UNIVERSITAT
JAUME•I

Master's Degree in Computational Mathematics
2016 - 2017

Bivariate Functional Archetypoid Analysis: An Application to Financial Time Series

Author:
Jesús Moliner

Tutor:
Irene Epifanio

Abstract

Archetype Analysis (AA) is a statistical technique that describes individuals of a sample as a convex combination of certain number of elements called Archetypes, which in turn, are convex combinations of the individuals in the sample. For it's part, Archetypoid Analysis (ADA) tries to represent each individual as a convex combination of a certain number of extreme subjects called Archetypoids. It is possible to apply these techniques to functional data applying a basis expansion function and performing AA or ADA to the weighted coefficients in the basis.

This document presents an application of Functional Archetypoids Analysis (FADA) to financial time series. The starting time series consists of daily equity prices of the SP500 stocks. From it, measures of volatility and profitability are generated in order to characterize listed companies. These variables are converted into functional data through a Fourier basis expansion function and bivariate FADA is applied. By representing subjects through extreme cases, this analysis facilitates the understanding of both the composition and the relationships between listed companies. Finally, a cluster methodology based on a similarity parameter is presented. Therefore, the suitability of this technique for this kind of time series is shown, as well as the robustness of the conclusions drawn.

Keywords

Archetype Analysis, Archetypoid Analysis, Statistical Learning, Clustering, Unsupervised Learning, Financial Time Series, Extremal point.

Contents

Chapter 1

Introduction

Archetypal analysis represents each observation in the data set as a convex combination of pure extremal types called archetypes. Archetypes themselves are restricted to be convex combinations of the individuals in the data set. This idea was presented for the first time by Cutler & Breiman (1994). However, archetype analysis may not be satisfactory in some fields since, being artificial constructions, nothing guarantees the existence of subjects in our sample with characteristics similar to those of the archetypes (Seiler & Wohlrabe, 2013). In order to solve this issue, the new concept of archetypoid was introduced in Vinue *et al.* (2015). Archetypoid analysis represents each observation in the data set as convex combination of a set of real extreme observations called archetypoids.

This process not only allows us to identify the subjects of the sample with extreme subjects but also facilitates the comprehension of the data set. As some authors affirm, humans understand the data better when the individuals are exposed through their extreme constituents or when features of an individual are shown as opposed to those of another. Because of this, archetypes and archetypoids analysis has aroused the interest of researchers working in different fields as for example astrophysics (Chan *et al.*, 2003), biology (D'Esposito *et al.*, 2012), genetics (Thøgersen *et al.*, 2013), market research (Li *et al.*, 2003), (Porzio *et al.*, 2008), (Midgley & Venaik, 2013), industrial engineering (Epifanio *et al.*, 2013), (**?**), multi document summarization (Canhasi & Kononenko, 2014) and machine learning (Mørup & Hansen, 2012),(Stone, 2002).

Functional data analysis (FDA) is characterized by dealing with data characterized through continuous functions instead of discrete vectors as in classical multivariate analysis. When dealing with time series, the impossibility of measuring most of these variables continuously over time, and the theoretical complexity of many of the statistical methods available for analysis, leads to the management of periodic summaries that constitute the data series that are commonly used in practice. Although there are many modeling and prediction techniques for discrete temporal data, most of them, such as the classic Box-Jenkins theory (see Box & Jenkins (1976)), require that a set of quite restrictive hypotheses be verified, such as stationarity, equally spaced observations or belonging to a specific kind of well known processes.

In their seminal, paper Cutler & Breiman (1994) already present an example with a data set composed by functions, although they resort to the discretization of these functions, taking values in some points, to perform their analysis. The first paper that combines AA and data expressed as functions is Costantini *et al.* (2012). An expansion basis function was applied on the data functions, to finish applying classical AA analysis to the coefficients in the basis. This method presents the disadvantage that it can only be applied if the basis is orthonormal. In Epifanio (2016), authors develop a methodology in order to obtain functional archetypes and archetypoids regardless of whether basis functions are orthonormal or not.

The characteristics of Functional Archetype Analysis (FAA) and Functional Archetypoid Analysys (FADA) make it especially suitable for financial time series. Classification of time series is a task that has been addressed with many approaches, depending on the features of the data to be analysed (see Liao (2005) for a survey). In particular, many authors have addressed the problem of the analysis and classification of financial time series, since these series have certain specific characteristics that should be considered when our objective is to classify them into homogeneous groups. A deeper review of this specific features can be found in Tseng & Li (2011). In this paper we apply FADA to the stock prices of companies in the S&P500 index in a time frame of 13 years, adjusting continuous functions that represent the profitability and volatility for each company. As will be shown, the algorithm will be able to extract qualitative information about the composition of the market and the relationships between listed companies. Finally we propose and apply a methodology to classify subjects in the sample in different classes according to their similarity. As will be seen, the properties of the archetypoid analysis facilitate the compression of the results, which is especially interesting if the result of the model will be interpreted by non-specialized agents.

The main novelties of this work consist of: 1. Applying for the first time ADA together with FDA, to the financial field; 2. Proposing a methodology for representing graphically the information returned by FADA through networks. The rest of the WORK is structured as follows: Chapter 2 presents the theoretical fundamentals of classical AA, ADA and their functional and multivariate versions. The computational details of the implemented algorithm are also detailed in this chapter. Chapter 3 describes the baseline data and details the manipulations carried out to obtain our risk and volatility indicators and the functions that approximate them. In chapter 4 results are shown. On the one hand, archetypoids and their weights in the sample companies are compared with the structure of sectors used by financial analysts. On the other hand, groups resulting from applying the proposed clustering algorithm are presented.

Chapter 2

Methodology

The goal of our analysis is to identify extremal observations corresponding to specific individuals of our sample, which we call archetypoids. In this way, we can express the rest of the individuals in the sample as a convex combination of our archetypoids. In order to achieve this, we use Functional Archetype and Archetypoid Analysis presented in Epifanio (2016). In section 2.1 Archetype and Archetypoid analysis are described. Section 2.2 details the methodology for applying Archetype and Archetypoid analysis to functional data. Finally, section 2.3 addresses Multivariate FAA and FADA.

2.1 Archetype and Archetypoid analysis

Let $\mathbf{X}$ be an $n \times m$ matrix that contains a multivariate dataset with n observations and m variables. The objective of archetype analysis (AA) is to find a $k \times n$ matrix $\mathbf{Z}$, whose rows are the k archetypes in those data, in such a way that data can be approximated by mixtures of the archetypes. To obtain them, AA computes two matrices α and β which minimize the residual sum of squares (RSS) that arises from combining the equation where $\mathbf{x}_i$ is approximated by a mixture of $\mathbf{z}_j$'s(archetypes) ($\sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j||^2$) and the equation where $\mathbf{z}_j$'s is expressed as a mixture of the data ($\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$)

$$RSS = \sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j||^2 = \sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l||^2$$

Under the constraints:

1. $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, ..., n$
2. $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1, .., k$.

First constraint implies that $\mathbf{x}_i$'s are convex combinations of archetypes $\hat{\mathbf{x}} = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$, where α_{ij}'s represent the weight of archetype j for the observation i. In other words, α_{ij}'s represent the percentage of contribution of each archetype to explain each observation.

Second constraint implies that archetypes $\mathbf{z}_j$ are mixture of observations, $\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$ where β_{jl}'s are the weight that each observation has on each archetype.

It is important to remark that archetypes are artificial constructions and they not necessarily match real observations. Specifically, this will only happen when one and only one β_{jl} is equal to one for each archetype, ie, each archetype is composed by only one observation that presents the entire weight. Adding this conditions leads us to archetypoid analysis.

In archetypoid analysis the continuous optimization problem of archetype analysis transforms into the following mixed-integer optimization problem:

$$RSS = \sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j||^2 = \sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l||^2$$

Under the constraints:

1. $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, ..., n$
2. $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, ..., k$

Here, second constraint implies that $\beta_{jl} = 1$ for one and only one l and $\beta_{jl} = 0$ otherwise.

For values of $k > 1$ archetypes belong to the boundary of the convex hull of data (Cutler & Breiman, 1994). By contrast, archetypoids are not restricted to this region (Vinue *et al.*, 2015). If $k = 1$, the archetype coincide with the mean and the archetypoid with the medoid. (Rousseeuw & Kaufman, 1990).

In their seminal paper, Cutler & Breiman (1994) presented an “alternating optimization algorithm”. This method consists in two steps: Finding the best α's for a given set of x-mixtures, and finding the best x-mixtures for a given set of α's. At each step, the sum of squares is reduced, and the algorithm stops when the reduction is sufficiently small.

The convex least squares problems are solved using a penalized version of the non-negative least squares algorithm by Lawson & Hanson (1974). As Cutler & Breiman (1994) claim, this method is quite slow but it allows analysing 'wide' data with much more variables than subjects.

To compute archetypoid analysis, Vinue *et al.* (2015) presented an algorithm which was subsequently implemented in R by Vinue (2015). This method is based on the Partitioning Around Medoids clustering algorithm proposed by Kaufman & Rousseeuw (1990). That algorithm consists of two stages. In the first one, the BUILD step, an initial set of archetypoids is computed. In the second one, the SWAP step, selected archetypoids are exchanged by unselected observations and it is taken into account if these replacements reduce the RSS. In the R implementation, three set of initial candidates are determined according different criteria. The first one is composed by the nearest observations in Euclidean distance to the k archetypes. On the second set of candidates, observations with the maximum α value for each archetype are picked. Thus, this set is composed by candidates that present the largest weight for each archetype. Finally, the third set consists of the observations with the maximum β value for each archetype, i.e., the largest contributors to the generation of archetypes.

