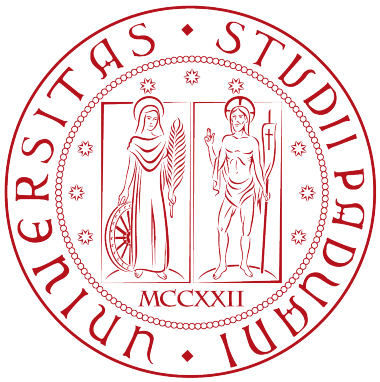

Università degli Studi di Padova

Scuola Galileiana di Studi Superiori
Classe di Scienze Naturali

From data to firm's ecology

Laureando:
Riccardo Della Vecchia

Relatore:
Prof. Matteo Marsili
ICTP, Trieste

Contents

Chapter 1

Introduction

The recent financial crisis and subsequent spillovers in the real economy suggests that conventional economic policy measures have lost their grip on the real economy. One important factor is that the economic ecosystem has undergone profound changes: "habitats" that were once fragmented have now become part of the same global economy and the IT revolution, besides creating economic niches for new types of firms, has reshaped the way in which firms interact as well as the pace at which interaction takes place.

The rise of finance, securisation and credit markets seems to have provoked a decoupling of the financial sector from the real economy. The graph below shows the dynamic of the total assets for US firms in the financial and non financial sector. The dataset spans the period 1979-2014. The dynamics of the firm size distribution exhibits a marked difference in the distribution of asset sizes in the two sectors after 1995.

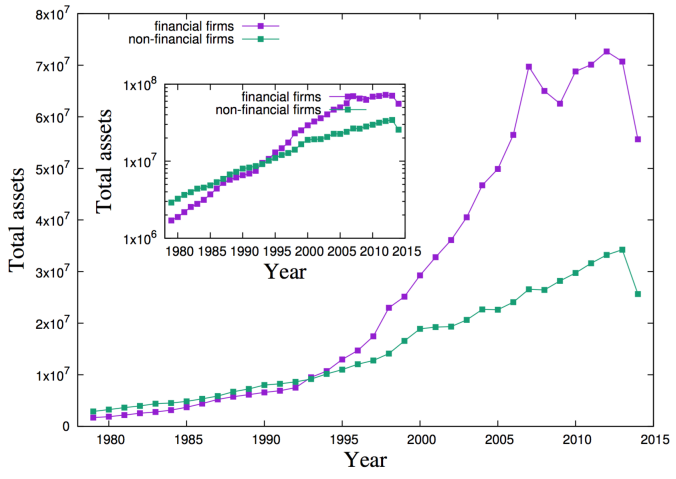

Figure 1.1: Result of the analysis of a large dataset of US firms over the period 1979-2014

Furthermore, both economic sectors are becoming more and more complex and far less predictable than they used to be. Yet many firms still pursue classic approaches to strategy that were designed for more stable times, emphasizing analysis and planning focused on maximizing short-term performance rather than long-term robustness. The result is that firms are dying younger than they used to do (Figure 1.2). This problem has been studied in detail by the authors of [8] and in this paper they describe some principles that confer robustness to companies exploiting analogies with biology and natural systems. Their claim is that understanding these principles can mean the difference between survival and extinction.

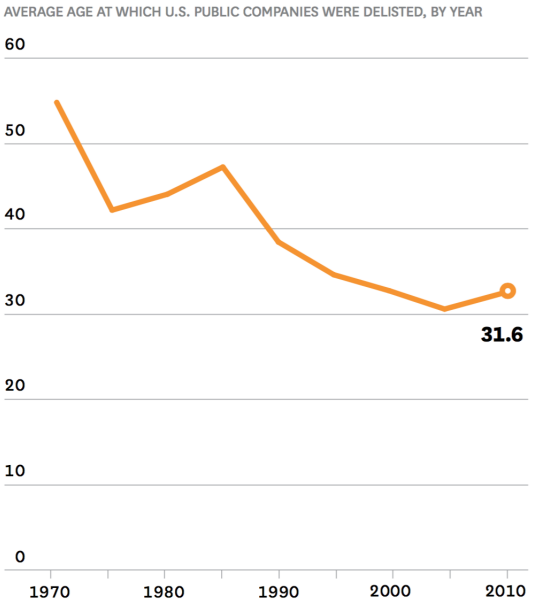

Figure 1.2: Companies are dying younger because they are failing to adapt to increasing complexity. They have a one in three chance of being delisted in the next five years, whether because of bankruptcy, liquidation, or other causes. That's six times the delisting rate of companies 40 years ago. This graph is taken from [8].

