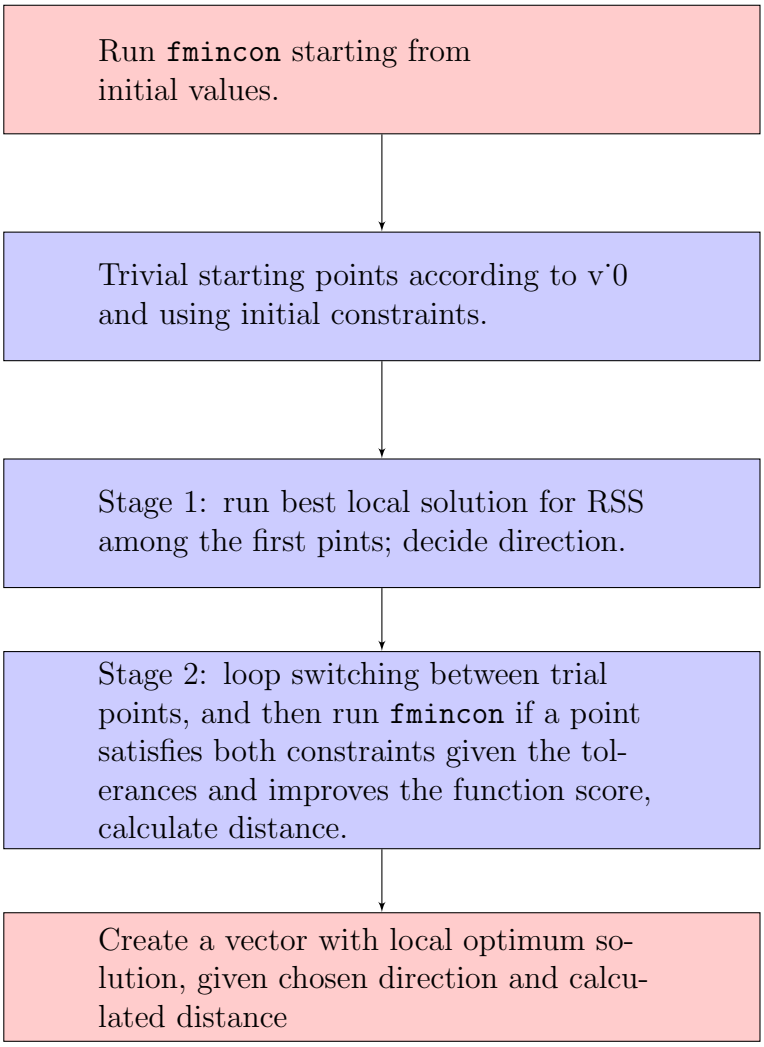


Figure 4.12: Flowchart for line search strategy in `fmincon` routine for each iteration.

Chapter 5

Italian School System benchmarking by means of Histogram AA

The evaluation of the performance of the educational system using several quantitative methods is becoming, year by year, increasingly important. For what concerns school system, the evaluation is performed in an *quantitative internal benchmarking* perspective (Binder, Clegg, and Egel-Hess, 2006; J. Zhu, 2014). The definition of such analysis referes to the fact that benchmarking process has been applied also into contexts that are supposed to by apart from the logic of profit. For that reason, it has been adopted not only in private companies, but also in public sectors (Kouzmin et al., 1999). This new field of application has lead to establish the self-assessment concept, that is the core of internal quantitative benchmarking in public sector. The idea is that a public sector is able to manage a huge amount of data regarding its sub-entities, and it aims to evaluate their performances with respect to each others and with respect to the overall sector tasks to be achieved, defining standard of excellence to face challenges. On the other hand, when private companies perform benchmarking analysis, they are interested in comparing themselves with peers, in order to adopt best practices of competitors to increase profitability. Therefore, in this context the assumption is that all the information regarding sub-entities are available and collected, while in external benchmarking information about competitors are usually hard to get.

In literature, several authors have exploited statistical methods to make assessments about educational system evaluation in an internal quantitative benchmarking perspective, as Kelly, 2004 and Goldstein, Bonnet, and Rocher, 2007 among others. For what concerns Italian School System sce-

nario, the evaluation is in charge of the Italian national institute for the evaluation of the school system *INVALSI*, whose main aim is to gather data from various sources and provide tools and comprehensive analyses for the evaluation of the school system as a whole. Data that are treated and made available by INVALSI, would enable policy makers, politicians, administrators in general, but also citizens, to assess if the Italian school system is achieving its objectives. Usually, INVALSI institute publishes itself most important results of their own analyses, and also several publications or technical reports are available in its own website http://www.invalsi.it/invalsi/index.php.
According to Fondazione Giovanni Agnelli, 2014, several intermediate steps have been necessary to reach to the actual INVALSI system, and the path have not been always linear. Following the chronological order, most important steps have been:

i Centro Europeo dell'Educazione (CEDE) has been estabilished in the May of 1974 and has become fully operative in 1982, with a public selection of teachers to involve in the organization administration task. Main aim of CEDE was to put Italian education assessment to an European standard in terms of methodology.

ii Embedded in CEDE framework, the Servizio Nazionale per la Qualitàdell'Istruzione (SNQI) has been founded as a temporary institution, waiting for a decisive reform in Italian education system.

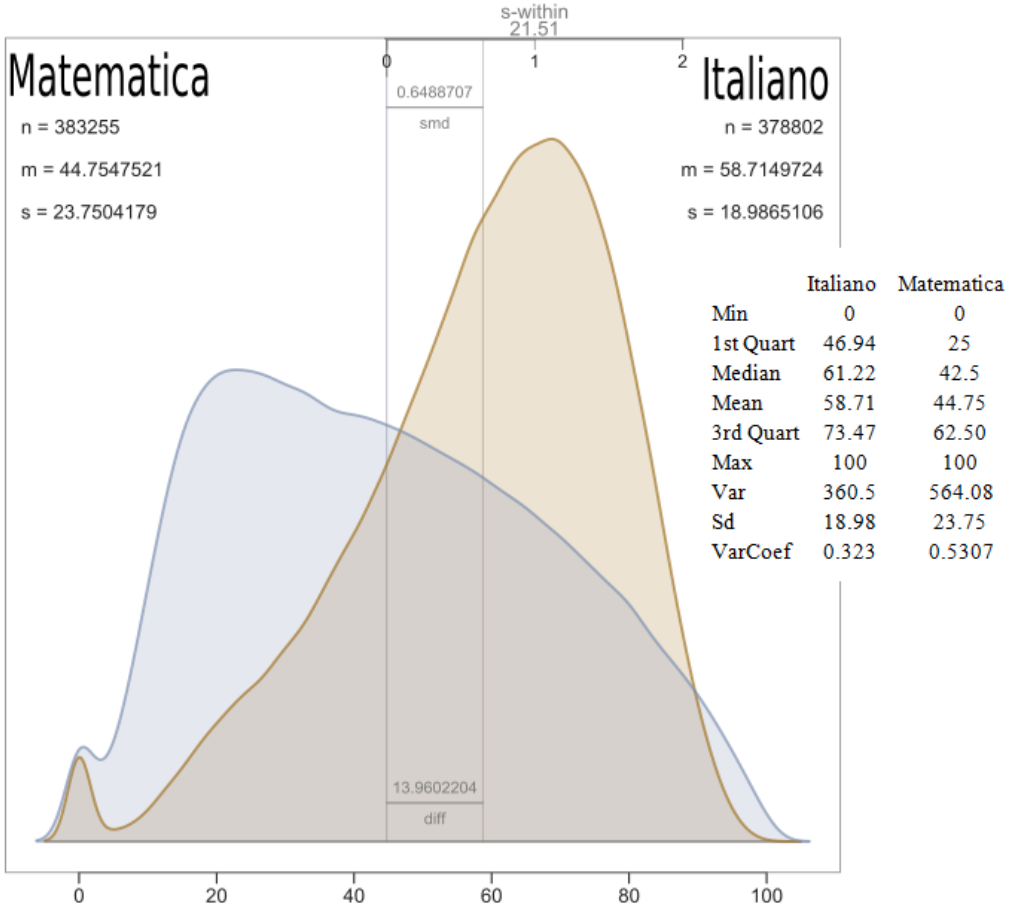
iii Two years later, in 1999, CEDE is definitely transformed into Istituto nazionale per la valutazione del sistema dell'istruzione (INVALSI).

iv Five year later, in 2004, the organization has changed its full name in Istituto Nazionale per la Valutazione del Sistema educativo dell'Istruzione e della formazione, but the acronym remains still INVALSI

Through the years, the terminology used, as well as the long term objectives of the institution, has changed. Looking at the terms involved in the definition, institutes has moved from "educazione", to "istruzione" and lastly to "formazione". These terms are not so different in Italian (all of them could be translated with education), but are highlighting different aspects of the pupils' scholastic experience. This is consistent with the changes in the concept of what Italian school system represents in terms of sector aims, from the 70's up to the recent years. The word "educazione" refers to what is needed to provide a proper process of adaptability for the next generations; "istruzione" refers to what is needed to provide a proper process of adaptability for the next generations by means of acquiring new specific knowledge;

"formazione" refers to what is needed to provide a proper process of adaptability for the next generations by means of acquiring new specific skills useful to the future recruitment to get a job. Even the shift from "Qualità" (quality) to "Valutazione" (evaluation or assessment) underlines how the whole conception of the system has been affected by the social, political, cultural and economic dynamic environment. It is possible to assume as key year the 1990, cause from that moment on the concept of "Assessment culture" has been introduced in Italian public system. In that year a National School Conference was opened to discuss scientific and political issues, in order to build a reform design to give a different direction to the education system and its autonomy. The idea was to align Italy to other western Countries that already had an effective system of evaluation of public sectors.
INVALSI institute makes use of a set of standardized tests to evaluate the proficiency of students attending different schools at different years. Several domains are tested, and main domains are mathematical skills and italian language proficiency, in terms of both reading and writing skills (often this domain is simply called "Italian"). For what concerns overall state of art about INVALSI tests and about the debate of INVALSI role in providing tools to improve scholastic performances, a thorough discussion can be found in Trinchero, 2014. But many authors have focused their works directly on the analysis of pupils' proficiency. Several contribution have been made using multilevel models to analyze data from primary school INVALSI tests to evaluate, in a regression perspective, the profiency and the skills variability of pupils (Grilli and Sani, 2010; Sani and Grilli, 2011). Multilevel model, given the natural hierarchical structure of pupils (pupils nested in classes, in turn nested in schools, in turn nested in provinces and so on) seems to be a proper methodological choice to evaluate patterns in such data. In Petracco-Giudici, Vidoni, and Rosati, 2010 it has been deepen the state of art for what concerns primary school as a whole, while in Capperucci, 2017 the focus has been on pupils' mathematical skills. Among the diverse set of quantitative, and statistical in particular, techniques that have been proposed and implemented for INVALSI tests, we assume that there is opportunity to conceive and experiment a new technique able to preserve more information than traditional techniques.
The main aim of this section is, therefore, to use an innovative statistical tool, archetypal analysis for histogram-valued data, to provide a data-driven benchmarking analysis of a part of INVALSI tests data. As a consequence, taking as references previous known results and patterns already available mainly from the mentioned works, but also from international literature about scholastic proficiency, it will be possible to discuss findings from exploratory archetypal analysis with respect to such previously known theses.

Data under consideration refers to INVALSI standardized tests for the academic year 2015-2016. with respect to two domains, italian (reading and writing) and mathematics, for pupils attending the 2nd year of high school. These standardized scores range from 0 to 100. The total number of mathematics test observations is 383, 255, while for italian test the total number of observations is 378, 802. Overall, data length is 762, 057. INVALSI tests



	Italiano	Matematica
Min	0	0
1st Quart	46.94	25
Median	61.22	42.5
Mean	58.71	44.75
3rd Quart	73.47	62.50
Max	100	100
Var	360.5	564.08
Sd	18.98	23.75
VarCoef	0.323	0.5307

Figure 5.1: Two Kernel density estimations, for both domains if INVALSI tests, pupils' level. Mathematical domain has lower scores distribution and higher variation. It is highlighted also, consistently, by most common descriptive statistics computed for such distributions.

are provided in a census-like fashion, so all the pupils belonging to the population target are involved in the test. Pupils population is made up by:2nd year of primary school, 5th year of primary school, 3nd year of secondary school first grade, 2nd year of high school. Given the census survey collec-

tion strategy, let consider that some schools, classes and then pupils are also included in a sampling procedure. The sample is extracted in two stages: in the first one schools are extracted,and in the second one, usually, two entire classes belonging to each sampled school are included in the final sample. Schoolchildren included in the sample have to carry out the same identical tests, but the difference is that an external official observer is present during the test progress. The aim of the sample is, thus, to ensure the regularity of the test and an higher degree of reliability of final scores. For this reason, some INVALSI official reports are based only on data collected in this sampling framework. Moreover, several other variables are collected with respect to pupils and for what concerns classes and schools, other than proficiency scores. Demographic and social background variables are collected for pupils (gender, nationality, parents' educational level, residence and so on). For the class and the school collected variables are related to size, in terms of number of pupils, and about program typology (general for school, specific for class). Moving to data proposed in 5.1, Italian proficiency has, considering all pupils togheter, higher mean value, higher median and higher quartiles, compared to mathematical proficiency. It has also lower variability (as showed by coefficient of variation in 5.1), and one of the aim of educational system is to allow for equal skills among pupils, so a high values in variation measures shows a challenge that has to be faced. As mentioned also in some works such as Costanzo and Desimoni, 2017, that have proposed a quantile approach to analyze scores distribution, the interest in other quantiles of the proficiency level other than the average or the median value is increasing,and this is an hint that tools able to retain a larger part of the original information could be more useful in the next future, going into the direction of an approach like SDA or others with similar logic. In the following, rather than analysing scores with respect to pupils, in this section the focus will be on the distribution of scores with respect to both domains in each school, merging together pupils nested in same school creating an histogram object. Therefore, from $762,057$ different scores, final data-set is made up by $3,882$ schools, and each school is defined by a bivariate histogram-object, with a distribution for each domain.
Given this data structure, working hypothesis to be discussed by means of histogram-valued archetypal analysis, in an internal quantitative benchmarking perspective, can be summarised as follows:

1 Is it possible to identify abstract entities (archetypes) that act as benchmark units, in terms of good, bad or unusual performances?