Archetypes are not necessarily nested and so are archetypoids. Therefore, changes in k will yield different conclusions. This is why selection criterion is particularly important. Thus, if the researcher has a priori knowledge of the structure of the data, k value can be chosen based on that information. Otherwise, it will be necessary to calculate the RSS for different k values and choose the point where the elbow is found, as in Cutler & Breiman (1994), Eugster & Leisch (2009) or Vinue *et al.* (2015).

2.2 Archetype and Archetypoid analysis for functional data

The defining quality of functional data is that they consist of functions (Ramsay & Silverman, 2002). In this context, the values of the m variables in the multivariate context become function values with a continuous index t, adopting the form $\{x_1(t), ..., x_m(t)\}$ with $t \in [a, b]$. It is assumed that these functions belong to Hilbert space, i.e., they satisfy reasonable smoothness conditions and are square-integrable functions on that interval. In addition, in the definition of inner product, the sums are changed by integrals. Again, the goal of archetype analysis is to find k archetypes so that our data samples can be approximated as a convex combination of that archetypes. The main difference is that now both archetypes and observations are functions. In functional archetypes analysis (FAA) two matrices α and β are calculated minimizing the RSS. The similarities with the method for general multivariate data are evident. However, a couple of features should be highlighted. On one hand, RSS are now calculated with a functional norm instead of a vector norm. On the other hand, observational and archetype vectors $\mathbf{x}_i$ and $\mathbf{z}_i$ now correspond to observational and archetype functions $x_i(t)$ and $z_i(t)$. Anyway, the interpretation of matrices α and β is the same as in standard multivariate case. Functional archetypoid analysis (FADA) is also an adaptation of ADA changing vectors by functions. In this sense, FADA aims to find k functions of the sample (archetypoids) so that it is possible to approximate functions on the sample through mixtures of these functional archetypoids. Again, vector norms are replaced by functional norms. Interpretation of matrices is the same as before.

2.2.1 Computational details

As detailed above, variables are now represented by continuous functions. Therefore, we need to define the L^2-norm ($||f||^2 =< f, f >= \int_a^b f(t)^2 dt$) in order to compute RSS. However, it is easier to work with functions observed in a finite set of points. In their seminal paper Cutler & Breiman (1994) discretize functions over a grid with m values equidistributed between a and b. This way, an $n \times m$ $\mathbf{X}$ matrix is constructed and then standard multivariate AA is applied. Of course, applying archetypoid analysis from this matrix is also possible.

But this approach presents two main drawbacks. In one hand, depending on the nature of the functions large number of m values have to be calculated. The problem is that the computational cost goes up quickly. More specifically, increasing the number of variables implies a polynomial increase of the computation time per iteration in AA (Eugster & Leisch, 2009). In the other hand, this method is approximating the integral $\int_a^b f(t)^2 dt$ as a sum of discrete values, a method less precise than desired. While it is true that it is possible to use more accurate and sophisticated numerical integration techniques, this would also imply higher computational costs.

To solve this problems, another approach is proposed: to represent functions as a linear combination of well known basis functions. This allows us to perform a more efficient analysis since the number of coefficients of the basis functions that we use will normally be smaller than the number of sampling points. This feature is especially interesting when the data are composed of extensive series with many sampling points. Examples of this kind of framework are Thurau & Bauckhage (2009), Mørup & Hansen (2012), Thurau *et al.* (2012), Feld *et al.* (2015), Steinschneider & Lall (2015), Tsanousa *et al.* (2015) or Zhao *et al.* (2015).

Representing functions as combinations of base functions allows us to work with data in which the sampling points are not evenly distributed. It also presents the advantage that it can be applied to datasets where measuring time points, number of data observations and frequency of observations vary across subjects.

Each function x_i is expressed as a linear combiantion of known basis functions B_h with $h = 1, ..., m : x_i(t) = \sum_{h=1}^{m} b_i^h B_h(t) = \mathbf{b_i'B}$ where $\mathbf{b_i'}$ is the transposed vector of coefficients and $\mathbf{B}$ is the functional vector whose elements are the basis functions. The RSS minimization problem with the restrictions for AA and ADA respectively can be written as:

$$
\begin{aligned}
RSS =& \sum_{i=1}^{n} ||x_i - \sum_{j=1}^{k} \alpha_{ij} z_j||^2 \\
=& \sum_{i=1}^{n} ||x_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l||^2 \\
=& \sum_{i=1}^{n} ||\mathbf{b_i'B} - \sum_{j=1}^{k} a_{ij} \sum_{l=1}^{n} \beta_{ij} \mathbf{b_i'B}||^2 \\
=& \sum_{i=1}^{n} ||(\mathbf{b_i'} - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{b_l'})\mathbf{B}||^2 \\
=& \sum_{i=1}^{n} ||\mathbf{a_i'B}||^2 = \sum_{i=1}^{n} < \mathbf{a_i'B}, \mathbf{a_i'B} >= \sum_{i=1}^{n} \mathbf{a_i'Wa_i}
\end{aligned}
$$

where $\mathbf{a_i'} = \mathbf{b_i'} - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{b_l'}$ and $\mathbf{W}$ is an m symmetric matrix with elements $w_{m1,m2} = \int B_{m1} B_{m2}$, namely, the inner products of the pairs of basis functions. If functions selected as basis functions are orthonormal, such as Fourier, $\mathbf{W}$ is the identity matrix of dimension m. In this case, FAA and FADA become AA and ADA of the coefficients of the functions. In all other cases, it will be necessary to use numerical integration to compute $\mathbf{W}$.

2.3 Multivariate FAA and FADA

It is common to analyse data of dimension greater than one. In our context, this means working with samples in which we analyse more than one function for each individual. So that each function describes a characteristic of the subject.

The key is to define an inner product, which is computed simply as the sum of the inner products of the multivariate functions. Therefore, the squared norm of an M multivariate function is the sum of the squared norms of the M components. Consequently, FAA or FADA for M multivarite functions is equivalent to M independent FAA or FADA with shared parameters α and β. Proposed algorithm works with a composite function formed by stringing the M functions together.

Without loss of generality, let $f_i(t) = (x_i(t), y_i(t))$ be a bivariate function. So, its squared norm is $||f_i||^2 = \int_a^b x_i(t)^2 dt + \int_a^b y_i(t)^2 dt$. In order to compute FAA and FADA, let us consider $\mathbf{b_i^x}$ and $\mathbf{b_i^y}$, vectors of length m of the coefficients for x_i and y_i for the basis functions B_h. Thus,

$$
\begin{aligned}
RSS = &\sum_{i=1}^{n} ||f_i - \sum_{j=1}^{k} \alpha_{ij} z_j||^2 \\
= &\sum_{i=1}^{n} ||f_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} f_l||^2 \\
= &\sum_{i=1}^{n} ||x_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l||^2 + \sum_{i=1}^{n} ||y_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} y_l||^2 \\
= &\sum_{i=1}^{n} \mathbf{a^{x}{}'_{i} W a^{x}_{i}} + \sum_{i=1}^{n} \mathbf{a^{y}{}'_{i} W a^{y}_{i}}
\end{aligned}
$$

where $\mathbf{a^x}'_i = \mathbf{b^x}'_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{b^x}'_l$ and $\mathbf{a^y}'_i = \mathbf{b^y}'_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{b^y}'_l$ with the corresponding AA or ADA constraints for α and β. Again, a penalized version of the non-negative least squares algorithm is used to solve the minimization. Observations are now formed by joining $\mathbf{b^x_i}$ and $\mathbf{b^y_i}$. In the case where the basis functions are orthonormal, FAA and FADA can be computed joining the coefficient matrix for x and y components and applying the standard multivariate algorithm to the $n \times 2m$ coefficient matrix.

Chapter 3

Data

3.1 Data sources

In this paper we use data from three different sources. The bulk of the information is provided by QuantQuote (Quantcuote, 2017). This dataset is composed by a collection of daily resolution data with the typical open, high, low, close, volume (OHLCV) structure. This collection, goes back from 01/01/1998 to 07/31/2013 for 500 currently active symbols in the S&P500. The company ensures that provided tickers are reviewed and error-free. In addition, the time series containing the aggregate S&P500 index OHLCV daily ticks has been extracted from the yahoo finance service (Yahoo, 2017). The length of this time series has been selected so that his range coincides with that of the QuantQuote disaggregated series. The last source that has been appealed is SectorSPDR database (ALPS Portfolio Solutions Distributor, Inc., 2017). SectorSPDR classifies stocks on the S&P500 index within ten major sectors, namely: Consumer Discretionary (XLY), Consumer Staples (XLP), Energy (XLE), Financials (XLF), Health Care (XLV), Industrials (XLI), Materials (XLB), Real Estate (XLRE), Technology (XLK), and Utilities (XLU).

3.2 Data manipulation

3.2.1 Calculating return and variance indicators

Our initial data set is composed by 501 tables (500 stocks and the aggregated index) with dimension 3927×6. Each row stores the data of a day for the variables: date, open, high, low, close, and volume, which are represented in the columns.

We are interested in extracting relevant information from a financial point of view. With regard to investments and more specifically to portfolio theory, it is widely accepted that the two key variables are risk and profitability. As Markowitz (1952) pointed out in his seminal paper “the investor considers expected return a desirable thing and variance of return an undesirable thing”. On the one hand, we will choose as a measure of profitability in a time t the aggregate returns over a period of length N, R_N. So, for each price series, s_i, $i = 1, ..., 501$

$$R_N^{s_i}(t) = \frac{X_t^{s_i} - X_{t-N}^{s_i}}{X_{t-N}^{s_i}}$$

where $X_t^{s_i}$ is the value of the stock i at the time t. In our case, we have chosen N = 250, approximately the days that stock markets remain open in a year. It is noteworthy that the logarithmic approximation has been discarded as a result of taking a full year as the basis period. It seems more reasonable to calculate the benefits as an annual discrete rate of return rather than a continuous compounded rate. It is true that there is not a big difference between the two approaches when growth rates are not very high. However, this is not the case that concerns us since we can find in our data inter annual variations that exceed 50%.