How, then, can companies flourish and persist? In [8] the authors suggest that companies are identical to biological species in an important respect: both are what's known as complex adaptive systems. Therefore, the principles that confer robustness in these systems, whether natural or manmade, are directly applicable to business.

In a complex adaptive system, local events and interactions among the "agents", whether ants, trees, or people, can cascade and reshape the entire system, a property called emergence. The system's new structure then influences the individual agents, resulting in further changes to the overall system. Thus the system continually evolves in hard-to-predict ways through a cycle of local interactions, emergence, and feedback. In nature we see this play out

when ants of some species, for example, although individually following simple behavioural rules, collectively create "supercolonies" of several hundred million ants covering more than a square kilometer of territory. In business we see workers and management, through their local actions and interactions, shape the overall structure, behavior, and performance of a firm.

Complexity therefore exists at multiple levels, not just within organizational boundaries; and at each level there is tension between what is good for an individual agent and what is good for the larger system.

Motivated by our interest for the changing environment in which firms live and struggle to survive, we start with the presentation of the results of our analysis. In this analysis we have tried to get a better understanding of the global economy through a classification of the firms that contribute to shape it. Our objective is to understand what makes firms similar and correlated. Therefore, our analysis is heavily based on clustering techniques and more generally on a rigorous statistical analysis of the data.

Chapter 2

A quantitative perspective into firms' ecology

Data clustering deals with the problem of classifying a set of N objects into groups so that objects within the same group are more similar than objects belonging to dfferent groups. Each object is identified by a number D of measurable features: Hence object $i = 1, \ldots, N$ can be represented as a point $\vec{x}_i = \left(x^{(1)}, \ldots, x^{(D)}\right)$ in a D-dimensional space. Data clustering aims at identifying clusters as more densely populated regions in this vector space. More precisely, a configuration of clusters is represented by a set $\mathcal{S} = \{s_1, \ldots, s_N\}$ of integer labels, where s_i is the cluster to which object i belongs.

The classic approaches to data clustering are partitioning methods and hierarchical clustering. Partitioning methods are based on two elements: (1) a distance between objects, which allows one to measure their similarity and (2) a cost function whose minima correspond to "optimal" clustering configurations. For example, a typical K-means (KM) approach takes as cost function the sum of squared distances of objects to the centroid $\vec{X}_s$ of the cluster in which they are classified [5]:

$$H_{KM}\{\mathcal{S}\} = \sum_s \sum_{i:s_i=s} \|\vec{x}_i - \vec{X}_s\|^2, \quad \vec{X}_s = \frac{1}{n_s} \sum_{i:s_i=s} \vec{x}_i, \tag{2.1}$$

where $n_s = \sum_{i:s_i=s} 1$ is the number of objects in cluster s, and $\|\vec{x}\|^2 = \sum_j (x^{(j)})^2$ defines the euclidean distance.

Apart from cases where a cost function is naturally suggested by the problem itself, the choice of the cost function or of a distance is an element of arbitrariness. Also the problem of finding the cluster structure which minimizes the cost function may be quite difficult: depending on the form of the cost function and on the data set the "cost landscape" can be either simple, with a single minimum which is easily accessible dynamically, or complex, with many metastable minima. Data clustering methods also differ for the specific algorithms used to reach a (local) minimum.

A second, very popular approach to data clustering, called hierarchical clustering, is based on the definition of a distance between objects and clus-

ters of objects and a very simple algorithm: Given a configuration with $K > 1$ clusters, it merges the two closest clusters into a single one. In this way, starting from the configuration with $K = N$ clusters, the algorithm generates a sequence of configurations as K varies from N to 1. This sequence of configurations and their hierarchic organization, can be represented by a convenient and compact graphical tool called dendrogram [4].

Expectation Minimization is a further approach [6] where the density of points is modelled as a mixture of Gaussians whose centering and scale parameters are fit by maximum likelihood.

The algorithm that we use for the clustering has been devised by the authors of [3] and has some similarity with Expectation Maximization data clustering [6], which is also based on likelihood maximization. However, the statistical hypothesis in [6] is very different from the one in our model. The algorithm is a fully unsupervised, parameter-free approach to data clustering which derives from a maximum likelihood principle.

2.1 The model and expectation maximisation algorithm

The data contains a set of firms, each characterised by a time series $x_i(t)$ where $i = 1, \ldots, N$ runs over the firms and t over the years covered in the dataset. Firm i is active from year t_i, which is the first year for which data is available, to year $t_i + \tau_i$ where τ_i is its lifetime. $x_i(t)$ is a multidimensional vector, but it mainly covers two dimensions: *i)* how the firm interacts with the economy (sales, profits) and *ii)* how it interacts with the financial market (share price/return).