2 Is it possible to perform a categorization of units (schools) using these benchmark units as reference points?

3 Are these identified categories consistent or not, and to what extent, with respect to previous studies?

5.1 Archetypes identification

Given the high computational cost of the archetypes identification for histogram-valued data, in this context the analysis that is presented takes into account a simple random sample of 200 schools out of the original 3,882. As mentioned, the objective function of the archetypal analysis takes into account a distance between histogram objects that is a function of centers and radii notation. Each histogram object is expressed in deciles. Deciles are chosen

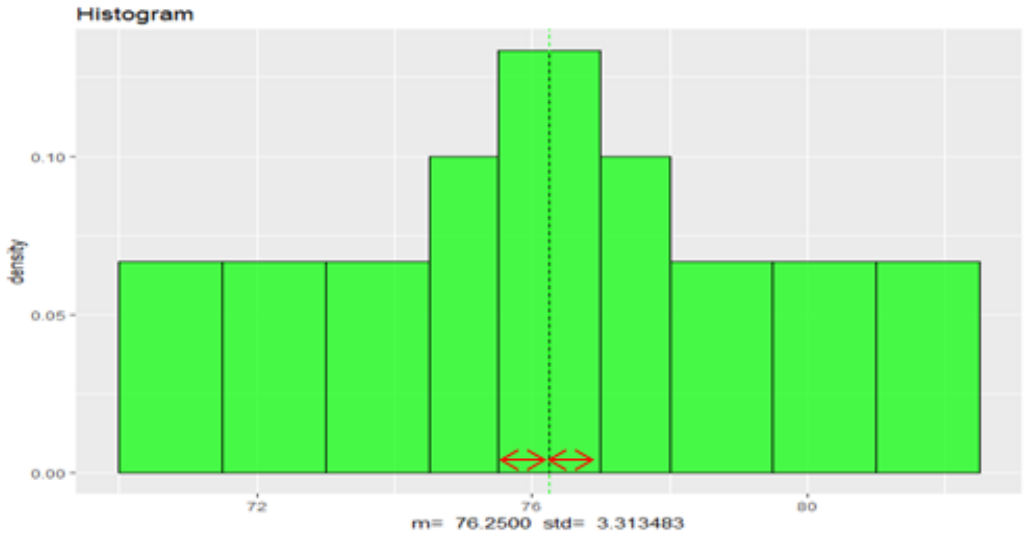


Figure 5.2: Definition of a unit (histogram) of a symbolic data-table in centers and radii notation. Each bin is defined by its center (mean value, the vertical line in the middle) and by its radius (half-width, each red arrow)

cause they are a good trade-off between computational complexity (the more bins in each histogram, the more complex is the optimization) and interpretative power. The final symbolic data table has the following dimensions: 200 rows, 40 columns. Columns are: 10 centers for Italian scores deciles, 10 centers for Mathematics scores deciles, 10 radii for Italian scores deciles, 10 radii for Mathematics scores deciles.
Optimization procedure as presented in 4.2 is successfully achieved, and constraints are satisfied within the tolerance level. Given some exploratory analyses, a good choice for the number of K archetypes seems to be either 3, 4 or 5, and so in the following findings refer to these 3 different cases. Let

Table 5.1 First 6 entries (schools) of symbolic data table and some columns: 40 columns, 20 centers and 20 radii, 2 variables in total. In this table, I. is Italian scores, M. is Mathematics scores, C. stays for Centers, R. for Radii

	I.C.1	I.C.2	I.C.10	M.C.1	M.C.10	I.R.1	I.R.2	I.R.10	M.R.1	M.R.10
1	18.2	29.0	74.7	5.0	55.0	5.9	4.9	9.0	5.0	17.5
2	51.6	65.9	85.0	48.8	87.1	12.9	1.4	2.8	11.2	7.9
3	30.6	43.9	80.6	20.0	77.5	10.2	3.1	3.1	10.0	7.5
4	22.9	32.9	74.3	6.0	65.0	6.5	3.5	3.3	1.0	10.0
5	68.2	70.4	76.7	22.0	58.0	0.8	1.4	0.8	2.0	7.0
6	27.6	37.8	79.6	11.2	65.0	9.2	1.0	4.1	1.2	7.5

start with $K = 4$. Using centers and radii estimated by the procedure in MATLAB, by means of `fmincon` non-linear programming solver, is possible to consistently reconstruct the 4 archetypes. As can be seen in 5.2, the

Table 5.2 Mean, Median, Standard Deviation, Skewness and Kurtosis for the four histogram archetypes

	It.Mean	Mat.Mean	It.Med	Mat.Med	It.Std	Mat.Std	It.Skw	Mat.Skw	It.Kur	Mat.Kur
Archetype1	75.68	77.55	75.77	77.55	3.70	9.43	-0.96	-0.00	4.30	1.80
Archetype2	91.45	31.25	92.85	31.25	5.06	13.71	-1.24	0.00	4.34	1.80
Archetype3	18.87	58.22	11.48	58.61	14.22	20.59	0.50	-0.52	1.82	2.58
Archetype4	11.53	48.80	7.21	48.58	10.19	24.16	1.04	-0.00	3.01	2.08

four archetypes are able to sum up different behaviour in terms of italian scores and mathematics scores. Linking this to the graphical displays of the archetypes as densities estimation in 5.3, it is possible to drawn some conclusions about their role and about what they represent.
The 1st archetype has a mean value for italian score, as well as for mathematics score, higher than the average, and very low standard deviations for both domains. It is an exceptional distribution for math domain, pretty good for italian domain. We can call this archetypes the *good performer all around; best in mathematics*. The 2nd archetype has even better scores for italian domain, while mathematics domain has a distribution below the average, and mathematics standard deviation is increasing as well if compared to the previous archetypes. We can call this archetype the *best in italian; bad in mathematics*. The 3rd archetype shows a bad performance for what concerns italian scores, but a mathematics proficiency above the average, with a negative skewness as well; the left tail is longer and most of the distribution is at the right, with respect to the median. We can call this archetypes the *bad in italian; good in mathematics but with long negative tail*. The last, 4^{rt}

archetype, the is similar to the previous, but with lower values for both italian and mathematics. It is just average for mathematics, and the worst for italian. We can call this archetypes the *worst in italian; average in mathematics*. For the complexity of the problem and for the fact that the algorithm

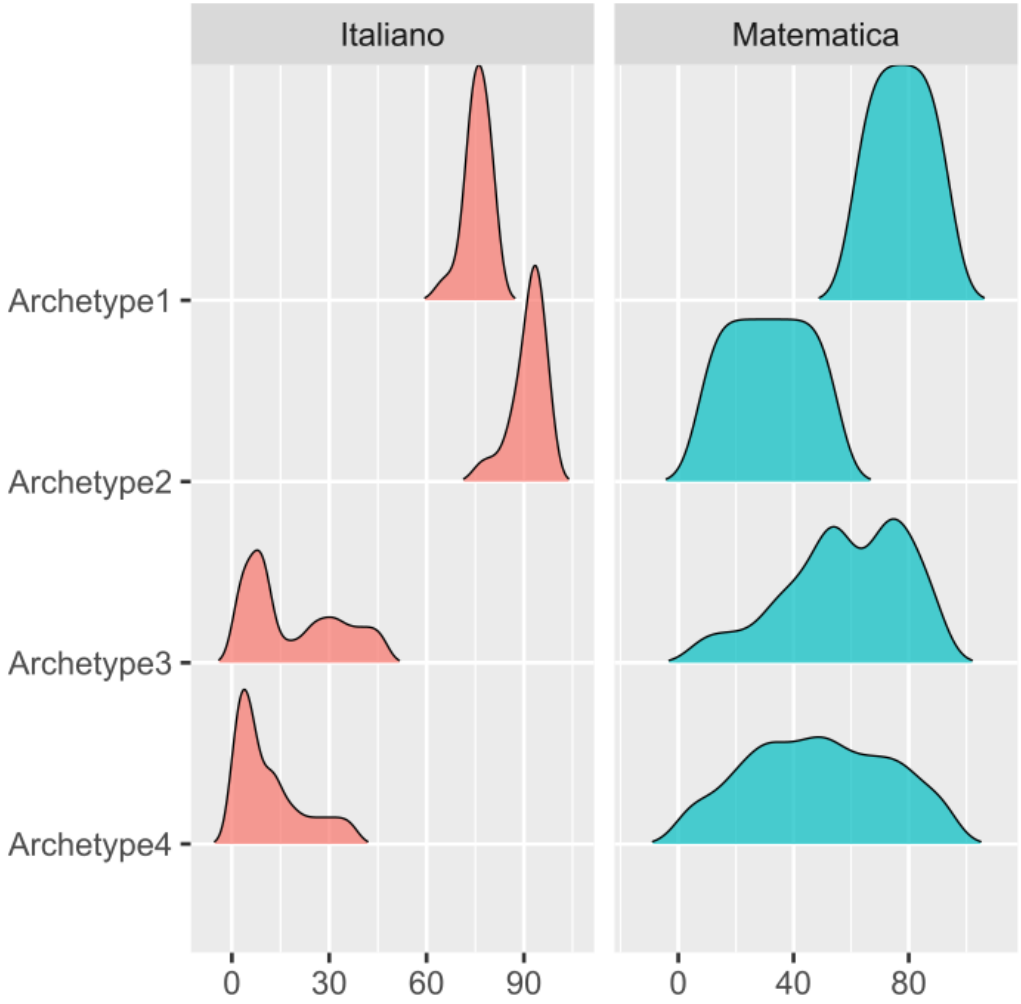


Figure 5.3: The four bivariate histogram valued archetypes

is pretty time demanding, it is not trivial to figure out the proper number of archetypes to better describe and summarize the whole information contained in the original histograms. Other reasonable choices, as mentioned, are 3 or 5, as well as 4. Let first discuss the case in which only 3 archetypes are identified. Archetypes are shown in 5.4. In this case there is no one real archetype that can be assumed as excellence standard overall, neither one is the worst performer. First one and second one are marked by better scores

distribution for writing/reading proficiency, with the second one showing a lower calculated value for both the mean and the median for both domains, with respect to the first archetype (as can be seen in 5.3). Third archetype represents, on the other hand, a very good performance for math score and very poor performance for reading/writing proficiency. As can be seen in

Table 5.3 Mean, Median, Standard Deviation, Skewness and Kurtosis for the three histogram archetypes

	It.Mean	Mat.Mean	It.Med	Mat.Med	It.Std	Mat.Std	It.Skw	Mat.Skw	It.Kur	Mat.Kur
Archetype1	63.83	16.95	67.25	17.48	18.19	9.58	-0.82	0.05	3.04	2.21
Archetype2	47.42	9.35	46.81	10.47	22.97	4.89	0.03	-0.35	2.20	1.85
Archetype3	20.91	66.90	14.40	70.22	13.76	18.55	0.66	-1.29	2.26	4.61

the table 5.3 and from graphic displays of the kernel density estimation of the archetypes in 5.4, differences among archetypes are not only marked by central tendency indexes like the mean or the median, but also by different behaviour in terms of variation, skewness and kurtosis. First archetype has a long left tail (negative skewness) for "italiano" domain, while second one is almost perfectly symmetric for that domain. Therefore, it is an additional element that highlights different behaviour in terms of distribution. It comes out as a more interesting analysis, especially in terms of interpretation, the archetypal analysis when 5 salient units are identified. In this case, visualizing the kernel density estimation of the archetypes in 5.5, the 4^{th} archetype is the one with very high score distribution in both domains. According to the mean value, the best in mathematics is the 5^{th}, while the best in reading/writing is the 3^{rd}. Other important differences between archetypes behaviour, and so about the distributions they represent, arise from the table 5.4. Also in this case, no one can be considered as the absolute archetype of very bad performances overall in both domains. The 4^{th} can be considered as good distribution above the average for both domains, but is it not the best neither in reading/writing nor in mathematics.

Table 5.4 Mean, Median, Standard Deviation, Skewness and Kurtosis for the five histogram archetypes

	It.Mean	Mat.Mean	It.Med	Mat.Med	It.Std	Mat.Std	It.Skw	Mat.Skw	It.Kur	Mat.Kur
Archetype1	9.88	58.33	5.03	62.50	9.77	24.46	1.13	-0.47	3.11	2.41
Archetype2	80.41	28.02	80.25	19.25	2.28	25.28	-0.25	0.49	2.65	1.88
Archetype3	82.08	18.94	82.74	9.14	3.69	22.44	0.65	1.3	2.76	3.75
Archetype4	77.55	74.83	77.55	74.49	9.43	1.35	-0.00	0.50	1.80	1.80
Archetype5	31.25	97.07	31.25	97.49	13.71	1.81	0.00	0.81	1.80	2.98

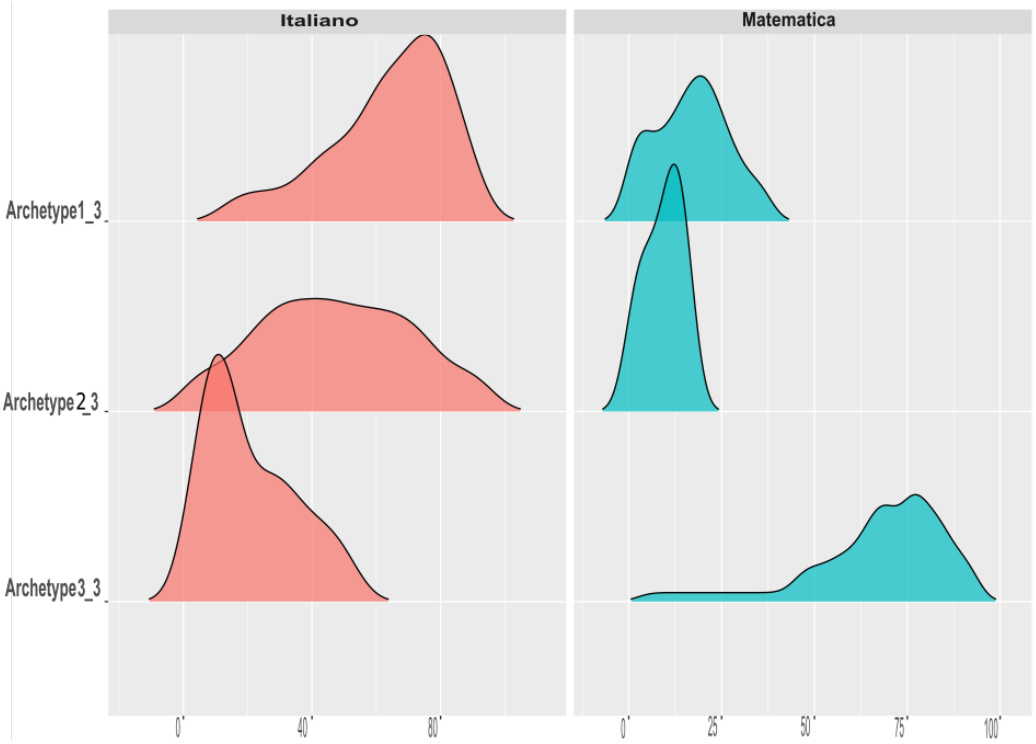


Figure 5.4: The three bivariate histogram valued archetypes

5.2 Archetypes as benchmarking units

To better understand the role of such archetypes, according to their relative positions and with respect to the real observed 200 units (schools), two different analysis will be proposed: the first one, as in 4.4, is a PCA that has been performed on the 200 units while archetypes act as supplementary individuals, and the second one is compositional analysis in the space spanned by archetypes, using α's coefficients as coordinates, visualizing it by means of a quaternary plot.
Principal Component Analysis has been achieved using all the 40 variables: 10 centers for Italian scores deciles, 10 centers for Mathematics scores deciles, 10 radii for Italian scores deciles, 10 radii for Mathematics scores deciles. As can be seen in the bi-plot 5.6, the patterns of the variable, in clock-wise order, starts with radii of italian scores (that are usually smaller), then mathematics radii, then all the centers that are pretty close each others without a clear order between italian and mathematics. By the way, both radii and centers, follow the trajectory from the last (10^{nt} center or radius) to the first (1^{st} center or radius). From the factorial map coordinates, three archetypes (1^{st},

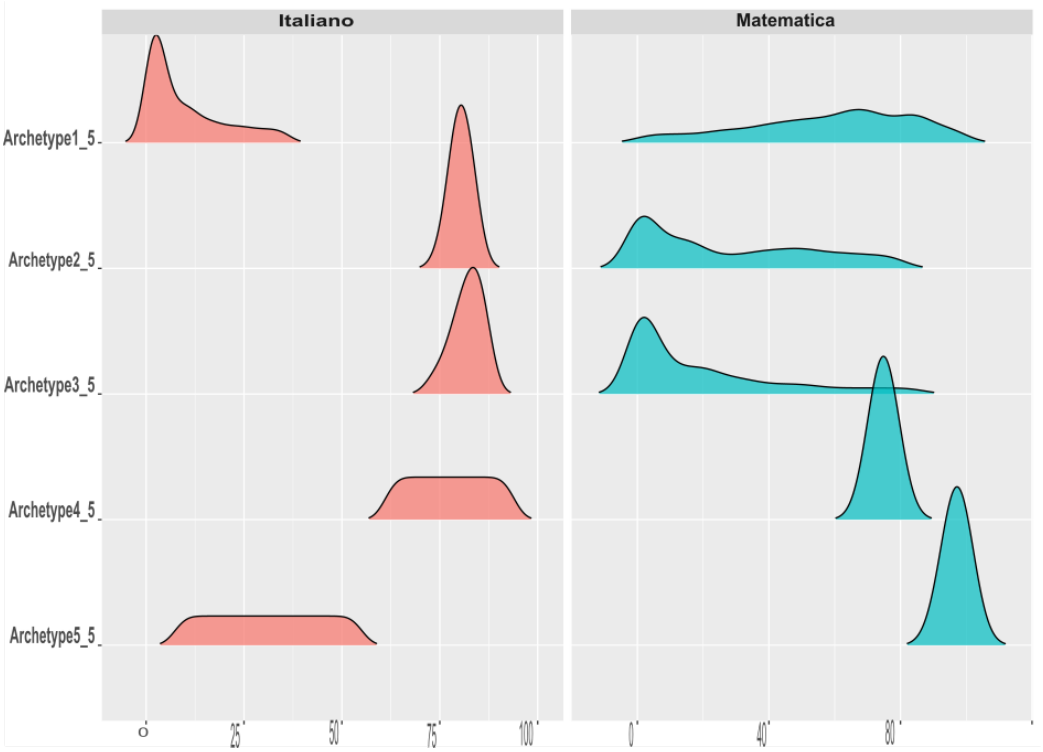


Figure 5.5: The three bivariate histogram valued archetypes

2^{nd} and 4^{th}) have external position with respect to the data-cloud, while the 3^{rd} archetype, even if not so close to a cluster of points, is located in a more central position. Let recall that also here, as in 4.4, PCA is able to explain a considerable amount of overall original variation (in this case about 60%), but it is still an approximation of the first two dimension, in terms of explanatory power, of a problem with way more variables; so, the position of archetypes is a good hint about their real location, even if the conclusion to be drawn could be slightly different.
As said, it is possible to look at the archetypes as vertices of a new compositional space. Each school, in this space, is a point having 4 different coordinates, each of them being a number between 0 and 1, that represent the "percentage" of that archetype that is used to reconstruct the original bivariate histogram. The four α's of each school will always sum to 1. In this case, the plot is a quaternary plot instead of a ternary plot, but the interpretation is similar. In this compositional space, most of the points are pretty close to 4^{rt} archetype and to the 2^{nd}, while areas close to the top (1^{st} archetype) and to the right (3^{rd}) are sparser. To better understand the pattern of points inside that quaternary plot, it is possible to exploit the com-