On the other hand, we will chose as a measure of volatility the beta or β coefficients, widely used in portfolio theory. Defined for a particular stock s_i in a time t as

$$\beta_N^{s_i}(t) = \frac{Cov(R_N^{s_i}(t), R_N^{index}(t))}{Var(R_N^{index}(t))}$$

where $R_N^{s_i}(t)$ stands for the aggregated returns of s_i in time t over the last N days and $R_N^{index}(t)$ are the returns of the aggregated S&P500 index in the same period. To be consistent, we perform the calculations with a temporary window that is also composed of 250 days. It may be easier to understand the ratio intuitively if we rewrite it as:

$$\beta = \rho_{s_i,index} \frac{\sigma_{s_i}}{\sigma_{index}}$$

where $\rho_{s_i,index}$ is the correlation between the stock s_i and the index and σ_{s_i} and σ_{index} are the variances of each item. In plain words, beta coefficient, or correlated relative volatility, indicates if allocating an asset to a well diversified portfolio will increase or decrease its relative volatility compared with the index volatility. Obviously, beta coefficient of all the market as a whole is 1. As a thumb rule, if a stock has a beta of 2, it means that it has returns that change, on average, by twice the magnitude of the market. In addition, the level of correlation with the market $\rho_{s_i,index}$ will determine the variance over that mean.

After calculations, we have 3677 observations of each variable (r_{250}, and β_{250}) for each stock.

3.3 From discrete to functional data

The missing data do not present a problem for the functional data as will be explained below. However our data sample presents an added problem: some stocks do not exist during the entire time series. Not because we have missing data, but because the companies where founded or started to be listed in the stock exchange after 01/01/1998, i.e, the ranges of the functions are different. Here we propose an alternative to deal with this problem trying to maximize the length of our sample and minimize the stocks that must be discarded, which is to take into account observations since 2000-01-01 and drop the stocks with more than 20% of missing values. In this way four companies are dropped out.

The first step is to decide the number of Fourier functions to use as a basis for representing our data series. In this way, we will transform our variables into two functions of the form

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

where ϕ_k are the Fourier basis functions and c_k are the coefficients for this basis. Selecting a large number of basis functions will give us more precision, but a reduced number of them will improve the computational efficiency. To address this dilemma, it is accepted to look at RMS explained as a function of the number of basis and choose a number of basis so that adding another one doesn't give much better modelling of the data. This general method is known as elbow method and its origins go back to Thorndike (1953).

The graphic analysis shown in Figure 3.1 is certainly not clear. However, following the aforementioned elbow rule, a number of Fourier basis $K = 11$ is chosen.

Finally, for each stock s_i, both variables, return and beta coefficient in a 250 day time window could be expressed as the functions:

$$r_{250}^{s_i}(t) = \sum_{k=1}^{11} \mathbf{a}_k^{s_i} \phi_k(t)$$

and

$$\beta_{250}^{s_i}(t) = \sum_{k=1}^{11} \mathbf{b}_k^{s_i} \phi_k(t)$$

where $\mathbf{a}^{s_i}$ stands for the vector of coefficients on the basis functions for $r_{250}(t)$ corresponding to the particular stock s_i and, in the same way, $\mathbf{b}^{s_i}$ stands for the vector of coefficients for $\beta_{250}(t)$ corresponding to the particular stock s_i.

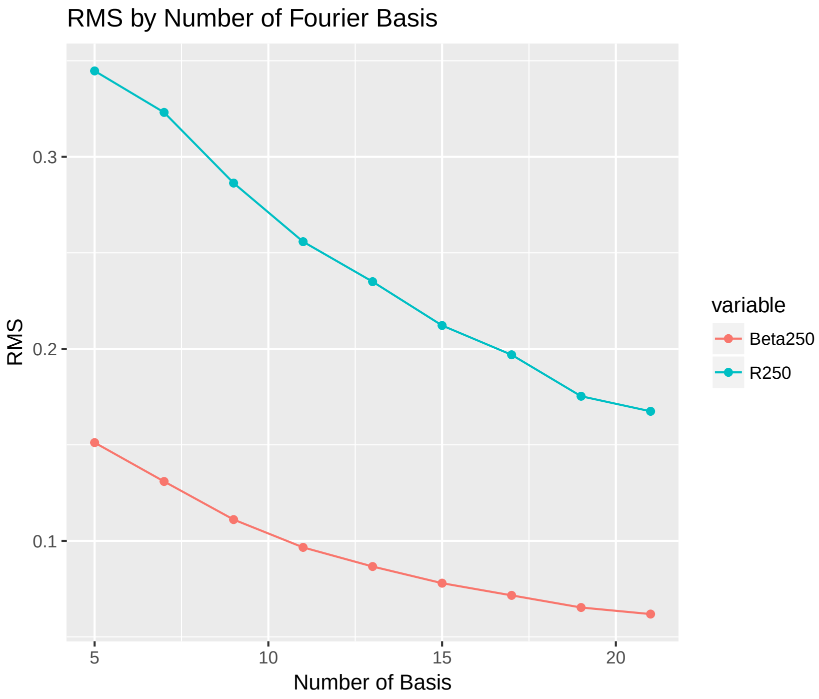

Figure 3.1: RMS depending on the number of Fourier basis used.

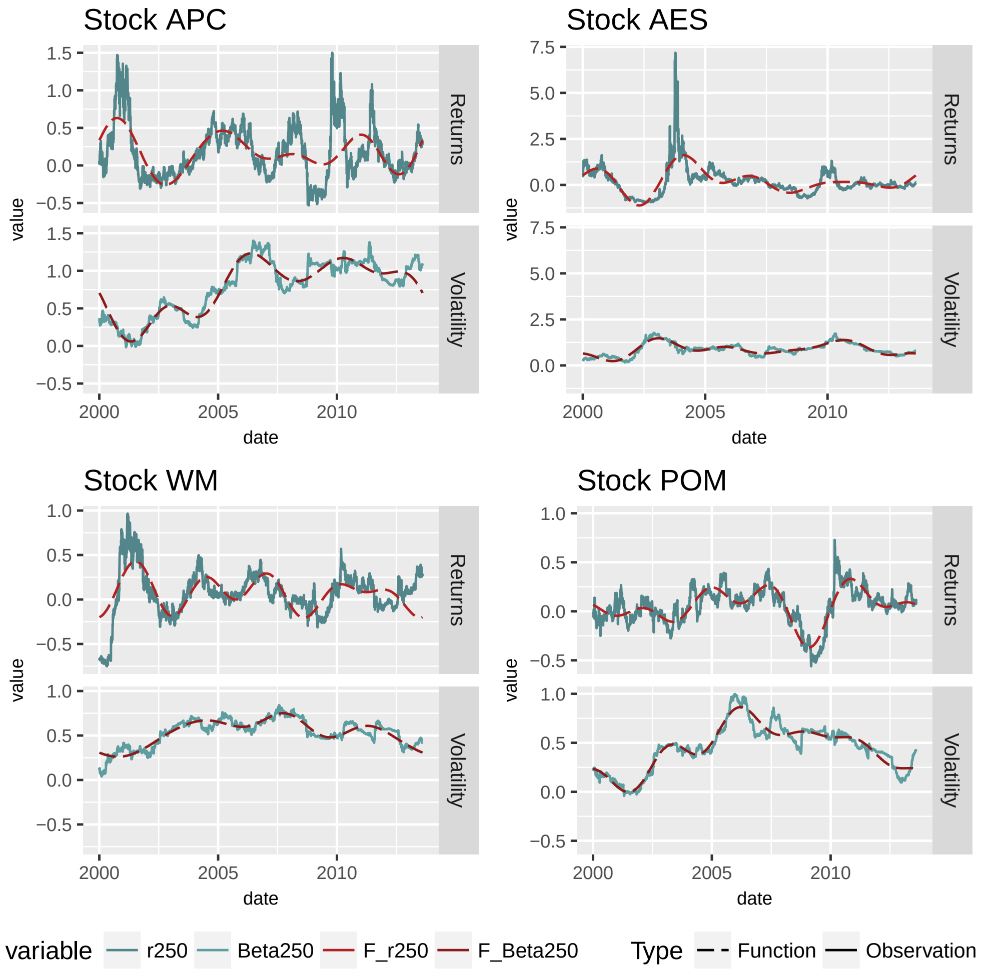

Figure 3.2: Four examples of r_{250} and β_{250} and their functional approximations

It should be noted that we can now store the data, i.e., the coefficients of the functions, in a two-dimensional array with two matrices of size $n \times K$ instead of the original $m \times n$ matrices. In this application, this means reducing the size of the data set from $3422 \times 496 \times 2$ to $496 \times 11 \times 2$.

As a graphic support, Figure 3.2 shows our observed variables (solid line) for four random selected stocks as well as their functional approximations using 11 Fourier basis functions(dashed line).

Chapter 4

Results

4.1 Archetypoid structure

An essential question in AA and ADA, and by extension in FAA and FADA is how many k archetypes or archetypoids we want to choose. Since archetypoids are not nested if we vary our k value, the extreme individuals that we take as archetypoids may not coincide at all. Table 4.1 shows the individuals chosen as archetypoids for different k values in our bivariate functional archetypoids implementation. To facilitate understanding, the economic sector to which each company belongs has been added in parentheses. It can be appreciated that when k increases, the number of sectors represented in the set of archetypoids increases too. As we have already pointed out, the archetypoids are subjects that present extreme characteristics. Moreover, we expect companies that belong to the same sector to have similar behaviours, so when k tends to the number of sectors we expect that an archetype of each sector would appear. However, we see that not all sectors end up having representation, which indicates that some sectors are more likely to contain extreme individuals than others. Regarding companies, not all archetypoids are maintained when k increases, but some of them are quite persistent as for example NTAP or FLIR. However, we find a nested structure in the order in which sectors appear.

k	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
3	NBL (Eneregy)	USB (Financial)	BRCM(Technology)							
4	RRC (Energy)	LNC (Financial)	BRCM(Technology)	CLX (C.Staples)						
5	RRC (Energy)	LNC (Financial)	EBAY (Technology)	NTAP (Technology)	ED (Utilities)					
6	BTU (Energy)	HST (R.Estate)	FLIR (Technology)	EBAY (Technology)	PFE (HealthCare)	CELG (HealthCare)				
7	EOG (Energy)	LNC (Financial)	FLIR (Technology)	NTAP (Technology)	BMC (Materials)	ED (Utilities)	AKAM (Technology)			
8	HP (Energy)	LNC (Financial)	FLIR (Technology)	NTAP (Technology)	BMC (Materials)	PNW (Utilities)	AKAM (Technology)	MNST (C.Staples)		
9	HP (Energy)	XL (Financial)	FLIR (Technology)	NTAP (Technology)	DOW (Materials)	EBAY (Technology)	AKAM (Technology)	MNST (C.Staples)	KR (C.Staples)	
10	RDC (Energy)	SLM (Financial)	FLIR (Technology)	NTAP (Technology)	BMC (Materials)	GIS (C.Discretionary)	AKAM(Technology)	MNST (C.Staples)	IPG (C.Discretionary)	ATI (Materials)

Table 4.1: Functional archetipoyds for different k values.

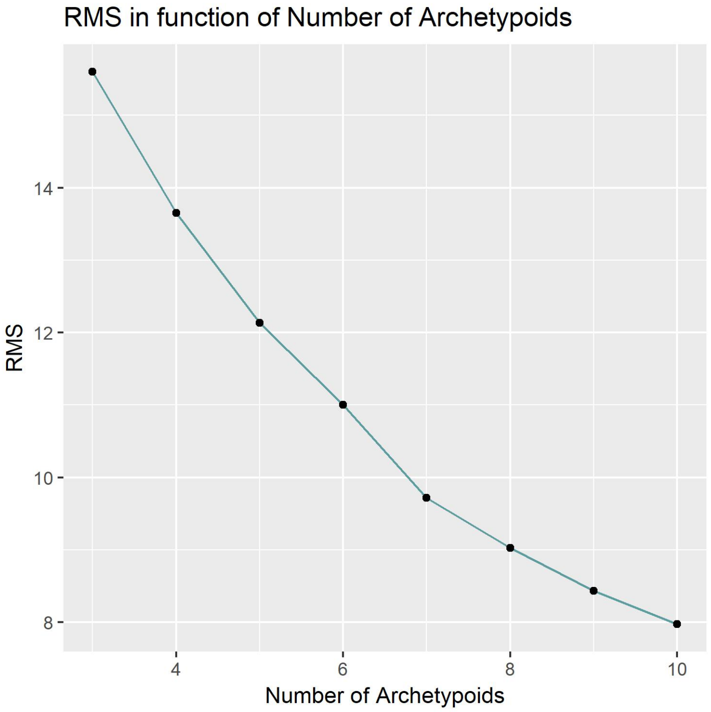

Figure 4.1: Root Mean Square of error in function of the number of Archetypoids chosen

4.2 Selected Archetypoids

We must choose a specific number of archetypoids to continue with our analysis. Once again, the elbow method can help determine which number k is best suited to our problem. Let's consider the Root Mean Square (RMS) of error in function of the number of Archetypoids chosen. The RMS generated by archetypoids calculated starting from candidates of each method are quite similar, therefore, the error of the best adjustment is represented in Figure 4.1. As can be seen, introducing new archetypoids reduces the RMS in any case, because it allows us to discover new patterns that had not previously appeared. So, the explanatory capacity of the convex combination of the archetypoids will improve as long as the number of archetypoids increases, but the idea is to find a point where the curve draws an elbow. As shown in Figure 4.1 there are two candidates, i.e, two points ($k = 5$ and $k = 7$) with significant changes in the slope of the curve. At this point it should be remembered that a desirable feature of the FADA is that it generates easy to understand results. To preserve this feature and keep the analysis in the smallest possible dimension we will choose $k = 5$. In this way it will be easier to understand the nature of our data and see how our subjects are related to each other.

Hence, our five archetypoids are RRC (Energy), LNC (Financial), EBAY (Technology), NTAP (Technology), ED (Utilities). In order to improve the compression, we include a brief description of the selected companies, extracted from the SectorSPDR website.

- Range Resources Corp.(RRC) acquires, develops, and finances oil and gas properties in the U.S. The company is mainly focused on lower-risk development drilling and acquisitions. The company also provides financing to other small oil and gas producers.
- Lincoln National Corp.(LNC) is a holding company. Through subsidiary companies, the company operates multiple insurance and investment management businesses.
- eBay(EBAY) is a powerful marketplace that host one of the largest on-line trading communities for the sale of goods and services.
- NetApp,(NTAP) is a company that provides a full range of hybrid cloud data services that simplify management of applications and data for big enterprises.
- Consolidated Edison, Inc (ED) provides a wide range of energy-related products and services to its customers through regulated utility subsidiaries and competitive energy and telecommunications businesses.

  Beyond the qualitative description of the companies, Figure 4.2 shows the values of the observations for our r_{250} and β_{250} variables as well as the functional approximation for each of the archetypoids.

To be able to compare them, we represent in Figure 4.3 the functions of each variable for all the subjects together. It can be seen that ED is a company that, in comparison with the rest of the archetypoids, presents low and constant values for both variables. Looking at NTAP, it presents high returns at the beginning and at the end of the time series, while its volatility decreases over time. LNC presents a typical financial company profile, with moderate profitability and volatility during the first three quarters of the time series. Once the crisis broke out in 2007, volatility shot up to unprecedented levels while profitability plummeted. Regarding volatilities, EBAY presents just the opposite profile, with great beta values in the first years that stabilize over time. About the returns, we see that this company has a moderate profitability level compared to the other archetypoids, but with slightly higher oscillations at the first half of the time series. Finally, RRC is characterized by having bell-shaped functions, that is, with relatively low values at the extremes of the temporal domain and higher values at the centre.

4.3 Contribution of the archetypoids to the economic sectors

FADA is an unsupervised learning algorithm. However, it is possible to compare the taxonomy of companies obtained by this method with the structure by sectors managed by market analysts.

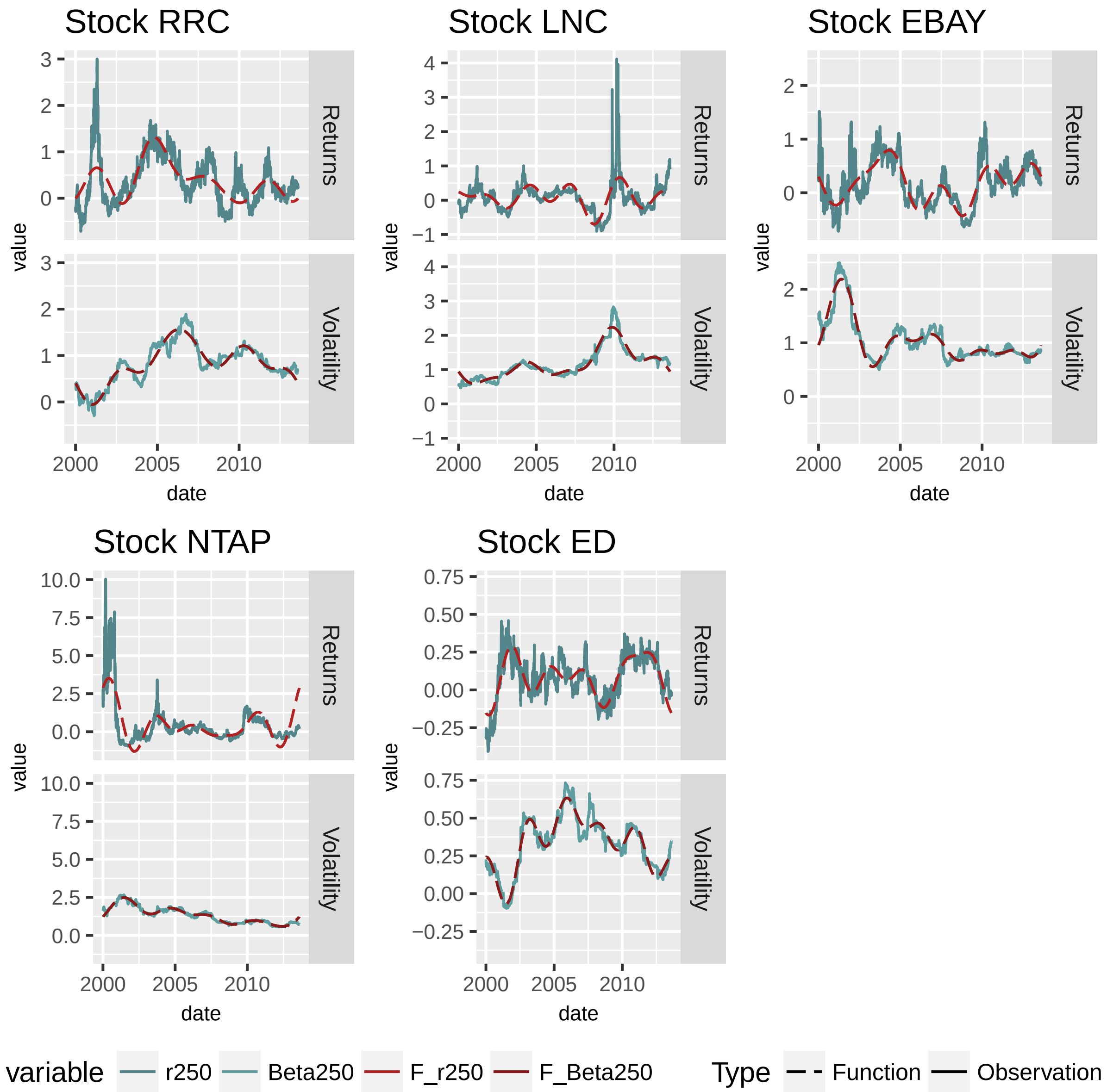

Figure 4.2: r_{250} and β_{250} and their functional approximations of the 5 archetypoids