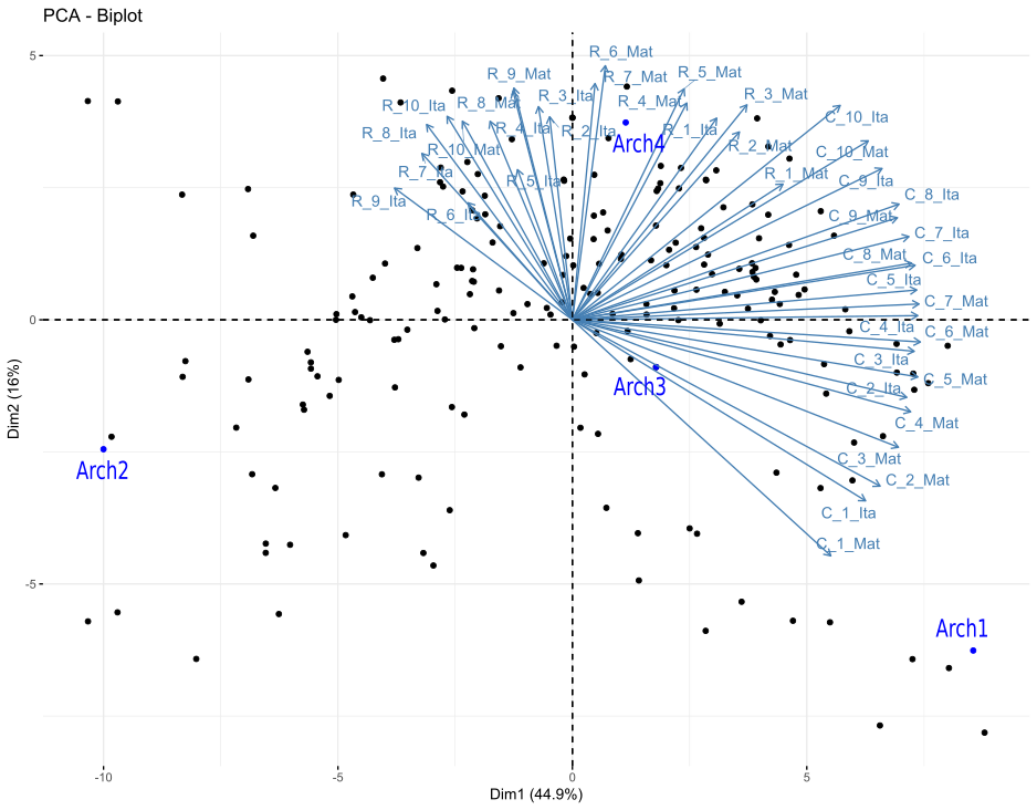


Figure 5.6: Biplot of PCA estimated on 200 schools, 4 archetypes as supplementary rows.

positional space properties, and finding clusters according to the Aitchinson distances. In this context, 5 clusters are identified, as can be seen in the quaternary plot 5.8. These groups are clusters of school sharing similar percentage of archetypes in their reconstruction. As shown in 5.9, analysing their centers, some of them are built for the most part as function of only one archetype, while others are a mixture of several archetypes in a more equal fashion, with similar weights among archetypes. Schools belonging to the 1st group (black colour), for example, are a mixture of the 2nd, the 3rd and the 4^{rt} archetypes, with no contribution at all of the 1st archetype. It is possible to look at the previous PCA including the information retrieved from clustering in the space spanned by archetypes. Are the patterns of the compositional space somehow in line with what can be viewed in the factorial maps?

As shown in 5.10, even the PCA factorial map follows similar patterns that have been already highlighted in compositional space. The 1st groups, in red

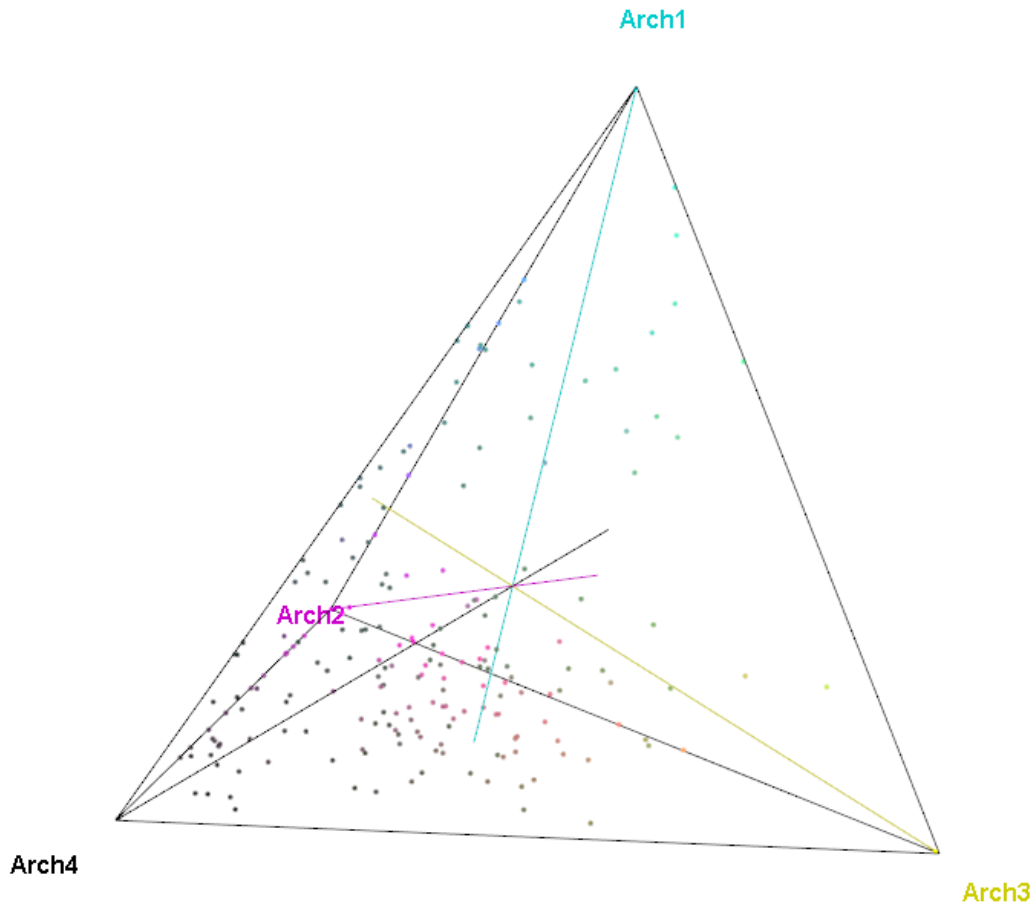


Figure 5.7: Quaternary plot representing the compositional space spanned by 4 archetypes. Points are located and coloured according to coordinates α's.

dots, is made by schools that are closer to the origin of axis, lying between three archetypes and very far only from the 1^{st} archetype. It is consistent with the results shown in the barplot in 5.9.
As said, the right number K of archetypes to be involved in the analysis is not an easy task. For this purpose, let consider the factorial map in which archetypes deriving from archetypal analysis with $K = 3, K = 4$ and $K = 5$ act as supplementary rows. In 5.11, 5.12 and 5.13 is possible to visualize how archetypes tend to be always at the external part of data cloud. Adding the 5^{th} archetypes, it lies in the first quadrant in clockwise order, in the upper-left part. This is an other hint that suggest to keep the additional information provided by additional archetype. Two archetypes, "Arch5" and "Arch4.2", looks on the first 2 dimensions of the PCA in 5.11, to be somehow

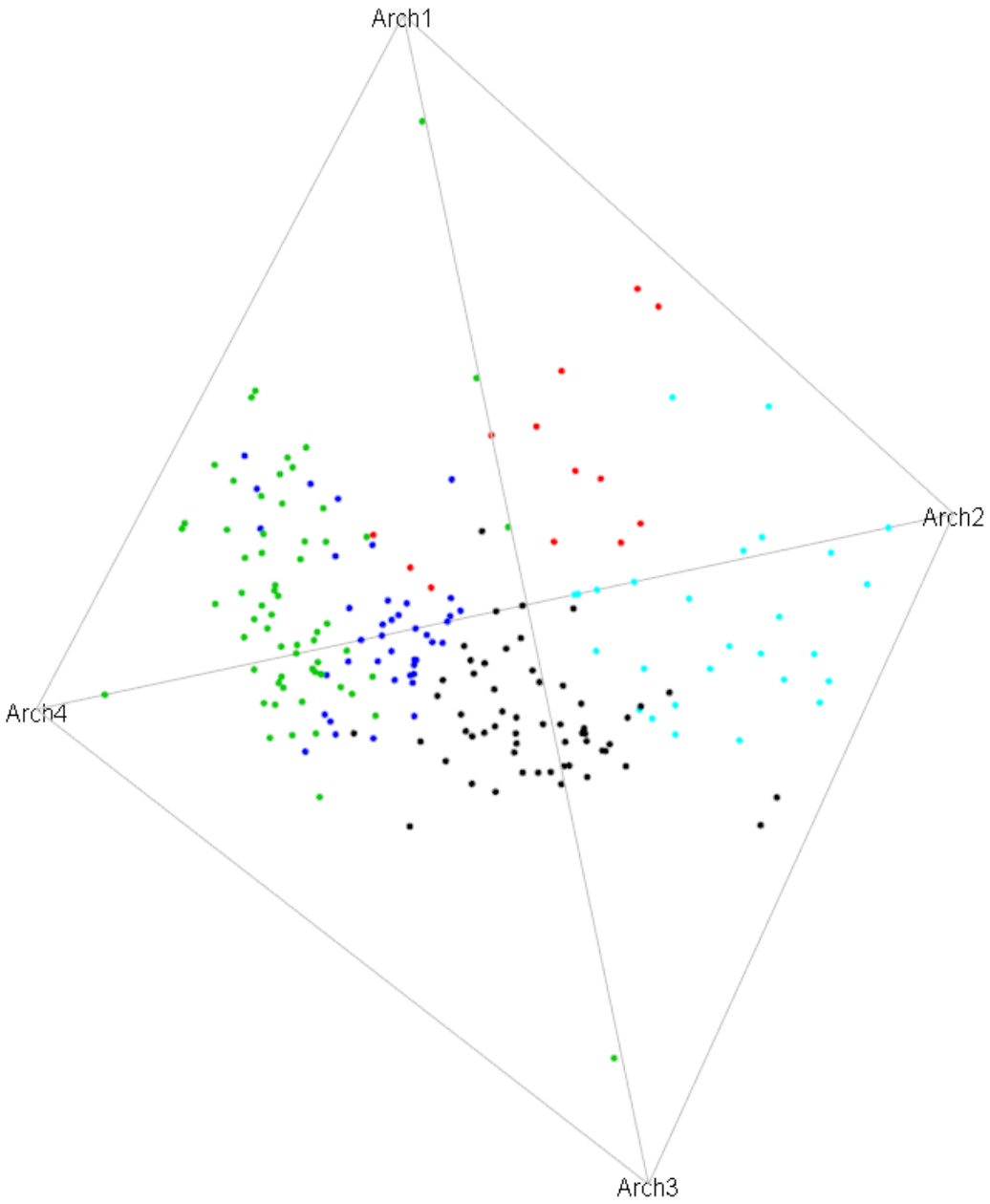


Figure 5.8: Quaternary plot representing the compositional space spanned by 4 archetypes. Points are located according to coordinates α's, and coloured according to clusters found by means of Aitchinson distance.

internal. Looking at their behaviour on the third dimension, so looking at 5.12 and 5.13, they assume a more extreme location. PCA is just a visual tool to have a flavour of the "real position/location" of the archetypes, but

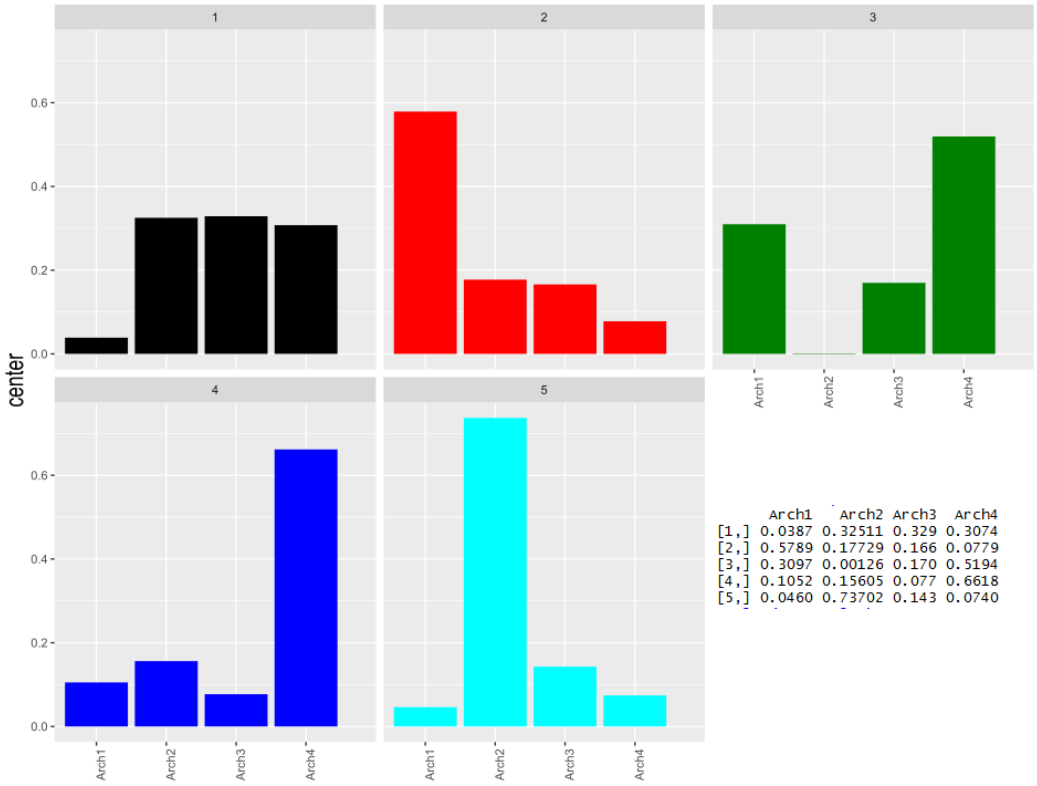


Figure 5.9: Centers of 5 groups identified in space spanned by archetypes. Barplot and real values.

as in the toy example, also in this context they seem to be proper entities to describe extreme patterns in data, cause they are naturally inclined to be at the edges of the cloud, given a pair of dimensions. Given all these reasoning about archetypes role in this context and in this kind of analysis, and given all these different tools, both graphical and analytic, is possible to assume that:

1 Archetypes are well separated and gain informative power, showing peculiar behaviour in histograms' distribution; therefore, they can be used as benchmarking abstract entities.

2 It is possible to consistently using them as reference points to make categorization by means of clustering procedures. The consistency between the space spanned by archetypes and the PCA factorial map is a strong connection that goes in such direction.