In this way, we can evaluate the performance of this algorithm in qualitative terms. Figure 4.4 shows the normalized relative weight of archetypoids in each one of our ten sectors. It can be seen that each archetypoid represents the component with the greatest weight of the sector to which it belongs. Thus, RRC represents half the weight of the Energy sector, LNC weighs more than 43% of the financial sector, ED represents almost 80% of the weight of the Utilities sector and in the technology sector, the sum of the EBAY and NTAP weights exceeds the weight of

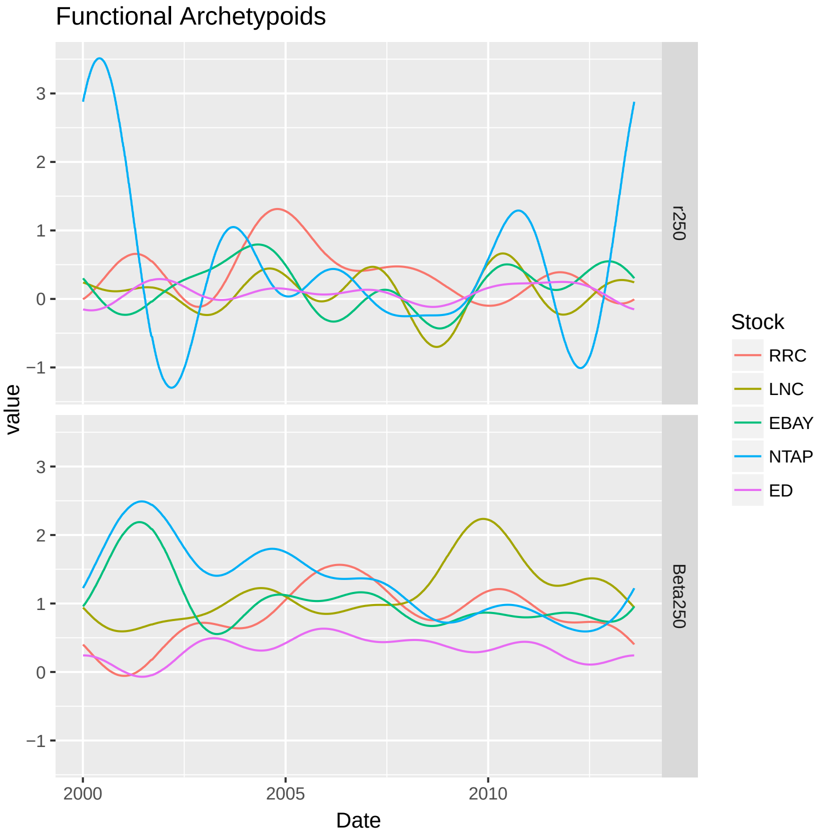

Figure 4.3:

the other three archetypoids.

But it is not only interesting to analyse the weights of the components with greater relevance. The composition of the mixtures also gives clues about the similarities and dependencies of the different sectors among themselves. For example, if we compare Consumer Discretionary sector with Consumer Staples sector, we see that weights keep a proportion that we would expect. For example, the companies that manufacture durable goods, which are included in the Consumer Discretionary sector, have a direct relationship with those that provide the investment to finance these purchases and with the companies that provide these goods, represented by the LNC and EBAY archetypoids respectively, and that is why these archetypoids have higher weights in this sector.

On the other hand, companies that provide basic or non-durable goods, which belong to the Consumer Staples sector, have a minor relationship with the financial sector. It may be obvious, but it is worth emphasizing that, by definition, non-durable goods are those that are purchased without resorting to financing. This sector has instead a great similarity with the Utilities sector

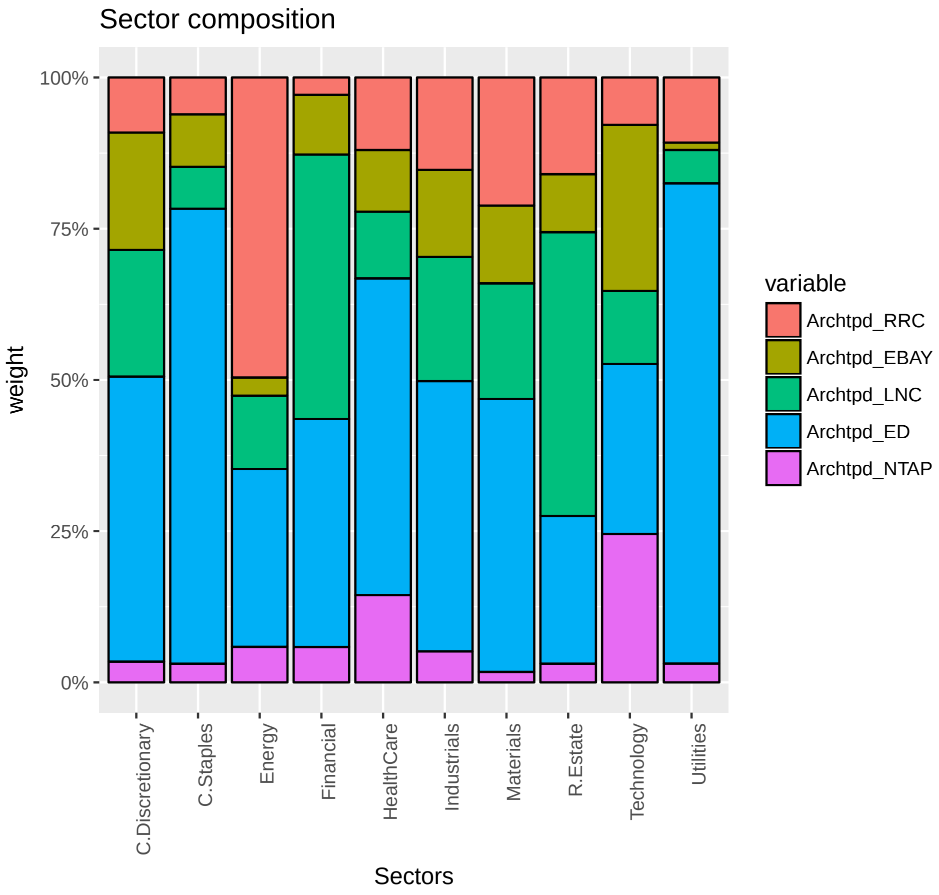

Figure 4.4: Relative weight of archetypoids in each sector

(distributors of electricity, water, gas, etc.). This makes economic sense, since basic goods and services distributed by companies in the Utilities sector have similar demand curves. In other words, in the expansive cycles of the economy, consumers decide to increase their investments in goods that require financing such as a car, a washing machine or a computer. However, spending on electricity, water, gas or telecommunications of households will remain relatively constant as well as spending on basic products such as bread, milk, oil, soap or toothpaste.

Regarding the composition of the energy sector (extractors of oil, gas etc) calls our attention the small weight of the technological archetypoids. This may point to the little relationship between the technologies used in each sector. On the one hand the Energy sector develops activities where the heavy machinery and in general the mechanical operations (drilling, mining, extraction etc.) are fundamental. This is completely opposite to the dynamics that prevail in the technological universe, where the main elements are computer applications, digital technology or patents.

Industrials and Materials sectors have similar profiles, which shows the great interrelation between them. Regarding the Real Estate sector, we see how the LNC archetypoid, belonging to the sector of financial companies, has the greatest weight outside its own sector. The relationship between these two sectors is also evident. Finally we can say that the Utilities sector is in some way the purest, in the sense that it presents a very high proportion between the weight of the archetypoid of this sector and the weight of the other four archetypoids in the companies of this group.

4.4 Taxonomy of the S&P500 stocks based in a simple FADA clustering method

Another usefulness that we can give to the analysis of archetypoids is to build clusters according to the α_{ij} of each individual. In this way, we will group the individuals that present similar coefficients in the linear convex combinations of archetypoids that represent them. A simple method to do this is to establish a threshold U, so that if the weight of an archetypoid for a given individual is greater than U we will say that this subject is in the cluster generated by this archetypoid. If we repeat this process for each archetypoid, we will generate 5 "pure" clusters of subjects, i.e., clusters of subjects that are represented mostly by a single archetypoid. To go a little further, this process is repeated with combinations of two archetypoids. Thus, we will group in the same cluster those individuals whose sum of weights for two concrete archetypoids is greater than U, generating this way $\binom{5}{2} = 10$ additional clusters. In this case it might happen that for a given subject, there are more than one combination of archetypoids whose sum exceeds U. To not complicate the graphical representation we will classify these subjects generically as mixtures, even though these mixtures will be composed of different sets of archetypoids. Figure 4.5 condenses all the information extracted. Archetypoids are highlighted with a grey square, the lines and colour codes allow us to differentiate the structure of the clusters and different sectors are represented through different geometrical shapes.

Starting with the pure clusters we see that EBAY is alone in its own cluster and NTAP generates a small group with 2 more companies. This would suggest that although technology companies present extreme patterns, that patterns are not widely shared by the rest of the companies in the sample.

ED is the archetype that generates the largest cluster on its own. Most of the companies in this cluster belong to the Utilities sector although we also find some of the non-durable goods sector, and some health companies.

LNC generates a small cluster with four other companies. Two of them, XL Group LTD (XL) and Hartgord Financial Services Group INC. (HIG) belong to the same sector and have similar profiles. The other two are Prologis (PLD) related to the logistics of building materials

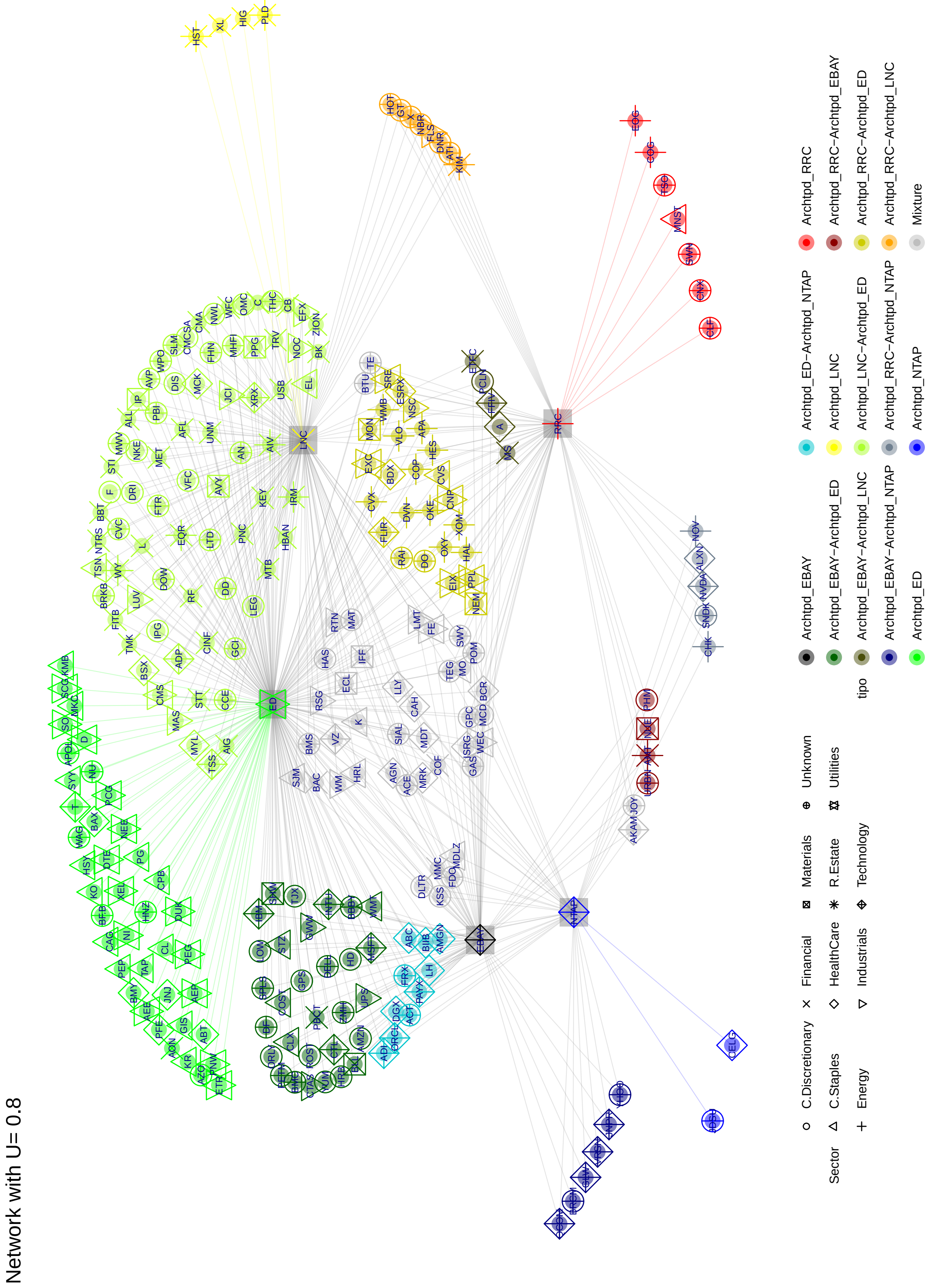

Figure 4.5: Cluster structure with U=0.8

and Host Hotels and Resorts Inc (HST), a real estate investment trust.

For its part, RRC generates a larger cluster in which 7 other companies are included. Regarding the sectors of these companies, we see how the database consulted has not classified some of them, but most of these companies are from the energy sector. Among those not classified are the Tesoro Corporation (TSO) with interests in the petroleum and natural gas sectors, Southwestern Energy (CNX) and Cleveland Cliffs (CLF) dedicated to the extraction of natural gas or Consol Energy (CNX) dedicated to the extraction of coal.

Regarding mixed clusters, we will point out some general characteristics since detailing all the relationships shown in Figure 4.5 would be too much extensive.

A first aspect is that the combination of NTAP and LNC do not form a cluster. In contrast, ED and LNC form the largest cluster for this level of U. Focusing on it, the ED-LNC cluster has just a few companies of the Utilities sector, since these companies are mainly classified in the cluster generated by ED alone. Many of the companies classified in the ED-LNC cluster belong to the financial, real estate and consumer staples sectors.

Apart from those already mentioned, the following most important clusters according to their size are ED-RRC and ED-NTAP. The ED-RRC cluster stands out for its economic sense, since the two generating archetypoids belong to the energy extraction and energy distribution sectors. Thus, this cluster is generated by archetypoids that have a direct economic relationship, so it is expected that the companies in this cluster should also be strongly related. The result is consistent with expectations and we see how most of the companies are part of the Utilities, Energy or Industrial sectors. At a financial level, it may be more interesting to see which companies in these clearly bounded sectors do not belong to this cluster, which would allow to diversify the investments and reduce the risk. However, this type of analysis goes beyond the objective of this paper.

The EBAY-ED cluster is certainly heterogeneous in terms of the sectors that comprise it. However, looking at the companies in the cluster, we see that the algorithm is capable of portraying not so direct economic relationships. Some examples of companies in this cluster are Amazon (AMZN), which also bases its business model on online sales, UPS which is a parcel company, or Ball Corporation (BLL) that is mainly dedicated to the manufacture of packaging for soft drinks, and food. We see that although the sectors to which they belong are different, companies are related since they are all influenced by the development of online sales.

The remaining clusters are characterized by being smaller. To highlight some of them we can mention the one formed by LNC-RRC, which contains companies such as Allegheny Technologies Incorporated (ATI), one of the largest producers of specialty materials in the world or Kimco Realty Corporation (KIM) a holding owner of more than 500 open air shopping centres, or the cluster formed by EBAY-NTAP that is composed only by technological companies.