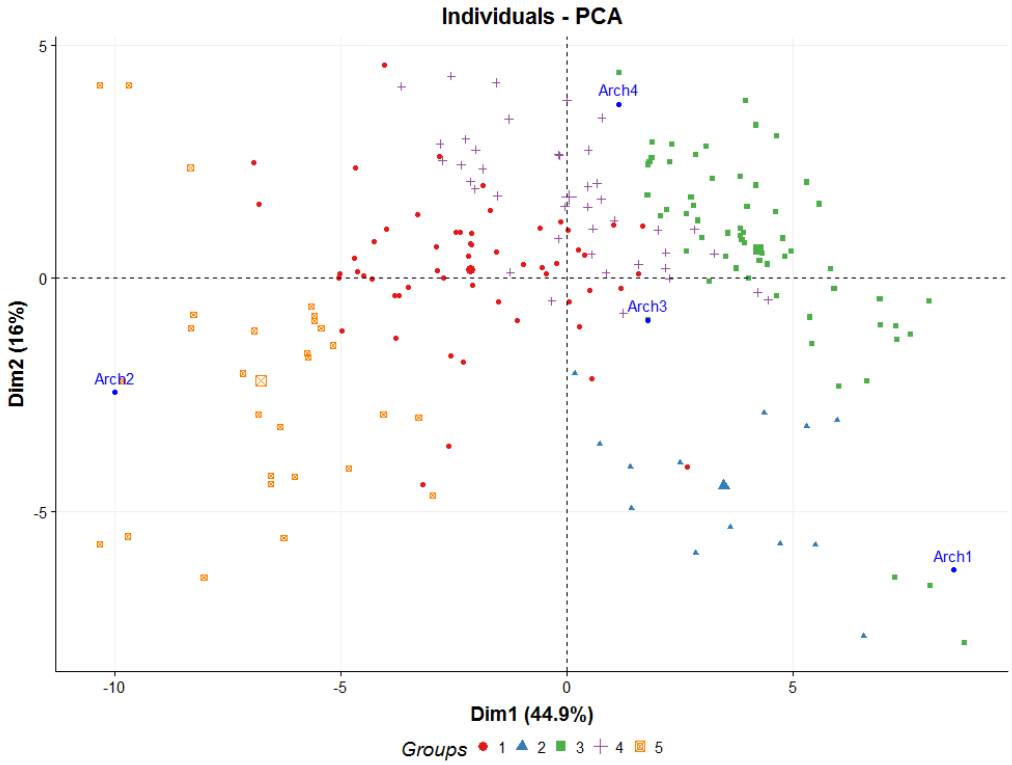


Figure 5.10: PCA with archetypes as supplementary rows: 5 groups of schools according to clusters in space spanned by archetypes

5.3 Working Hypotheses about INVALSI Test: using archetypes

In this last section, the aim is to try to, first of all, understand if some patterns about school characteristics, already discussed in literature, are confirmed or not by the archetypal analysis, and if it is possible to link clusters identified in the compositional space spanned by archetypes to some specific features with respect to schools. Results refer to the case of 3 archetypes identified, as summed up in the table 5.3 with their main descriptive statistics and in their histogram representation in 5.4. In this compositional space, is possible to perform, as said, a clustering, identifying group of schools that are similar in terms of both domains proficiency. The cluster procedure is performed using 8 number of cluster, using the Aitchinson distance as distance measure and without scaling data before the procedure. The size of the identified cluster is presented in the table 5.5. The teal is the bigger, with 43 schools inside it,

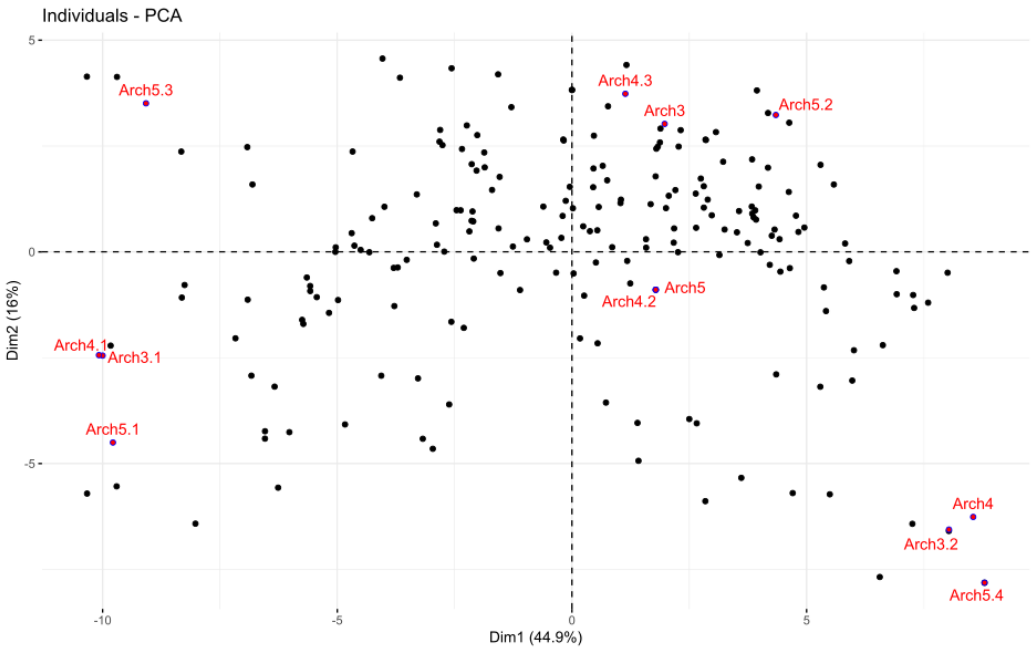


Figure 5.11: PCA with archetypes as supplementary rows: $K = 3, K = 4$ and $K = 5$. First 2 dimensions

while black and gray are the smaller, with only 14 schools each. The position

Table 5.5 Clusters size in compositional space spanned by archetypes, 8 clusters identified

Black	Red	Green	Blue	Teal	Purple	Yellow	Gray
14	15	30	22	43	29	33	14

of clusters in the ternary plot follows clearly the pattern highlighted by the cluster membership. Some groups of schools are very close to the corner of the triangle, and so they are grouped accordingly. They will likely share with the respective archetypes a good amount of similarity in their distributions scores, both for writing/reading and for mathematics. Other clusters are somehow in an intermediate positions with respect to the corners; others are lying at the middle of one external line, and so they have one α that is basically zero, being located between only 2 archetypes, without contribution at all of one archetype. The ternary plot, with schools coloured according to the cluster they belong to, is in 5.14.
Using archetypal coefficients α's as weights, and the 3 archetypes as variables, it is possible to exploit the properties of the Wasserstein distance to reconstruct, as a weighted mean of the archetypes, the centroid of each clus-

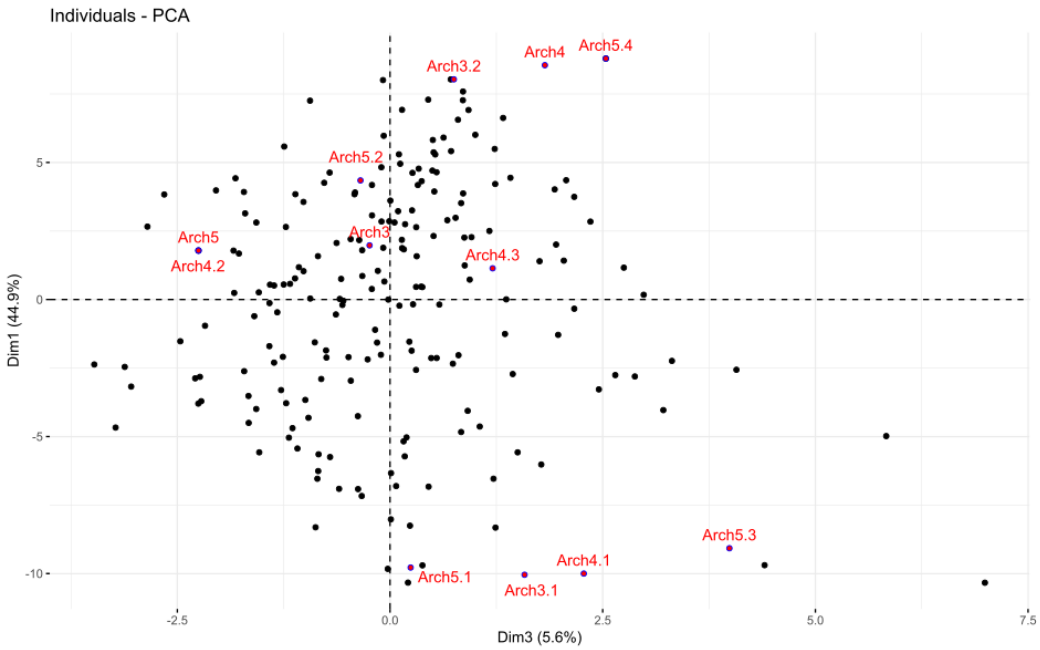


Figure 5.12: PCA with archetypes as supplementary rows: $K = 3, K = 4$ and $K = 5$. Dimensions 1 and 3.

ter, in order to better understand what they represent in terms of bivariate histogram. Centroids of the clusters, as presented in 5.15, are related to very different behaviour. To the extreme, we have the teal group (best performances in both domains) and yellow group (worst performances in both domains). In the middle, a wide range of different distributions.

Moving on to some interesting aspects of INVALSI data and about scholastic achievement literature in general, let's introduce the working hypothesis that will be discussed, trying to answer to what extent is possible to proof or disconfirm them using histogram archetypes analysis in this sample of 200 schools. The strategy is to use information retrieved from original data as an explanatory tool, so that it could be possible to explain why schools sharing a number of factors belong, or not, to the same cluster. To sum up, we will try to deepen the following issues:

- (i) Geographical gap, especially the comparison between South and North.
- (ii) Gender gap.
- (iii) Difference in the type of school (Liceo, Professionale and so on).

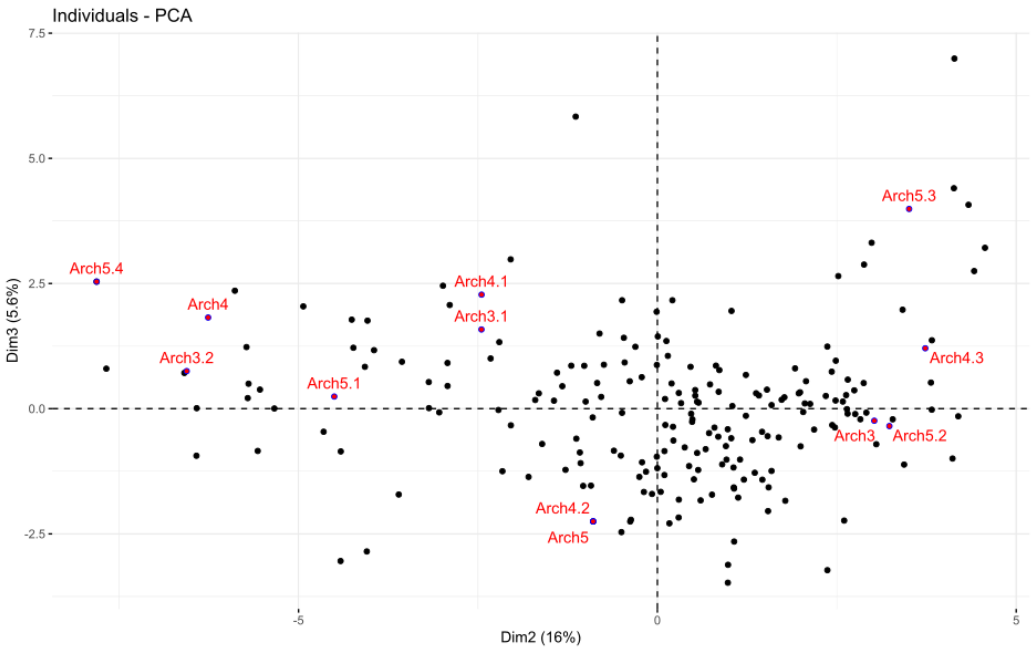


Figure 5.13: PCA with archetypes as supplementary rows: $K = 3, K = 4$ and $K = 5$. Dimensions 2 and 3.

- (iv) The impact of school size, in terms of number of classes, to the proficiency.

Therefore, first item is to discuss if the regional environment, in terms of macro-area, has an influence on the scholastic proficiency. As pointed out by several authors, there is a gap between North and South in proficiency with respect to several domains. While the gap is not significant in early age, it increases showing a peak in the difference around 15 years, when usually students from North outperform students from South, in particular for what concerns mathematics. However, it is possible to classify in a different fashion the Italian 20 Regions. Most common are: in 3 macro-area (North, Center and South) and in 5 (North-West, North-East, Center, South, Islands). Other have proposed to divide it still into 5 classes, but moving Calabria region to the "Islands" group, modifying this group into what is possible to rename "Extreme South", cause not only the big islands of Sicily and Sardinia are included. Crossing this information with the cluster membership of the schools, using 3 macroareas (5.6) and then using 5 macro-areas (5.7),it is possible to state some patterns. Looking at the 3 macroareas, schools from the North are the most present in the "best" group, the teal, while schools from South are the most present in the "worst" group,

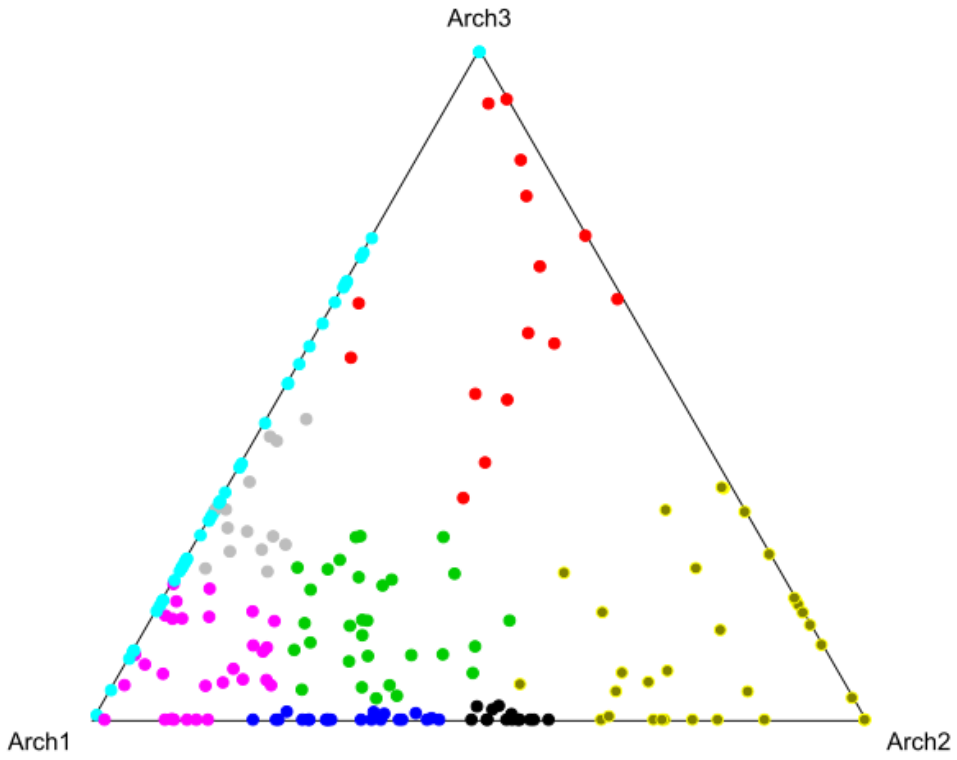


Figure 5.14: Ternary plot representation of space spanned by $K = 3$ archetypes. Two hundreds schools with different colours consistently with the clusters as in the table 5.5

Table 5.6 Contingency table crossing clusters with 3 macro-areas.

	Black	Red	Green	Blue	Teal	Purple	Yellow	Gray
North	3	5	15	6	24	18	7	9
Center	2	3	4	6	8	4	5	2
South	9	7	11	10	11	7	21	3

Table 5.7 Contingency table crossing clusters with 5 macro-areas.

	Black	Red	Green	Blue	Teal	Purple	Yellow	Gray
N-West	1	3	7	2	15	8	3	4
N-East	2	2	8	4	9	10	4	5
Center	2	3	4	6	8	4	5	2
South	3	6	8	5	6	6	7	0
Ex.South	6	1	3	5	5	1	14	3

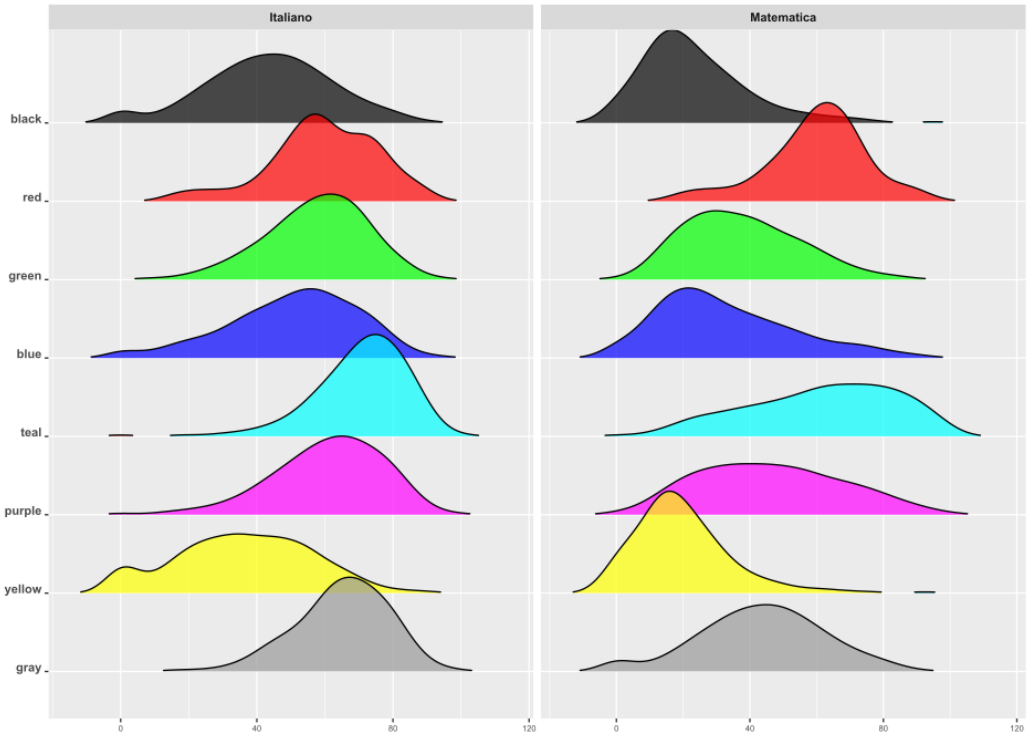


Figure 5.15: Centroid of the 8 clusters identified in the compositional space spanned by the $K = 3$ archetypes in 5.4. Each of them is a weighted mean, using *alpha*'s as weights, of such archetypes.

the yellow. Furthermore, southern schools are way more present in the black group, compared to Center and North, that is the second worst performable in both domains. Schools from Center seem to act in an intermediate role with respect to South and North, being almost equifrequent in the 8 groups. But what happens if we analyze the table in 5.7, so using a higher level of geographical detail? Yellow group is now a prerogative of Sardinia, Sicily and Calabria, being 3 out of 20 regions but accounting for almost the half of the schools in that cluster. Same happens in black group. North-West is outperforming North-East, according to the relative frequency in teal group. South shows a very similar behaviour to the Center in terms of distribution among clusters. Green cluster, characterized by a strong gap between reading/writing (average) and mathematics (below the average), has 8 schools from South and 8 schools from North-East, highlighting that mathematics skills are likely the most dangerous threat to achieve overall good proficiencies.
For what concerns gender gap, given that we are deepening schools and not

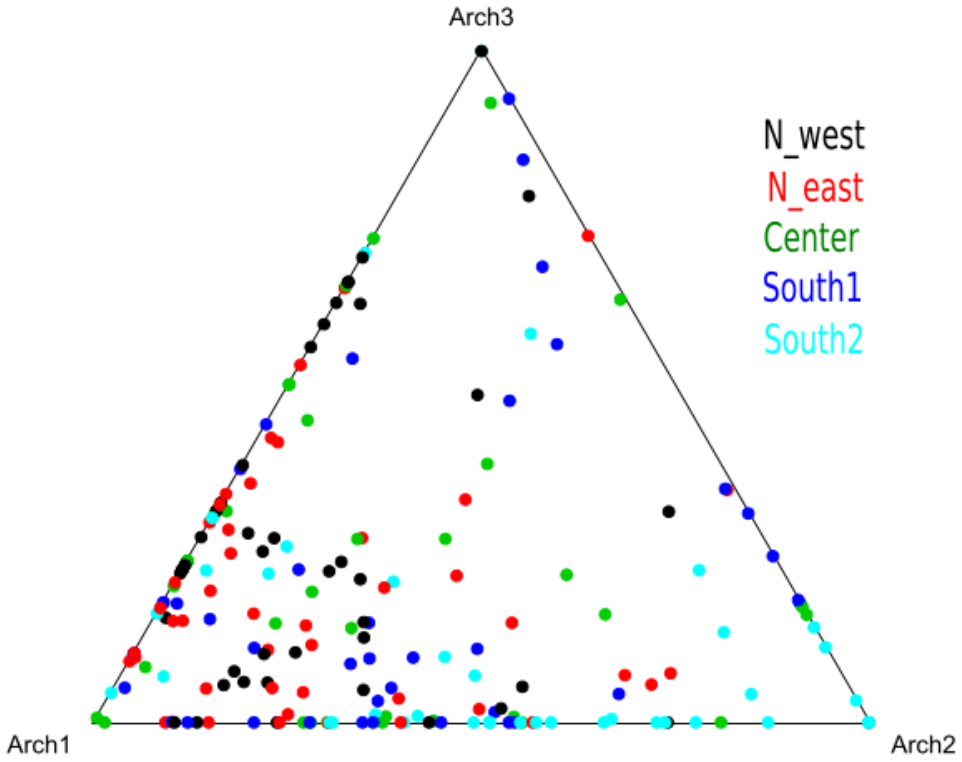


Figure 5.16: Ternary plot representation of space spanned by $K = 3$ archetypes. Two hundreds schools with different colours, according to the 5 macro-areas. Teal, extreme South, mostly close to the worst archetype (the second). Black schools, North-West, in the "optimal external line" between first and third archetype.

individuals, what is possible to do is to use an index, computed for each school, that is the proportion of female on the total. So, school with a value over 0.5 have more females than males, and school with a value below 0.5 have more males than females. Several studies have discussed the discrepancy in proficiencies between females, doing better in reading/writing, and males, doing better in mathematics.

From the ternary plot in 5.17, schools with a majority of males are pretty close to the worst performable corner, the second archetypes, and on the "optimal external line" between first and third archetype. This behaviour likely means that most of the average performances are achieved in schools with an equal proportion of males and females or with a majority of females,

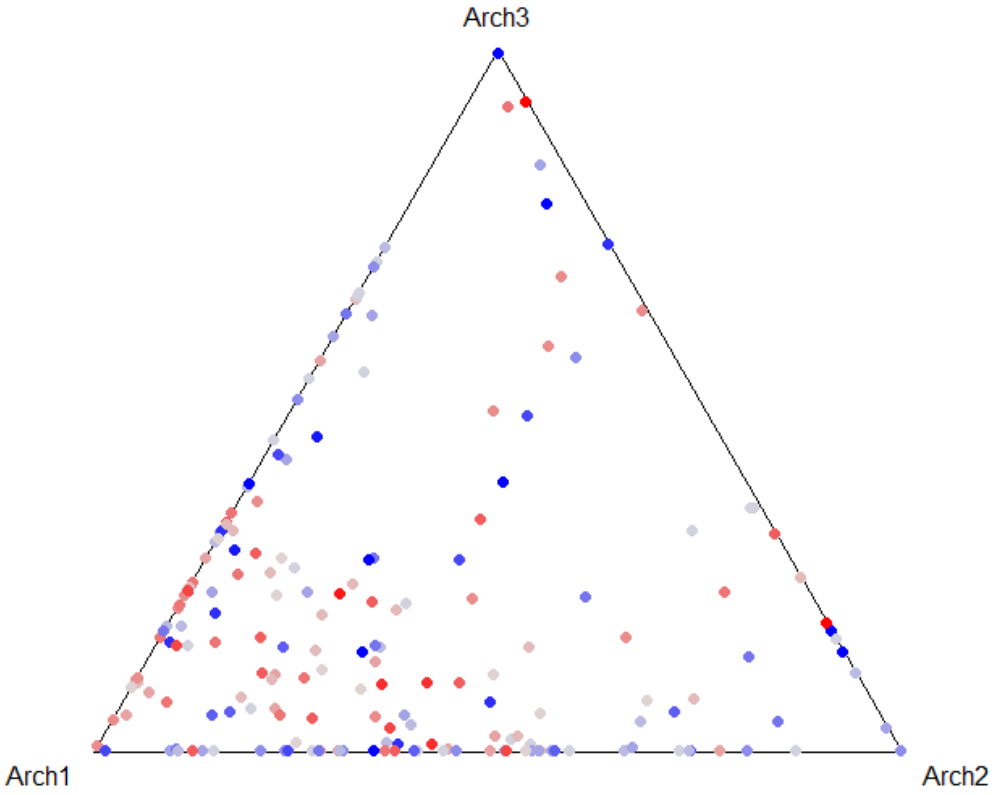


Figure 5.17: Ternary plot representation of space spanned by $K = 3$ archetypes. Two hundreds schools with different colours, from blue to red, according to the percentage of females in that school (the more red, the more females).

and schools with more males are or very good or very bad. As shown in the table with the mean values of the female proportion (5.8), the highest relative number of females are in the green cluster, that is not a bad performer but with a very high gap between reading/writing and mathematics, highlighting how the gap in this domain is somehow confirmed. However, overall, the mean proportion in the best group, the teal, is of 52% females and 48% males, and in the worst group, the yellow, males are 55% of the total on the average. For what concerns different type of schools, some issues arise in the definition of what is the proper classification of Italian high-schools. Some schools have all the classes attending the same programme, but others, often called "Istituti Comprensivi", are able to provide different programmes for different pupils in several classes. For the purpose of this

Table 5.8 Mean value of female proportion for each cluster.

Clusters	ProportionFemale
Black	0.50
Red	0.48
Green	0.54
Blue	0.45
Teal	0.52
Purple	0.51
Yellow	0.45
Gray	0.42

work, the following distinction between school types will be made. "Liceo", where the education received is mostly theoretical and not aimed to people who wants to get a job just after the high school. There is a specialization in a specific field of studies (humanities, science, or art), and the general aim is to prepare students to university and higher education rather than introduce them into a professional position. "Liceo" can be seen as the Italian equivalent of University-preparatory school. "Istituto Tecnico" that provide some theoretical education but mostly a highly qualified technical specialization in a specific field of studies (e.g.: economy, technology, humanities, law, accountancy, administration, tourism, information). Most of the time, it is aimed to give a few months intern-ship in a company, association or university, during the last year or the second last year of study. "Istituto Professionale" is specifically structured only for practical activities. It is not aimed to theoretical knowledge, but rather to facilitate the direct entry of the pupil to the labour market. When more than one programme is available, in the following we will assume that the schools is "Mixed". As can be seen in 5.18, most of the black schools (Licei) are on the "optimal external line" between first and third archetype. Most of the schools lying somehow close to the second archetypes are Green, so Professionali; and Tecnici, in Blue, are spread all over the ternary plot. Even if Mixed schools, in Red, are not defined as schools with a clear and unambiguous programme, they are located as a cluster in the left-down side corner. From the graphic display and from the contingency table 5.9, some patterns are clear. The best performer, teal group, are basically all Liceo-type school. Mixed-type schools are basically restricted to the Blue and the Purple groups, clusters with an average behaviour. So, the red cluster of mixed schools in the ternary plot in 5.18, highlights how this type of schools are in an intermediate position between Licei and Professionali, the worst performer. More than half of the Professionali are located, indeed, in the Yellow cluster. Most of the Tecnici

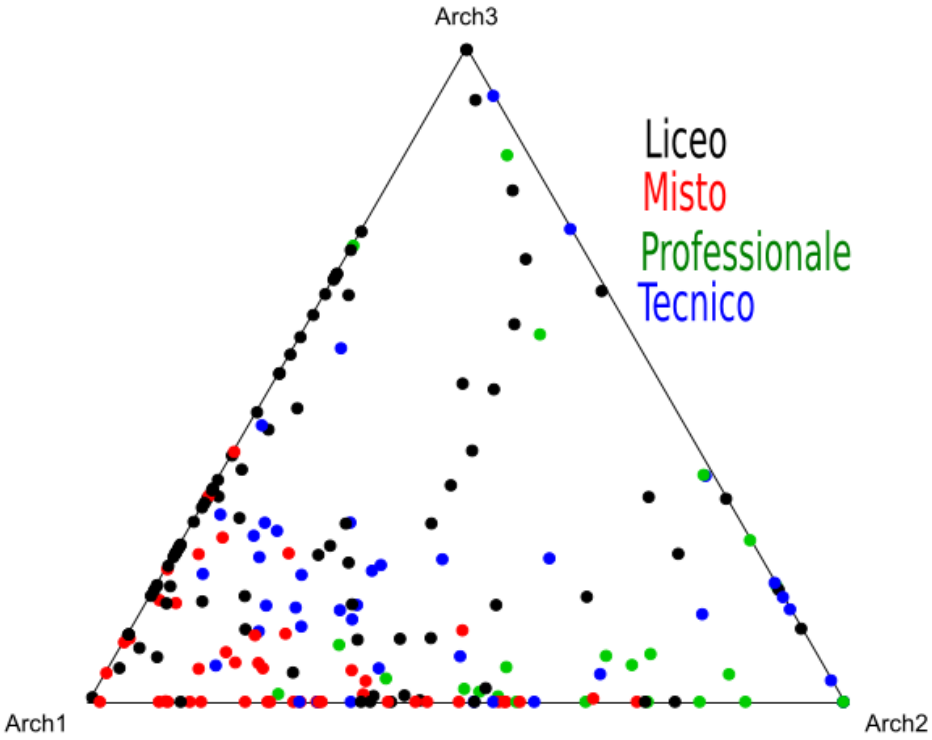


Figure 5.18: Ternary plot representation of space spanned by $K = 3$ archetypes. Two hundreds schools with different colours, according to the different type of programme. Licei in black, Professionali in green, Tecnici in blue, Mixed in red.

are in the Green cluster, the one with a very high difference between writing/reading and mathematics.
Last issue is related to the school size. A good proxy is the number of classes involved in the analysis. As said, the standard procedure of INVALSI is to test all the classes belonging to the school, so on the average this information is a valid approximation of the school size. Are big schools better than small schools? Many studies involve statements about school population. Some authors proof that having a daily and deeper touch with teachers, as it happens in Small schools, improve student's skills, while other says that only big schools with heterogeneity in terms of pupils and teachers is able to provide good proficiencies. Looking at the ternary plot in 5.19, most of the schools on the "optimal external line" between first and third archetype are either Green or Red, so they are Medium sized or Big sized. A lot of Small-sized schools are, on the other hand, located to the right corner, close to the second

Table 5.9 Contingency table of type of school and clusters.

	Black	Red	Green	Blue	Teal	Purple	Yellow	Gray
LICEO	2	10	11	6	39	9	9	5
MISTO	5	0	5	12	3	16	2	3
PROFESSIONALE	4	2	2	1	1	0	13	0
TECNICO	3	3	12	3	0	4	9	6

archetype. Looking at the contingency table in 5.10, it is notable that Small

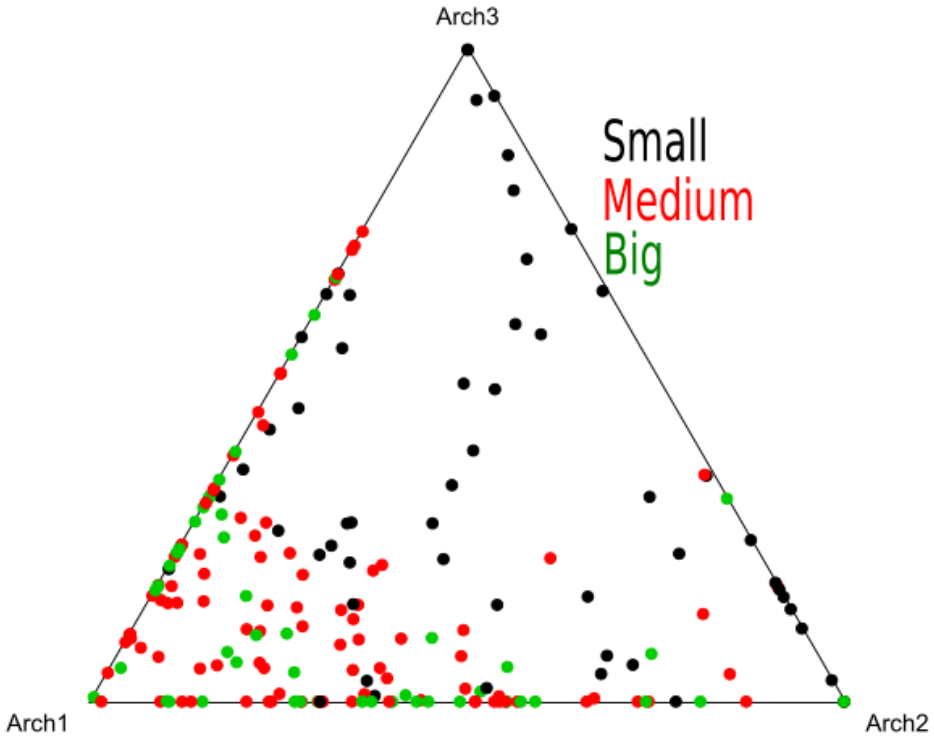


Figure 5.19: Ternary plot representation of space spanned by $K = 3$ archetypes. Two hundreds schools with different colours, according to the different school size. Small schools in black, Medium size schools in red, Big schools in green.

schools are split in only 3 groups: Red, Green and Yellow. Worst performers are so small schools, while the Teal group, with the best schools, is made up Medium and Big schools for the majority. Red clusters, only made up by 15 Small schools, is a group of schools with similar distribution scores for both domains, and so it is just average for reading/writing domain but is the second best performer for what concerns mathematics. It seems that Small

schools are not the best, but are at least suffering not so dramatically by the mathematics threat. Coming back to the four original research questions,

Table 5.10 Contingency table of school sizes and clusters.