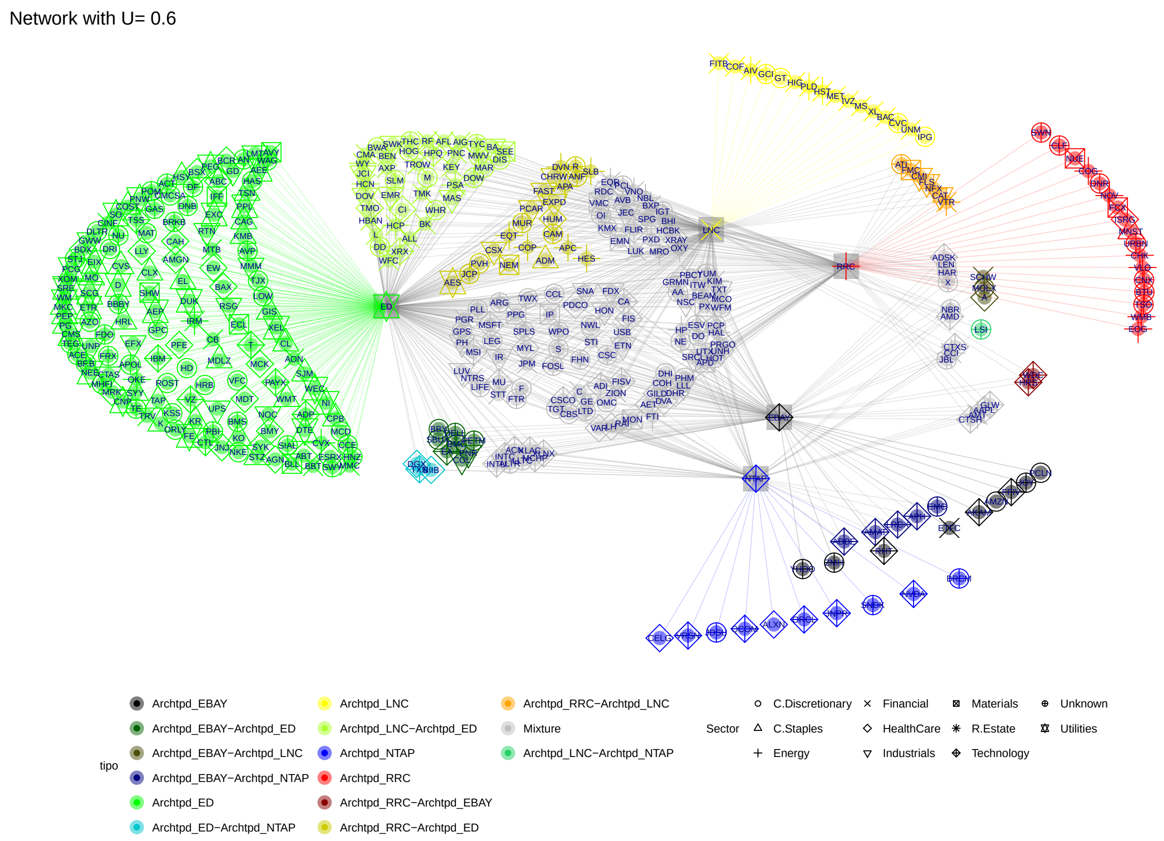

Figure 4.6: Clusters generated with $U = 0.6$

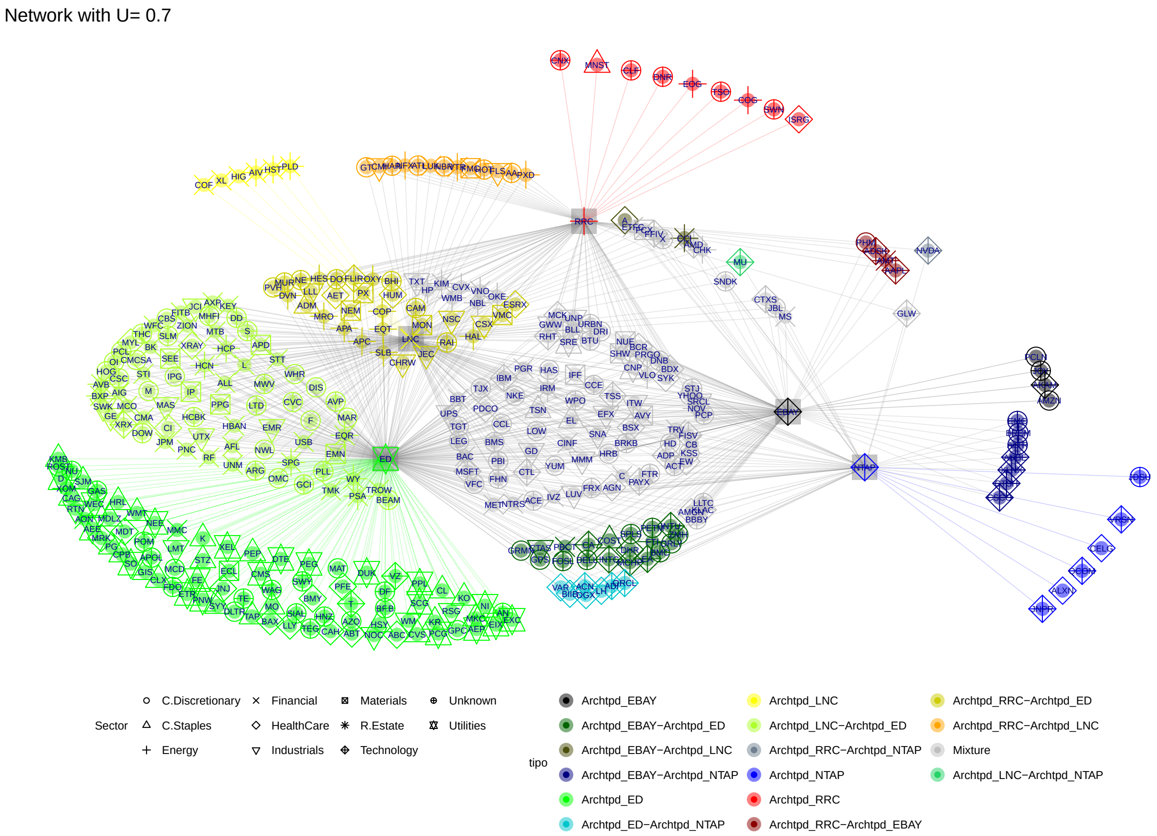

Figure 4.7: Clusters generated with $U = 0.7$

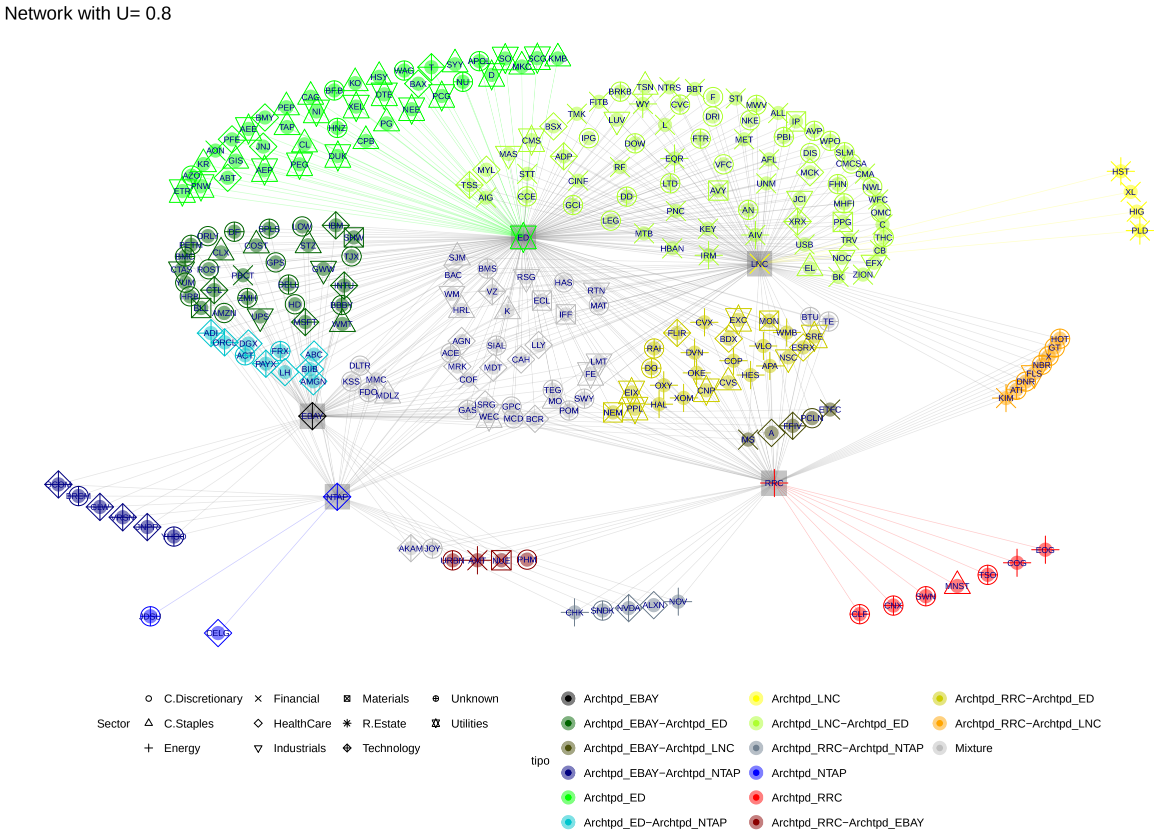

Figure 4.8: Clusters generated with $U = 0.8$

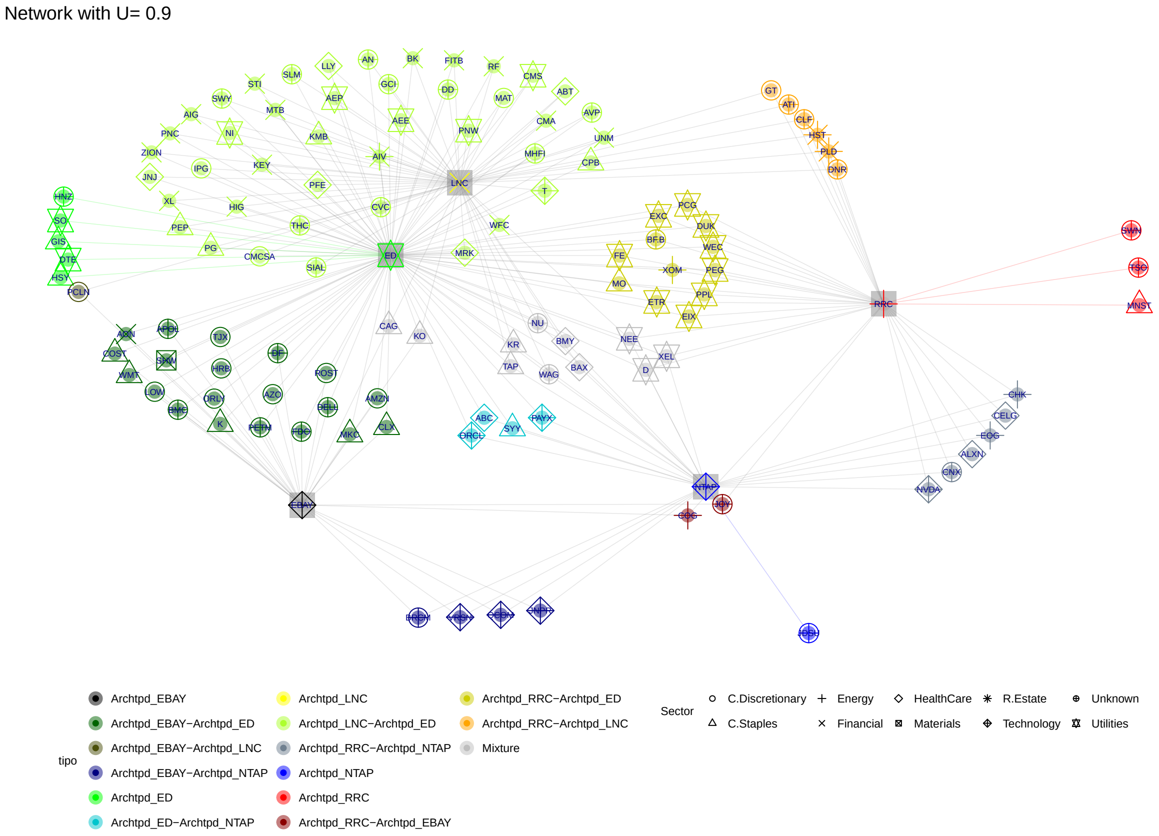

Figure 4.9: Clusters generated with $U = 0.9$

The last issue that we want to illustrate is how the structure of clusters varies depending on the selected U level. Figures 4.6, 4.7, 4.8 and 4.9, show the generated structures using $U = 0.6$, $U = 0.7$, $U = 0.8$ and $U = 0.9$ respectively.