	Black	Red	Green	Blue	Teal	Purple	Yellow	Gray
Small	1	15	10	2	5	1	17	5
Medium	8	0	17	11	21	20	10	6
Big	5	0	3	9	17	8	6	3

main findings are:

- (i) Geographical gap is still an issue. But North is not so homogeneous. North-West is able to reach better peak than Nort-East. In the South, we can observe two completely different behaviour: one for "Extreme South", and one for the "South", that is very similar to the performances in Center Italy.
- (ii) Gender gap is still an issue, and mathematics is a threat for females. By the way, the pattern is not so clear, given that the worst performers are for the majority schools with more males than females.
- (iii) Difference in the type of school. Licei are by far the best, followed by Mixed schools. Tecnici have issues with mathematics skills, while Professionali are the worst overall.
- (iv) The impact of school size: small schools are more likely to be in the worst performers group, but compared to other "weak categories", such as Professionali or females or Southern schools, Small schools tend less to be involved in very poor scores for what concerns mathematics. Medium and Big schools don't show an high level of differentiation.

To sum up, to profile the pure type of schools in both cases of best and worst scenario, the best schools are big-sized Licei, especially from North-West, while worst are Professionali from Extreme South with a majority of males.

Conclusions and further developments

In this work it has been presented a first, at least to the best of our knowledge, attempt to define archetypes for histogram-valued data. Defining archetypes for complex data has always been a non-trivial operation. While for simpler data, like points, most of the properties can be pretty easily proofed by means of analytical/mathematical tools, for complex data this procedure becomes harder and harder. For this reason, in this context, the focus has been indeed on the definition of the archetypes for histogram-valued data and on the algorithm developed to figure out them given the constraints and the function to be optimized. Thus, most of the findings and of the discussion are derived directly from empirical evidences rather than mathematical proofs. In 4 section, as well as in 5, notions about the algorithm structure, the pseudocode and its outcomes have been shown. Histogram archetypes, consistently with archetypes identified for data of different natures, seem to act as entities with an extreme/peculiar behaviour. This is confirmed using different tools of graphic displays, and exploiting properties of other statistical techniques in comparison with archetypes.
Reconstructing archetypes, as for the toy data about France regions proposed in 4.2, is the first way to visualize what they represent in terms of histogram object. In the histograms plots drawn in 4.5, for example, the 2^{nd} archetype represents two distributions shifted to the minimum cost for both variables; the 1^{st} archetype represent a distribution with a very unusual long right-tail for the second variable, and this is an other sign that this histogram archetype is external to the data cloud. It should be recalled that, with this kind of data, one of the aim is to deal with units that are made up by a complex internal structure and to exploit analyses able to keep such structure. The assumption is that it is not worthy to reduce the units' description using only one index, like the mean value or the median value, for each complex unit. Therefore, describing what histogram archetypes represent, should be made taking in account the relative outcoming distribution in all its aspect,

using even several indexes to summarise it, like kurtosis, skewness and so on. If the aim is to not reduce complexity but to keep it, the results have to be deepen consistently with this reasoning.
Statistical learning techniques able to identify salient entities found, at the end, abstract units that share the same nature of the original data, and this is the case of histogram archetypes. While from empirical evidences the technique seems to work properly and to keep the promises about usual archetypes location and functionality, more issues are related to the algorithm. Giving it enough time, it is able to solidly identify archetypes, without shifting their position in an unstable fashion, and achieving the tolerance about constraints. The full algorithm, presented as pseudocode in 4.11 and as MATLAB code in the appendix A, is satisfying in terms of both flexibility (different number of data and of archetypes) and ability to accomplish the final results (constraints about α's and β's are always fullfilled, in the experiments presented in this work).
The shortcomings are about the computational cost, that is much higher in comparison with other similar techniques for histogram-valued data, like k-means proposed in 2.70. The required time to accomplish the algorithm increases according to several conditions: if the number of quantiles increases, if the overall number of observations increases and if the K number of archetypes increases.
A future development will be, to solve this issue, first of all the implementation of an algorithm in a simpler programming language, given that MATLAB is still a meta-language, and this increases the computational cost. Most of the commands used in the code of MATLAB, roughly around 90%, are written in C++ (if they are not written directly in MATLAB). Some `Perl` and `Java` functions are used as well, but the best choice is likely to use C++ coding to improve computational cost of the algorithm. It has to be underlined that the whole procedure is nested inside the BigData approach framework, so it is reasonable that this technique could be applied to large dataset and finding a way to improve the algorithm speed is a crucial step in the next contributions, to make full use of the histogram archetypes analysis results. Other possible ways to improve such speed could be implementing a different design, using for example "Monte Carlo" techniques, introducing so a sampling step in the data rather than slavishly using every available data point. Some improvements should be mad in the algorithm choice itself.
One of the core concept in this work is that the complexity of the data is, in this approach, also an opportunity in terms of findings interpretation, as well as a challenge in terms of algorithmic procedure. For this reasons, the SDA paradigm allows to better deal with non-standard units and preserve such complexity. In the presented implementation, the insight has been to use

the relationships between interval-valued data and histogram-valued data in SDA, extending so some proofed findings with respect to interval-valued data to histogram-valued data case. This is consistent with the choice of distances derived from Wasserstein distance, and the general conceive of histogram objects in terms of centers and radii notation. But, in the next, a development could be to implement of a more complex and structured symbolic notation. These idea come straight from the application of the techniques to the real case of benchmarking in Italian school system. Therefore, in that context, a full description of the units could be made adding several school fixed factors, alongside its analytical part that has already lead to the presented histogram objects description. In this new approach, that could be exhaustive, even if particularly sophisticated, a unit is described by its score histogram, as well as by several additional features, such as its geographic location, its programme, its dimension/size and so on.
The advantage to work in SDA framework is that it allows to use much more many elements to describe units and category, in comparison with traditional approach. Schools seem to be notably suited to be described and analyzed using as many factors as possible, given the complex and hierarchical structure that is at the ground of their nature. Due to the nested structure they always show (pupils nested in classes, classes nested in school complexes, school complexes nested in schools, schools nested and in municipalities and so on), an other future focus could be likely to think about a SDA specific approach, able to identify archetypes linked to this hierarchical structure. This analysis could lead to define "multilevel archetypes", as abstract entities acting at different level of the same data, highlighting in each level extreme behaviours, but allowing for eventual interaction effects between layers, keeping so a sort of algebraic linkage between levels belonging to the same context of interest.
These new opportunities are a clue that the technique of histogram-valued data archetypes identification has still some open space ahead, given the wide range of options embedded in symbolic objects definition and the increasing number of fields of application of innovative methodologies linked to a Big Data framework.

Appendices

Appendix A

Appendix - Matlab Code

```
% initial constraints
function [c,ceq] = constr(v,X,m,K)

ceqalpha = zeros(1,m);
for i=1:m
        for j=0:p-1
        ceqalpha(i)= ceqalpha(i)+v(i+j*m);
    end
    ceqalpha(i)= ceqalpha(i)-1;
end

ceqbeta = zeros(1,K);
for i=1:K
        for j=0:m-1
                ceqbeta(i)= ceqbeta(i)+v(K*m+i+j*K);
    end
    ceqbeta(i)= ceqbeta(i)-1;
end


ceq=[ ceqalpha' ; ceqbeta'];

% for i=1:K
%       for j=1:m
%               alpha(i,j)=v(i+(j-1)*K);
%       end
% end
```

```
% for i=1:m
%       for j=1:K
%               beta(i,j)=v(K*m+i+(j-1)*m);
%       end
% end

%definition of the histogram archetypes function

function f = hist_fun(v,X,m,n,K)


%starting from v and reconstruct alpha and beta
   matrices
for i=1:m
        for j=1:K
                alpha(i,j)=v(i+(j-1)*m);
        end
end

for i=1:p
        for j=1:m
                beta(i,j)=v(K*m+i+(j-1)*K);
        end
end

A = alpha;
B = beta;
XBAr = abs(A*B)*X(:,n+1:2*n);
XBAc = A*B*X(:,1:n);
XBA = [XBAc XBAr];
%defining function to optimize
mat = (X(:,1:n)-XBAc).^2+(1/3)*(X(:,n+1:2*n)-XBAr)
   .^2;
% frobenius norm
f = norm(mat,'fro');



function [alphas,betas,Xric_c,Xric_r,fvalue,
   Archetypes_c,Archetypes_r]=Archetypes_hist(X,K)
```

```
%
% [alphas,betas,Xric_c,Xric_r,fvalue,Archetypes,
   Archetypes_r]=Archetypes_hist(X,K)
%
%
% INPUT
% X = data matrix , units * variables, of dimension
   m * n;
%   in first n/2 columns there are bins centers,
%   in last n/2 columns there are bins radii
%
% K=fixed number of archetypes
%
% OUTPUT
% alpha's and beta's coefficients, archetypes
   centers and radii,
% reconstruction of original X centers and radii
   given archetypes, fvalue is  %objective function
   final result


m = size(X,1);
n2 = size(X,2);
n = n2/2;


v0=rand(2*p*m,1);

lb=zeros(2*p*m,1);
ub=ones(2*p*m,1);

options=optimset('MaxFunEvals', 1000000, 'MaxIter',
   100000,'TolCon',1e-6,'TolFun',1e-6,'Display','
   iter');
% 'Display','iter' used to see iterations, not
   necessary
% constrained optimization
[v,fval,~,~] =fmincon(@(v)hist_fun(v,X,m,n,p),v0
   ,[],[],[],[],lb,ub,...
    @(v)constr(v,X,m,K),options);
```

```
for i=1:m
        for j=1:K
                alpha(i,j)=v(i+(j-1)*m);
        end
end
sum_alpha=sum(alpha);
for i=1:K
        for j=1:m
                beta(i,j)=v(K*m+i+(j-1)*K);
        end
end

% beta's used to build archetypes centers and radii,
    alpha's to build original data

Archetypes_c = beta*X(:,1:n);
Archetypes_r = beta*X(:,n+1:2*n);
Xric_c = alpha*Archetypes_c;
Xric_r = alpha*Archetypes_r;

end
```

Bibliography

Abbasi, Ahmed, Hassan, Ammar, and Dhar, Milan (2014). "Benchmarking Twitter Sentiment Analysis Tools." In: *LREC*. Vol. 14, pp. 26–31.

Aitchison, John (1982). "The statistical analysis of compositional data." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2, pp. 139–177. DOI: 10.1016/S0165-0114(02)00246-4. URL: http://www.jstor.org/stable/2345821.

Aitchison, John, Barceló-Vidal, Carles, Martín-Fernández, Josep Antoni, and Pawlowsky-Glahn, Vera (2000). "Logratio analysis and compositional distance". In: *Mathematical Geology* 32.3, pp. 271–275.

Anderson, Chris (2008). *The end of theory: The data deluge makes the scientific method obsolete. Wired, 23 June 2008.*

Azrieli, Yaron and Lehrer, Ehud (2007). "Categorization generated by extended prototypes—An axiomatic approach". In: *Journal of Mathematical Psychology* 51.1, pp. 14–28.

Barnes, Trevor J (2013). "Big data, little history". In: *Dialogues in Human Geography* 3.3, pp. 297–302.

Bauckhage, Christian and Thurau, Christian (2009). "Making archetypal analysis practical". In: *Joint Pattern Recognition Symposium*. Springer, pp. 272–281.

Belohlavek, Radim (2008). "Introduction to formal concept analysis". In: *Palacky University, Department of Computer Science, Olomouc*, p. 47.

Berk, Richard A. (2016). *Statistical Learning from a Regression Perspective*. 2nd ed. New York: Springer. ISBN: 978-3-319-44047-7.

Bertrand, Patrice and Goupil, Francoise (2012). "Descriptive statistics for symbolic data". In: *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Ed. by Hans-Hermann Bock and Edwin Diday. Berlin, Germany: Springer Science and Business Media, pp. 103–124.

Bezdek, James C., Ehrlich, Robert, and Full, William (1984). "FCM: The fuzzy c-means clustering algorithm". In: *Computers & Geosciences* 10.2-3, pp. 191–203.

Bien, Jacob and Tibshirani, Robert (2011a). "Hierarchical clustering with prototypes via minimax linkage". In: *Journal of the American Statistical Association* 106.495, pp. 1075–1084.

— (2011b). "Prototype selection for interpretable classification". In: *The Annals of Applied Statistics*, pp. 2403–2424.

Billard, Lynne and Diday, Edwin (2003). "From the statistics of data to the statistics of knowledge: symbolic data analysis". In: *Journal of the American Statistical Association* 98.462, pp. 470–487.

— (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. 1st ed. Chichester,West Sussex, UK: John Wiley and Sons. ISBN: 978-0-470-09016-9.

Billard, Lynne and Kim, Jaejik (2017). "Hierarchical clustering for histogram data". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 9.5, e1405.

Binder, Mario, Clegg, Ben, and Egel-Hess, Wolfgang (2006). "Achieving internal process benchmarking: guidance from BASF". In: *Benchmarking: An International Journal* 13.6, pp. 662–687.

Bock, H. and Diday, Edwin (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information From Complex Data*. 1st ed. Berlin Germany: Springer. ISBN: 978-3-540-66619-6.

Borgelt, Christian (2006). "Prototype-based classification and clustering". PhD thesis. Otto-von-Guericke-Universität Magdeburg, Universitätsbibliothek.

Bousquet, Olivier, Boucheron, Stephane, and Lugosi, Gabor (2004). "Introduction to statistical learning theory". In: *Advanced Lectures on Machine Learning. Lecture Notes in Computer Science* 3176, pp. 169–207. DOI: https://doi.org/10.1007/978-3-540-28650-9_8.

Boyd, Danah and Crawford, Kate (2012). "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon". In: *Information, communication & society* 15.5, pp. 662–679.

Brown, Michael (2011). "Learning analytics: The coming third wave (EDUCAUSE Learning Initiative Brief)". In: *Retrieved from EDUCAUSE library https://net. educause. edu/ir/library/pdf/ELIB1101. pdf*.

Brown, Stephen W. and Swartz, Teresa A. (1989). "A gap analysis of professional service quality". In: *The Journal of Marketing*, pp. 92–98.

Camp, Robert C. (1989). "Benchmarking-The Search for Best Practices that Lead to Superior Performance". In: *Quality Progress* 22.2, pp. 70–75.

— (1995). *Business process benchmarking: finding and implementing best practices*. Vol. 177. ASQC Quality Press Milwaukee, WI.

Capperucci, Davide (2017). "Prove del Servizio nazionale di valutazione e apprendimento della matematica: migliorare le performance della scuola

primaria a partire dai risultati". In: *Studi sulla Formazione* 20.1, pp. 43–67.

Cha, Sung-Hyuk (2007). "Comprehensive survey on distance/similarity measures between probability density functions". In: *International Journal of Mathematica Models and Methods in Applied Sciences* 1.4, pp. 300–307.

Chan, Ben H.P., Mitchell, Daniel A., and Cram, Lawrence E. (2003). "Archetypal analysis of galaxy spectra". In: *Monthly Notices of the Royal Astronomical Society* 338.3, pp. 790–795.

Chapelle, O., Schölkopf, B., and Zien, A. (2009). *Semi-Supervised Learning*. 3rd ed. Cambridge, MA; London, UK: MIT Press. ISBN: 978-0-262-03358-9.

Chavent, Marie (1998). "A monothetic clustering method". In: *Pattern Recognition Letters* 19.11, pp. 989–996.

— (2000). "Criterion-based divisive clustering for symbolic data". In: *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, pp. 299–311.

Chen, Songcan and Zhang, Daoqiang (2004). "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34.4, pp. 1907–1916.

Conover, William Jay (1999). *Practical nonparametric statistics*. 3rd ed. New York: Wiley.

Corsaro, Stefania and Marino, Marina (2010). "Archetypal Analysis of Interval Data". In: *Reliable Computing* 14.1, pp. 105–116.

Costantini, Paola, Porzio, Giovanni C., Ragozini, Giancarlo, and Romo, Juan (2012). "Archetypal functions". In: *Analysis and Modeling of Complex Data in Behavioural and Social Sciences. JCS CLADAG, Anacapri, Italy*, pp. 1–4.

Costanzo, Antonella and Desimoni, Marta (2017). "Oltre l'effetto "in media": uno studio sulle prestazioni degli studenti nei test INVALSI utilizzando l'approccio quantile. Thinking beyond the "average case": exploring students' performance in INVALSI test through a quantile regression perspective". In: *I dati INVALSI: uno strumento per la ricerca*, pp. 185–197.

Courrieu, Pierre (2008). "Fast computation of Moore-Penrose inverse matrices". In: *Neural Inf. Process.—Lett. Rev.* 8.2, pp. 25–29.

Cowan, Nelson (2010). "The magical mystery four: How is working memory capacity limited, and why?" In: *Current directions in psychological science* 19.1, pp. 51–57.

Cukier, Kenneth (2010). *Data, data everywhere: A special report on managing information*. Economist Newspaper.

Cutler, Adele and Breiman, Leo (1994). "Archetypal analysis". In: *Technometrics* 36.4, pp. 338–347.

D'Esposito, Maria Rosaria, Palumbo, Francesco, and Ragozini, Giancarlo (2006). "Archetypal analysis for interval data in marketing research". In: *Ital. J. Appl. Stat* 18, pp. 343–358.

— (2011). "On the use of archetypes and interval coding in sensory analysis". In: *Classification and Multivariate Analysis for Complex Data Structures*. Springer, pp. 353–361.

Dagan, Ido, Lee, Lillian, and Pereira, Fernando (1997). "Similarity-based methods for word sense disambiguation". In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 56–63.

deCarvalho, Francisco (1994). "Proximity Coefficients between Boolean symbolic objects". In: *New Approaches in Classification and Data Analysis*. Ed. by Edwin Diday, Yves Lechevallier, Martin Schader, Patrice Bertrand, and Bernard Burtschy. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 387–394. ISBN: 978-3-642-51175-2.

— (1998). "Extension based proximities between constrained Boolean symbolic objects". In: *Data Science, Classification, and Related Methods*. Ed. by Chikio Hayashi, Keiji Yajima, Hans-Herman Bock, Noboru Oshumi, Yutaka Tanaka, and Yasumasa Bada. Tokyo: Springer Japan, pp. 370–378. ISBN: 978-4-431-65950-1.

Deza, Michel-Marie and Deza, Elena (2006). *Dictionary of distances*. Elsevier.

Diday, Edwin (1971). "Une nouvelle méthode en classification automatique et reconnaissance des formes: la méthode des nuées dynamiques". In: *Revue de statistique appliquée* 19.2, pp. 19–33.

Diday, Edwin and Esposito, Floriana (2003). "An introduction to symbolic data analysis and the SODAS software". In: *Intelligent Data Analysis* 7.6, pp. 583–601.

Diday, Edwin and Simon, J. Claude (1976). *Digital Pattern Classification*.

Dodd, Edward L. (1940). "The Substitutive Mean and Certain Subclasses of this General Mean". In: *Annals of Mathematical Statistics* 11.2, pp. 163–176. URL: https://www.jstor.org/stable/i312722.

Domingos, P. (2012). "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10, pp. 78–87. DOI: 10.1145/2347736.2347755.

Drake, Jonathan and Hamerly, Greg (2012). "Accelerated k-means with adaptive distance bounds". In: *5th NIPS workshop on optimization for machine learning*, pp. 42–53.

Dubey, Rameshwar, Gunasekaran, Angappa, Childe, Stephen J., Papadopoulos, Thanos, Luo, Zongwei, Wamba, Samuel Fosso, and Roubaud, David (2017). "Can big data and predictive analytics improve social and environmental sustainability?" In: *Technological Forecasting and Social Change*.

Dunn, Joseph C. (1973). "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters". In:

Edelsbrunner, Herbert and Seidel, Raimund (1986). "Voronoi diagrams and arrangements". In: *Discrete & Computational Geometry* 1.1, pp. 25–44.

Elder, A. and Pinnel, J. (2003). "Archetypal analysis: an alternative approach to finding defining segments". In: *2003 Sawtooth Software Conference Proceedings*, pp. 113–129.

Epifanio, Irene (2016). "Functional archetype and archetypoid analysis". In: *Computational Statistics & Data Analysis* 104, pp. 24–34.

Epifanio, Irene, Vinué, G, and Alemany, Sandra (2013). "Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem". In: *Computers & Industrial Engineering* 64.3, pp. 757–765.

Eugster, Manuel J. A. (2012). "Performance profiles based on archetypal athletes". In: *International Journal of Performance Analysis in Sport* 12.1, pp. 166–187.

Eugster, Manuel J. A. and Leisch, Friedrich (2009a). "From Spider-Man to Hero – Archetypal Analsis in R". In: *Journal of Statistical Software* 30.8, pp. 1–23. URL: http://www.jstatsoft.org/v30/i08/.

— (2009b). "From spider-man to hero-archetypal analysis in R". In:

Eyrich, H.G. (1991). "Benchmarking to become the best of breed". In: *Manufacturing Systems* 9.4, pp. 40–47.

Fayyad, Usama, Piatetsky-Shapiro, Gregory, and Smyth, Padhraic (1996). "From data mining to knowledge discovery in databases". In: *AI magazine* 17.3, pp. 1–37.

Finesso, Lorenzo and Spreij, Peter (2004). "Approximate nonnegative matrix factorization via alternating minimization". In: *arXiv preprint math/0402229*.

Fletcher, Roger (2013). *Practical methods of optimization*. John Wiley & Sons.

Fondazione Giovanni Agnelli (2014). *La valutazione della scuola: A che cosa serve e perché è necessaria all'Italia*.

Franke, Beate, Plante, Jean-François, Roscher, Ribana, Lee, En-shiun Annie, Smyth, Cathal, Hatefi, Armin, Chen, Fuqi, Gil, Einat, Schwing, Alexander, Selvitella, Alessandro, et al. (2016). "Statistical inference, learning and models in big data". In: *International Statistical Review* 84.3, pp. 371–389.

Frühwirth-Schnatter, Sylvia (2006). *Finite mixture and Markov switching models*. 1st ed. berlin, Germany: Springer Science and Business Media. ISBN: 978-0387-32909-3.

Ganter, Bernhard and Wille, Rudolf (1996). *Formale BegrifJsanalyse. Mathematische Grundlagen*. 1st ed. Heidelberg, Germany: Springer–Verlag. ISBN: 978-3-540-60868-4.

Gärdenfors, Peter (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Gavin, Daniel G., Oswald, Wyatt W., Wahl, Eugene R., and Williams, John W. (2003). "A statistical approach to evaluating distance metrics and analog assignments for pollen records". In: *Quaternary Research* 60.3, pp. 356–367.

George, Liddell Henry, Robert, Scott, Stuart, Jones Henry, and Roderick, McKenzie (1996). *A Greek-English Lexicon*.

Gibbs, Alison L. and Su, Francis E. (2002). "On choosing and bounding probability metrics". In: *International statistical review* 70.3, pp. 419–435.

Gilchrist, Warren (2000). *Statistical modelling with quantile functions*. Chapman and Hall/CRC.

Gill, Philip E., Murray, Walter, and Wright, Margaret H. (1981). "Practical optimization". In:

Ginestet, Cedric E., Simmons, Andrew, and Kolaczyk, Eric (2012). "Weighted Frechet means as convex combinations in metric spaces: properties and generalized median inequalities". In: *Statistics and Probability Letters* 82.10, pp. 1859–1863. DOI: 10.1016/j.spl.2012.06.001.

Givens, Clark R., Shortt, Rae M., et al. (1984). "A class of Wasserstein metrics for probability distributions." In: *The Michigan Mathematical Journal* 31.2, pp. 231–240.

Goldstein, Harvey, Bonnet, Gerard, and Rocher, Thierry (2007). "Multilevel structural equation models for the analysis of comparative data on educational performance". In:

Golub, Gene H. and Van Loan, Charles F. (1996). *matrix computations, 3rd*.

Gowda, K.Chidananda and Diday, Edwin (1991a). "Symbolic clustering using a new dissimilarity measure". In: *Pattern Recognition* 24.6, pp. 567–578.

— (1991b). "Unsupervised learning through symbolic clus- tering". In: *Pattern Recognition* 12.5, pp. 259–264.

Gowda, K.Chidananda and Ravi, T.V. (1995). "Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity". In: *Pattern Recognition Letters* 16.6, pp. 647–652. ISSN: 0167-8655. DOI: https://doi.org/10.1016/0167-8655(95)80010-Q. URL: http://www.sciencedirect.com/science/article/pii/016786559580010Q.

Graziani, Rebecca and Veronese, Piero (2009). "How to compute a mean? The Chisini approach and its applications". In: *The American Statistician* 63.1, pp. 33–36. DOI: 10.1198/tast.2009.0006.

Grilli, Leonardo and Sani, Claudia (2010). "Valutazione degli apprendimenti degli studenti della scuola primaria italiana: un'Analisi Multilivello". In:

Haas, Arthur Erich (1925). *Introduction to theoretical physics*. Vol. 2. Constable Limited.

Halmos, Paul R. (2017). *Naive set theory*. Courier Dover Publications.

Han, Jiawei, Pei, Jian, and Kamber, Micheline (2011). *Data mining: concepts and techniques*. 3rd ed. Waltham, USA: TheMorgan Kaufmann Series in DataManagement Systems, Elsevier. ISBN: 978-0-12-381479-1.

Hand, David, Mannila, Heikki, and Smyth, Padhraic (2001). *Principles of data mining*. 2nd ed. Cambridge, England: The MIT Press. ISBN: 9780262082907.

Hartigan, John A. (1975). "Clustering algorithms". In:

Hathaway, Richard J. and Bezdek, James C. (1994). "NERF c-means: Non-Euclidean relational fuzzy clustering". In: *Pattern recognition* 27.3, pp. 429–437.

Hoerl, Arthur E. and Kennard, Robert W. (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67. DOI: http://dx.doi.org/10.2307/1267351.

Hofmann, Marko A. (2015). "Searching for effects in big data: Why p-values are not advised and what to use instead". In: *Winter Simulation Conference (WSC), 2015*. IEEE, pp. 725–736.

Howell, David C. (2011). "Confidence intervals on effect size". In: *University of Vermont*.

Hui, Sammy King Fai, Brown, Gavin T.L., and Chan, Sky Wai Man (2017). "Assessment for learning and for accountability in classrooms: The experience of four Hong Kong primary school curriculum leaders". In: *Asia Pacific Education Review* 18.1, pp. 41–51.

Ichino, Manabu and Yaguchi, Hiroyuki (1994). "Generalized Minkowski metrics for mixed feature-type data analysis". In: *IEEE Transactions on Systems, Man, and Cybernetics* 24.4, pp. 698–708. ISSN: 0018-9472. DOI: 10.1109/21.286391.

Irpino, Antonio (2018). *HistDAWass: Histogram-Valued Data Analysis*. R package version 1.0.1. URL: https://CRAN.R-project.org/package=HistDAWass.

Irpino, Antonio and Verde, Rosanna (2006). "A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data". In: *Data science and classification*. Springer, pp. 185–192.

Irpino, Antonio and Verde, Rosanna (2015). "Basic statistics for distributional symbolic variables: a new metric-based approach". In: *Advances in*

Data Analysis and Classification 9.2, pp. 143–175. DOI: 10.1007/s11634-014-0176-4.

Irpino, Antonio, Verde, Rosanna, and DeCarvalho, Francisco (2014). "Dynamic clustering of histogram data based on adaptive squared Wasserstein distances". In: *Expert Systems with Applications* 41.7, pp. 3351–3366.

Jaccard, Paul (1901). "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". In: *Bull Soc Vaudoise Sci Nat* 37, pp. 547–579.

Jain, Anil K. and Dubes, Richard C. (1988). "Algorithms for clustering data". In:

James, Gareth, Witten, Daniela, Hastie, Trevor, and Tibshirani, Robert (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Jee, Kyoungyoung and Kim, Gang-Hoon (2013). "Potentiality of big data in the medical sector: focus on how to reshape the healthcare system". In: *Healthcare informatics research* 19.2, pp. 79–85.

Jeffreys, Harold (1998). *The theory of probability*. OUP Oxford.

John Lu, ZQ (2010). "The elements of statistical learning: data mining, inference, and prediction". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173.3, pp. 693–694.

Jolliffe, Ian (2011). "Principal component analysis". In: *International encyclopedia of statistical science*. Springer, pp. 1094–1096.

Kaufman, Leonard and Rousseeuw, Peter J (2009). *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.

Kearfott, Baker (1996). "Interval computations: Introduction, uses, and resources". In: *Euromath Bulletin* 2.1, pp. 95–112.

Kelly, Anthony (2004). *Benchmarking for school improvement: a practical guide for comparing and achieving effectiveness*. Routledge.

Khamsi, Mohamed A. and Kirk, William A. (2011). *An introduction to metric spaces and fixed point theory*. 1st ed. Vol. 53. New York, Usa: John Wiley & Sons. ISBN: 0-471-41825-0.

Kim, Dongmin, Sra, Suvrit, and Dhillon, Inderjit S. (2010). "Tackling box-constrained optimization via a new projected quasi-newton approach". In: *SIAM Journal on Scientific Computing* 32.6, pp. 3548–3563.

Kim, Gang-Hoon, Trimi, Silvana, and Chung, Ji-Hyong (2014). "Big-data applications in the government sector". In: *Communications of the ACM* 57.3, pp. 78–85.

Kim, Jaejik (2009). "Dissimilarity measures for histogram-valued data and divisive clustering of symbolic objects". PhD thesis. Athens, GA, USA: University of Georgia.

Kim, Jaejik and Billard, Lynne (2011). "A polythetic clustering process and cluster validity indexes for histogram-valued objects". In: *Computational Statistics & Data Analysis* 55.7, pp. 2250–2262.

— (2013). "Dissimilarity Measures for Histogram-valued Observations". In: *Communications in Statistics - Theory and Methods* 42.2, pp. 283–303. DOI: 10.1080/03610926.2011.581785. URL: https://doi.org/10.1080/03610926.2011.581785.

Kitchin, Rob (2014). "Big Data, new epistemologies and paradigm shifts". In: *Big Data & Society* 1.1, p. 2053951714528481.

Koh, Lenny S.C., Gunasekaran, Angappa, and Saad, Syed M. (2005). "A business model for uncertainty management". In: *Benchmarking: An International Journal* 12.4, pp. 383–400.

Kouzmin, Alexander, Löffler, Elke, Klages, Helmut, and Korac-Kakabadse, Nada (1999). "Benchmarking and performance measurement in public sectors: towards learning for agency effectiveness". In: *International Journal of Public Sector Management* 12.2, pp. 121–144.

Krinidis, Stelios and Chatzis, Vassilios (2010). "A robust fuzzy local information C-means clustering algorithm". In: *IEEE transactions on image processing* 19.5, pp. 1328–1337.

Küçüktunç, Onur, Güdükbay, Uğur, and Ulusoy, Özgür (2010). "Fuzzy color histogram-based video segmentation". In: *Computer Vision and Image Understanding* 114.1, pp. 125–134.

Kullback, Salomon and Leibler, Richard (1951). "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86. ISSN: 00034851. URL: http://www.jstor.org/stable/2236703.

Labrinidis, Alexandros and Jagadish, Hosagrahar V (2012). "Challenges and opportunities with big data". In: *Proceedings of the VLDB Endowment* 5.12, pp. 2032–2033.