Imposing more restrictive conditions will leave out of our scheme some companies, so we see that the total number of companies that belong to any cluster decreases as U takes higher values.

Another effect of the increasing U is that the number of companies in clusters generated by a single archetypoid decreases. In the end, neither EBAY nor LNC have companies in their clusters. In other words, these two archetypoids do not have a weight greater than 90% in any of the subjects of the sample. However, this fluctuation does not happen equally in all groups. It can be seen how a lot of companies migrate from ED to ED-LNC as the threshold raises while other clusters like RRC-LNC remain relatively stable for any value of U.

Regarding sectors, we see that for $U = 0.9$, there are no companies that belong to the industrial sector, i.e., no industry in this sector presents a weight greater than 90% for any archetypoid, which is consistent with the sector analysis carried out in section 4.3.

Chapter 5

Conclusions

In this paper FADA has been applied to the time series of stock quotes in the S&P500 from 2000 to 2013. The objective of our paper was to show an alternative clustering method to those already used, such as Hausdorff Clustering (Basalto *et al.*, 2007) or the widely used Generalized Autoregressive Conditional Heteroskedasticity model(GARCH) (Bollerslev, 1986).

Understanding the results of these or other statistical learning models is not always an easy task. Additionally, if these results have to be explained to a public without mathematical knowledge, things can be even worse. In that sense, ADA stands out since its results can be interpreted in a very simple way by any non-expert person. For instance, anyone with minimal investment knowledge understands what we mean if we say that a certain stock behaves like a mixture of Consolidated Edison and Lincoln National stocks. Another advantage of the applied model is that the functional version of archetypoid analysis (FADA) allows us to condense vectors of observations of any length into a few coefficients which provides an improvement in computational efficiency and makes this method highly recommended when working with long time series.

With regard to the conclusions, in the first place, it has been seen that when we increase the number of chosen archetypoids, the Technology sector appears repeated while other sectors do not appear. Therefore, there are companies within this sector that exhibit very different behaviours, such as NTAP and EBAY FLIR and AKAM.

Secondly, we have analysed the sectors according to the normalized relative weight of archetypoids that compose each sector and we have seen that some sectors present certain similarities. It is worth mentioning that sectors like Consumer Discretionary, Materials or Industrials offer better opportunities to diversify risks, since their composition is more heterogeneous. On the other hand, sectors such as Utilities or Consumer Staples present more heterogeneous structures, where the weight of the dominant archetypoids of the sector can exceed 80%.

Finally, we have shown the graphic representations of the structure of clusters obtained. By definition, grouping by clusters means assigning classes to objects depending on how similar they are. In this sense, we have established different degrees of similarity through different U values in order to construct our clusters.

Graphic representations practically speak for themselves, and trying to explain with words all the information there reflected would be an extremely exhaustive task. However, we have analysed some aspects such as the effect of selecting different levels of U. It has been shown that some clusters such as LNC-RRC, are composed of the same stocks for any level of U and others, such as the cluster generated by ED, lose individuals gradually when U increases. Thus, the proportion between companies in both types of clusters can vary strongly depending on the threshold we choose.

As regards future work, the application of these models to the world of finance is still a relatively unexplored field. The application of models with functional data allows to take into account variables collected with different frequencies such as daily quotes, quarterly balances, or annual results, which makes these models especially suitable for financial time series.

Taking this into account, a next development may be to extend the implementation of the bivariate model to an n-variable model that allows working with a large amount of data from each company. From a financial point of view, a door opens to develop investment strategies that can use results shown here to improve performance and reduce the risk of investment decisions.

Bibliography

ALPS Portfolio Solutions Distributor, Inc. 2017 (09). *SectorSPDR web-site.* Retrieved on 15/09/2017 from http://www.sectorspdr.com/sectorspdr/sector/.

Basalto, Nicolas, Bellotti, Roberto, De Carlo, Francesco, Facchi, Paolo, Pantaleo, Ester, & Pascazio, Saverio. 2007. Hausdorff clustering of financial time series. *Physica A: Statistical Mechanics and its Applications*, **379**(2), 635–644.

Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, **31**(3), 307–327.

Box, George EP, & Jenkins, Gwilym M. 1976. *Time series analysis: forecasting and control, revised ed.* Holden-Day.

Canhasi, Ercan, & Kononenko, Igor. 2014. Multi-document summarization via archetypal analysis of the content-graph joint model. *Knowledge and information systems*, **41**(3), 821–842.

Chan, Ben HP, Mitchell, Daniel A, & Cram, Lawrence E. 2003. Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, **338**(3), 790–795.

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

Cutler, Adele, & Breiman, Leo. 1994. Archetypal analysis. *Technometrics*, **36**(4), 338–347.

D'Esposito, Maria R, Palumbo, Francesco, & Ragozini, Giancarlo. 2012. Interval archetypes: a new tool for interval data analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **5**(4), 322–335.

Epifanio, Irene. 2016. Functional archetype and archetypoid analysis. *Computational Statistics & Data Analysis*, **104**, 24–34.

Epifanio, Irene, Vinué, G, & Alemany, Sandra. 2013. Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem. *Computers & Industrial Engineering*, **64**(3), 757–765.

Eugster, Manuel, & Leisch, Friedrich. 2009. From spider-man to hero-archetypal analysis in R. *Computational Statistics & Data Analysis.*

Feld, Sebastian, Werner, Martin, Schönfeld, Mirco, & Hasler, Stefanie. 2015. Archetypes of alternative routes in buildings. *Pages 1–10 of: Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on.* IEEE.

Kaufman, L, & Rousseeuw, PJ. 1990. Finding groups in data, 1990. *New York*, 22–52.

Lawson, CL, & Hanson, RJ. 1974. Soling least square problems Prentice-Hall. *Englewood C liffs NJ, 340pp.*

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

Liao, T Warren. 2005. Clustering of time series data - a survey. *Pattern recognition*, **38**(11), 1857–1874.

Markowitz, Harry. 1952. Portfolio selection. *The Journal of Finance*, **7**(1), 77–91.

Midgley, David, & Venaik, Sunil. 2013. Marketing strategy in MNC subsidiaries: pure versus hybrid archetypes. *Pages 215–216 of: 55th Annual Meeting of the Academy of International Business.* AIB Executive Secretariat.

Mørup, Morten, & Hansen, Lars Kai. 2012. Archetypal analysis for machine learning and data mining. *Neurocomputing*, **80**, 54–63.

Porzio, Giovanni C, Ragozini, Giancarlo, & Vistocco, Domenico. 2008. On the use of archetypes as benchmarks. *Applied Stochastic Models in Business and Industry*, **24**(5), 419–437.

Quantcuote. 2017 (09). *QuantQuote Free Historical Stock Data web-site.* Retrieved on 15/09/2017 from https://quantquote.com/historical-stock-data.

Ramsay, James O, & Silverman, Bernard W. 2002. *Applied functional data analysis: methods and case studies.* Vol. 77. Springer New York.

Rousseeuw, Peter J, & Kaufman, L. 1990. *Finding Groups in Data.* Wiley Online Library.

Seiler, Christian, & Wohlrabe, Klaus. 2013. Archetypal scientists. *Journal of Informetrics*, **7**(2), 345–356.

Steinschneider, Scott, & Lall, Upmanu. 2015. Daily precipitation and tropical moisture exports across the eastern United States: An application of archetypal analysis to identify spatiotemporal structure. *Journal of Climate*, **28**(21), 8585–8602.

Stone, Emily. 2002. Exploring archetypal dynamics of pattern formation in cellular flames. *Physica D: Nonlinear Phenomena*, **161**(3), 163–186.

Thøgersen, Juliane Charlotte, Mørup, Morten, Damkiær, Søren, Molin, Søren, & Jelsbak, Lars. 2013. Archetypal analysis of diverse Pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. *BMC bioinformatics*, **14**(1), 279.

Thorndike, Robert L. 1953. Who belongs in the family? *Psychometrika*, **18**(4), 267–276.

Thurau, Christian, & Bauckhage, Christian. 2009. Archetypal images in large photo collections. *Pages 129–136 of: Semantic Computing, 2009. ICSC'09. IEEE International Conference on.* IEEE.

Thurau, Christian, Kersting, Kristian, Wahabzada, Mirwaes, & Bauckhage, Christian. 2012. Descriptive matrix factorization for sustainability adopting the principle of opposites. *Data Mining and Knowledge Discovery*, **24**(2), 325–354.

Tsanousa, Athina, Laskaris, Nikolaos, & Angelis, Lefteris. 2015. A novel single-trial methodology for studying brain response variability based on archetypal analysis. *Expert Systems with Applications*, **42**(22), 8454–8462.

Tseng, Jie-Jun, & Li, Sai-Ping. 2011. Asset returns and volatility clustering in financial time series. *Physica A: Statistical Mechanics and its Applications*, **390**(7), 1300–1314.

Vinue, Guillermo. 2015. Anthropometry: An R package for analysis of anthropometric data. *R package version*, **1.5**.

Vinue, Guillermo, Epifanio, Irene, & Alemany, Sandra. 2015. Archetypoids: A new approach to define representative archetypal data. *Computational Statistics & Data Analysis*, **87**, 102–115.

Yahoo. 2017 (09). *Yaho Finance web-site.* Retrieved on 15/09/2017 from https://es.finance.yahoo.com/lookup.

Zhao, Genping, Jia, Xiuping, & Zhao, Chunhui. 2015. Multiple endmembers based unmixing using archetypal analysis. *Pages 5039–5042 of: Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International.* IEEE.