Langfelder, Peter, Zhang, Bin, and Horvath, Steve (2007). "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R". In: *Bioinformatics* 24.5, pp. 719–720.

Lawson, Charles L and Hanson, Richard J (1995). *Solving least squares problems*. Vol. 15. Siam.

Lee, Daniel D. and Seung, H. Sebastian (1999). "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755, p. 788.

Lee, Lillian (1999). "Measures of distributional similarity". In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 25–32. URL: https://arxiv.org/pdf/cs/0001012.pdf.

Lesot, Marie-Jeanne and Kruse, Rudolf (2007). "Typicality degrees and fuzzy prototypes for clustering". In: *Advances in Data Analysis*. Springer, pp. 107–114.

Lesot, Marie-Jeanne, Rifqi, Maria, and Bouchon-Meunier, Bernadette (2008). "Fuzzy prototypes: From a cognitive view to a machine learning principle". In: *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*. Springer, pp. 431–452.

Li, Ming and Vitanyi, P.M. (2008). *An Introduction to Kolmogorov Complexity and its Applications*. 3rd ed. New York: Springer.

Li, Shan, Wang, PZ, Louviere, JJ, and Carson, Richard (2003). "Archetypal analysis: A new way to segment markets based on extreme individuals". In: *Australian and New Zealand Marketing Academy Conference*. ANZMAC.

Lin, Jianhua (1991). "Divergence measures based on the Shannon entropy". In: *IEEE Transactions on Information theory* 37.1, pp. 145–151.

Lloyd, Stuart (1982). "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2, pp. 129–137.

Macfadyen, Leah P and Dawson, Shane (2012). "Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan." In: *Journal of Educational Technology & Society* 15.3.

Macfadyen, Leah P, Dawson, Shane, Pardo, Abelardo, and Gaševic, Dragan (2014). "Embracing big data in complex educational systems: The learning analytics imperative and the policy challenge." In: *Research & Practice in Assessment* 9, pp. 17–28.

Mallinger, Mark and Stefl, Matt (2015). "Big Data Decision Making". In: *Graziadio Business Review* 18.2.

Mallows, Colin L. (1972a). "A note on asymptotic joint normality". In: *The Annals of Mathematical Statistics*, pp. 508–515.

— (1972b). "A note on asymptotic joint normality". In: *The Annals of Mathematical Statistics*, pp. 508–515.

Manning, Christopher D. and Schütze, Hinrich (1999). *Foundations of statistical natural language processing*. MIT press.

Manteiga, Wenceslao González and Vieu, Philippe (2007). "Statistics for functional data". In: *Computational Statistics & Data Analysis* 51.10, pp. 4788–4792.

Manyika, James, Chui, Michael, Brown, Brad, Bughin, Jacques, Dobbs, Richard, Roxburgh, Charles, and Byers, Angela H. (2011). "Big data: The next frontier for innovation, competition, and productivity". In:

Medin, Douglas L. and Schaffer, Marguerite M. (1978). "Context theory of classification learning." In: *Psychological review* 85.3, p. 207.

Miller, Renée J. and Yang, Yuping (1997). "Association rules over interval data". In: *ACM SIGMOD Record* 26.2, pp. 452–461.

Mining, Through Educational Data (2012). "Enhancing teaching and learning through educational data mining and learning analytics: An issue brief". In: *Proceedings of conference on advanced technology for education*.

Mittas, Nikolaos, Karpenisi, Vagia, and Angelis, Lefteris (2014). "Benchmarking effort estimation models using archetypal analysis". In: *Proceedings of the 10th International Conference on Predictive Models in Software Engineering*. ACM, pp. 62–71.

Moliner, Jesús and Epifanio, Irene (2018). "Bivariate functional archetypoid analysis: an application to financial time series". In: *Mathematical and Statistical Methods for Actuarial Sciences and Finance*. Springer, pp. 473–476.

Moore, Ramon E. (1962). "Interval arithmetic and automatic error analysis in digital computing". In: *Ph. D. Dissertation, Department of Mathematics, Stanford University*.

Moore, Ramon E., Kearfott, Baker, and Cloud, Michael (1979). *Methods and applications of interval analysis*. 1st ed. Vol. 2. Philadelphia, Pennsylvania: SIAM: Society for Industrial and Applied Mathematics. ISBN: 978-0-898716-69-6.

Moore, Ramon E. and Lodwick, Weldon (2003). "Interval analysis and fuzzy set theory." In: *Fuzzy sets and systems* 135.1, pp. 5–9. DOI: 10.1016/S0165-0114(02)00246-4.

Morris, Robert J.T. and Truskowski, Brian J. (2003). "The evolution of storage systems". In: *IBM systems Journal* 42.2, pp. 205–217. DOI: 10.1147/sj.422.0205.

Mørup, Morten and Hansen, Lars Kai (2010). "Archetypal analysis for machine learning". In: *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*. IEEE, pp. 172–177.

— (2012). "Archetypal analysis for machine learning and data mining". In: *Neurocomputing* 80, pp. 54–63.

Na, Shi, Xumin, Liu, and Yong, Guan (2010). "Research on k-means clustering algorithm: An improved k-means clustering algorithm". In: *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on*. IEEE, pp. 63–67.

Neumaier, Arnold (1990). *Interval methods for systems of equations*. Vol. 37. Cambridge university press.

Nickerson, Raymond S. (2000). "Null hypothesis significance testing: a review of an old and continuing controversy." In: *Psychological methods* 5.2, p. 241.

Nieddu, Luciano and Rizzi, Alfredo (2007). "Proximity measures in symbolic data analysis". In: *Statistica* 63.2, pp. 195–211.

Nielsen, F. and Bhatia, R. (2013). *Matrix information geometry*. 1st ed. Berlin, Germany: Springer Science and Business Media. ISBN: 978-3-642-30231-2.

Noirhomme-Fraiture, Monique and Brito, Paula (2011). "Far beyond the classical data models: symbolic data analysis." In: *Statistical Analysis and Data Mining: the ASA Data Science Journal* 4.2, pp. 157–170. DOI: 10.1002/sam.10112.

Palumbo, Francesco and Irpino, Antonio (2005). "Multidimensional interval-data: metrics and factorial analysis". In: *Proceedings ASMDA 2005*.

Palumbo, Francesco and Ragozini, Giancarlo (n.d.). "Statistical categorization through archetypal analysis". In: *SIS 2017 Statistics and Data Science: new challenges, new generations* (), p. 759.

Petracco-Giudici, Marco, Vidoni, Daniele, and Rosati, Rossana (2010). "Compositional effects in Italian primary schools: an exploratory analysis of INVALSI SNV data and suggestions for further research". In: *World Summit on Knowledge Society*. Springer, pp. 460–470.

Pfeiffer, Paul E. (1990). "Some Properties of the Quantile Function". In: *Probability for Applications*. Springer, pp. 266–271.

Phillips, David L. (1962). "A technique for the numerical solution of certain integral equations of the first kind". In: *Journal of the ACM (JACM)* 9.1, pp. 84–97. DOI: http://dx.doi.org/10.1145/321105.321114.

Porzio, Giovanni C., Ragozini, Giancarlo, and Vistocco, Domenico (2006). "Archetypal Analysis for Data Driven Benchmarking". In: *Data Analysis, Classification and the Forward Search*. Springer, pp. 309–318.

— (2008). "On the use of archetypes as benchmarks". In: *Applied Stochastic Models in Business and Industry* 24.5, pp. 419–437.

Priss, Uta (2006). "Formal concept analysis in information science". In: *Annual Review of Information Science and Technology, ASIST* 40.

Provost, Foster and Fawcett, Tom (2013). "Data science and its relationship to big data and data-driven decision making". In: *Big data* 1.1, pp. 51–59.

Qing, Ye Xiu, Hua, Huang Zhen, Qiang, Xiao, et al. (1992). "Histogram based fuzzy C-mean algorithm for image segmentation". In: *11th IAPR International Conference on Pattern Recognition. Vol. III. Conference C: Image, Speech and Signal Analysis,* IEEE, pp. 704–707.

Ragozini, Giancarlo and D'Esposito, Maria Rosaria (2015). "Archetypal networks". In: *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE, pp. 807–814.

Ragozini, Giancarlo, De Stefano, Domenico, and D'Esposito, Maria Rosaria (2017). "Prototyping and comparing networks through Archetypal Anal-

ysis". In: *ARS'17 International Workshop: Challenges in Social Network Research*. ARS'17, pp. 103–103.

Ragozini, Giancarlo, Palumbo, Francesco, and D'Esposito, Maria Rosaria (2017). "Archetypal analysis for data-driven prototype identification". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10.1, pp. 6–20.

Ramsay, James (2005). "Functional data analysis". In: *Encyclopedia of Statistics in Behavioral Science*.

Reddy, Damodar, Jana, Prasanta K., and Member, IEEE Senior (2012). "Initialization for K-means clustering using Voronoi diagram". In: *Procedia Technology* 4, pp. 395–400.

Rifqi, Maria (1996). "Constructing prototypes from large databases". In: *International conference on Information Processing and Management of Uncertainty in knowledge-based systems, IPMU'96*.

Robert, Serge (2005). "Categorization, reasoning, and memory from a neological point of view". In: *Handbook of categorization in cognitive science*. Elsevier, pp. 699–717.

Rocha, Luis Mateus (1999). "Evidence sets: modeling subjective categories". In: *International Journal of General System* 27.6, pp. 457–494.

Rockafellar, Tyrrell R. and Wets, Roger J.B. (2009). *Variational analysis*. Vol. 317. Springer Science & Business Media.

Rokach, Lior and Maimon, Oded (2005). "Clustering methods". In: *Data mining and knowledge discovery handbook*. Springer, pp. 321–352.

Rosch, Eleanor H. (1973). "Natural categories". In: *Cognitive psychology* 4.3, pp. 328–350.

Rosch, Eleanor H. and Lloyd, Barbara Bloom (1978). "Cognition and categorization". In:

Rüschendorf, L. (2001). *Wasserstein metric. IN: HAZEWINKEL, M.(Ed.) Encyclopedia of mathematics*.

Saaty, Thomas L. (1990). "The analytic hierarchy process". In: *European Journal of Operational Research* 48, pp. 9–26.

Samuel, Arthur L. (1959). "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3. Reprinted in E. A. Feigenbaumand J. Feldman, editors, Computers and Thought, McGraw-Hill, New York, 1963, pp. 211–229.

Sani, Claudia and Grilli, Leonardo (2011). "Differential Variability of Test Scores among Schools: A Multilevel Analysis of the Fifth-Grade INVALSI Test Using Heteroscedastic Random Effects." In: *Journal of applied quantitative methods* 6.4, pp. 88–99.

Schölkopf, Bernhard, Smola, Alexander, and Müller, Klaus-Robert (1998). "Nonlinear component analysis as a kernel eigenvalue problem". In: *Neural computation* 10.5, pp. 1299–1319.

Scott, John (2017). *Social network analysis*. Sage.

Seiler, Christian and Wohlrabe, Klaus (2013). "Archetypal scientists". In: *Journal of Informetrics* 7.2, pp. 345–356.

Shannon, Claude E., Weaver, Warren, and Burks, Arthur W. (1951). "The Mathematical Theory of Communication (Review)". In: *Philosophical Review* 60.3, pp. 398–400.

Singleton, Jr. Royce, Straits, Bruce C., Straits, Margaret M., and McAllister, Ronald J. (2005). *Approaches to social research*. 4th ed. New York: Oxford University Press.

Smith, Peter (1990). "The use of performance indicators in the public sector". In: *Journal of the royal statistical society. Series A (statistics in society)*, pp. 53–72.

Spendolini, Michael (1992). *The benchmarking book*. Tech. rep.

Stone, Emily (2002). "Exploring archetypal dynamics of pattern formation in cellular flames". In: *Physica D: Nonlinear Phenomena* 161.3-4, pp. 163–186.

Stone, Emily and Cutler, Adele (1996). "Introduction to archetypal analysis of spatio-temporal dynamics". In: *Physica D: Nonlinear Phenomena* 96.1-4, pp. 110–131.

Sun, Weiwei, Yang, Gang, Wu, Ke, Li, Weiyue, and Zhang, Dianfa (2017). "Pure endmember extraction using robust kernel archetypoid analysis for hyperspectral imagery". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 131, pp. 147–159.

Talluri, Srinivas (2000). "A benchmarking method for business-process reengineering and improvement". In: *International Journal of Flexible Manufacturing Systems* 12.4, pp. 291–304.

Tan, Pang-Ning et al. (2007). *Introduction to data mining*. Pearson Education India.

Taneja, Inder J. (2001). "Generalized Information Measuresand Their Applications". In: *Department de Mathematics, University Federal de Santa Catarina, Florianópolis, SC, Brazil*.

Thøgersen, Juliane Charlotte, Mørup, Morten, Damkiær, Søren, Molin, Søren, and Jelsbak, Lars (2013). "Archetypal analysis of diverse Pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways". In: *BMC bioinformatics* 14.1, p. 279.

Timm, Heiko, Borgelt, Christian, Döring, Christian, and Kruse, Rudolf (2004). "An extension to possibilistic fuzzy cluster analysis". In: *Fuzzy Sets and systems* 147.1, pp. 3–16.

Towse, John N., Cowan, Nelson, Hitch, Graham J., and Horton, Neil J. (2008). "The recall of information from working memory: Insights from behavioural and chronometric perspectives". In: *Experimental Psychology* 55.6, pp. 371–383.

Trinchero, Roberto (2014). "Il Servizio Nazionale di Valutazione e le prove Invalsi. Stato dell'arte e proposte per una valutazione come agente di cambiamento". In: *Form@ re-Open Journal per la formazione in rete* 14.4, pp. 34–49.

Tversky, Amos (1977). "Features of similarity." In: *Psychological review* 84.4, p. 327.

Van der Laan, Mark, Pollard, Katherine, and Bryan, Jennifer (2003). "A new partitioning around medoids algorithm". In: *Journal of Statistical Computation and Simulation* 73.8, pp. 575–584.

Vapnik, Vladimir N. (2000). *The Nature of Statistical Learning Theory*. 2nd ed. New York: Statistics for Engineering and Informatic Science, Springer. ISBN: 978-1-4419-3160-3.

Vapnik, Vladimir N. and Chervonenkis, A-Ya (1971). "On the uniform convergence of relative frequencies of events to their probabilities". In: *Theory of Probability and Its Applications* 16.2, pp. 264–280. DOI: http://dx.doi.org/10.1137/1116025.

Vertan, Constantin and Boujemaa, Nozha (2000). "Using fuzzy histograms and distances for color image retrieval". In: *Challenge of Image Retrieval*, pp. 1–6.

Vinué, Guillermo and Epifanio, Irene (2017). "Archetypoid analysis for sports analytics". In: *Data Mining and Knowledge Discovery* 31.6, pp. 1643–1677.

Vinué, Guillermo, Epifanio, Irene, and Alemany, Sandra (2015). "Archetypoids: A new approach to define representative archetypal data". In: *Computational Statistics & Data Analysis* 87, pp. 102–115.

Wächter, Andreas and Biegler, Lorenz T. (2006). "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". In: *Mathematical programming* 106.1, pp. 25–57.

Ward, Joe H. (1963). "Hierarchical grouping to optimize an objective function". In: *Journal of the American statistical association* 58.301, pp. 236–244.

West, Darrell M. (2012). "Big data for education: Data mining, data analytics, and web dashboards". In: *Governance studies at Brookings* 4.1.

West, Douglas Brent et al. (2001). *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River.

Wiggerts, Theo A. (1997). "Using clustering algorithms in legacy systems remodularization". In: *wcre*. IEEE, p. 33.

Wilks, Daniel S. (2011). "Cluster analysis". In: *International geophysics*. Vol. 100. Elsevier, pp. 603–616.

Wille, Rudolf (1982). "Restructuring lattice theory: an approach based on hierarchies of concepts". In: *Ordered Sets*. Ed. by I. Rival (ed.) Dordrecht-Boston: D. Reidel Publishing Company, pp. 445–470.

Zadeh, Lotfi Asker (1968). "Probability measures of fuzzy events". In: *Journal of mathematical analysis and applications* 23.2, pp. 421–427.

Zhu, Joe (2014). *Quantitative models for performance evaluation and benchmarking: data envelopment analysis with spreadsheets*. Vol. 213. Springer.

Zhu, Xiaojin (2006). "Semi-supervised learning literature survey". In: *Computer Science, University of Wisconsin-Madison* 2.3, p. 4.

Zikopoulos, Paul, Eaton, Chris, et al. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.