We want to identify firms who share a similar behaviour and classify them accordingly. In order to do this, we assume a set of models $s = 1, \ldots, S$. If a firm i is described by model s, then

$$x_i(t) = \mu_s + g_s \eta_s(t) + \sigma_s \epsilon_i(t) \tag{2.2}$$

where $\eta(t)$ and $\epsilon_i(t)$ are Gaussian i.i.d. random variables with zero mean and unit variance and μ_s, g_s and σ_s are parameters (μ_s and g_s are vectors, with a component corresponding to each of the components of x). Furthermore, firms in model s die with probability d_s each year. So each model is described by the parameters $\theta_s = (\mu_s, g_s, \sigma_s, d_s)$. The idea of this model is that it splits the dynamics of the firm into a common part $\mu_s + g_s \eta_s(t)$, which is the same for all $i \to s$, and an individual one $\sigma_s \epsilon_i(t)$.

This allows us to compute the log-likelihood

$$\mathcal{L}(i \to s) = \log P\{x_i | \theta_s\} \tag{2.3}$$

Leaving the technical issue of computing $P\{x_i|\theta_s\}$ to the next section, the problem becomes that of *i)* finding the optimal assignment $i \to s$ for each firm and *ii)* estimating the optimal parameters θ_s^* for each model. This is done following an expectation maximisation procedure:

1. start from an initial assignment $i \to s$ for all i (e.g. a random one)

2. Estimate the optimal parameters θ_s^* for each model

$$\theta_s^* = \arg\max_{\theta_s} \mathcal{L}(i \to s). \tag{2.4}$$

3. For each firm update the assignment as the assigned that maximises the likelihood:

$$i \to \arg\max_{s} \mathcal{L}(i \to s) \tag{2.5}$$

with the updated parameters computed in the previous step.

4. go back to step 2 if the new assignment differs from the old one for at least one firm. Otherwise exit.

Once the optimal assignment is determined, one can look into each cluster s in order to characterise its defining properties (e.g. which firms $i \to s$? How many of these are active in year t, what are the parameters?). It is also possible to estimate the dynamical pattern $\hat{\eta}_s(t)$ that is common to all firms in cluster s, e.g. as

$$\hat{\eta}_s(t) = \frac{1}{g_s |\{i \to s\}|} \sum_{i \to s} x_i(t), \tag{2.6}$$

and to understand what is the interaction between different clusters. Thinking of each cluster as a species, this provides a detailed characterisation of the ecology.

2.2 Maximum likelihood estimation

In section 2.2 at step 2 we stated that our goal is to estimate the optimal parameters θ_s^* for each model. The way in which we want to do it is maximum likelihood estimation (MLE). Let us call the data set $\hat{X} = \{x_i(t)\}_{i=1\ldots N}^{t=1\ldots T}$, while $\mathcal{S} = \{s_i\}_{i=1}^N$ describes the clustering and the set $\Theta \equiv \{\theta_s\}$ contains all the parameters in the model. With a slightly different notation from before, the likelihood reads

$$P\{\hat{X}|\mathcal{S},\Theta\} = \prod_{t=1}^{T} \left\langle \prod_{i=q}^{N} \delta\Big(x_i(t) - \mu_s - g_s\eta_s(t) - \sigma_s\epsilon_i(t)\Big) \right\rangle, \tag{2.7}$$

where the average above is over all the η's and ϵ's. Gaussian integration and elementary algebra leads to the final expression for the log-likelihood,

$$\begin{aligned} \mathcal{L}(\mathcal{S},\Theta) \quad = \quad & \log P\{\hat{X}|\mathcal{S},\Theta\} = -\frac{1}{2}\sum_t \sum_s \Bigg\{ \log\left(2\pi \frac{\sigma_s^2 + n_s(t) g_s^2}{\sigma^{2(1-n_s)}}\right) \\ & + \frac{\sum_{i:s_i=s} x_i(t)^2 + n_s(t)\mu_s^2 - 2\mu_s \sum_{i:s_i=s} x_i(t)}{\sigma_s^2} \\ & - \frac{(\mu_s n_s(t) - \sum_{i:s_i=s} x_i(t))^2 g_s^2}{\sigma_s^2(\sigma_s^2 + n_s(t) g_s^2)} \Bigg\}. \end{aligned} \tag{2.8}$$

As we said, firm i is active from year t_i, which is the first year for which data is available, to year $t_i + \tau_i$ where τ_i is its lifetime. Therefore, in expression (2.8) $n_s(t)$ indicates the number of firms which are active in sector s at time t. Furthermore, summations over the firms in a certain sector should a priori comprise only those firms that are active at time t in that sector. So, one way to make notation in equation (2.8) rigorous is to put to zero all the points in the time series which lie outside the lifespan of firm i, $x_i(t) \equiv 0$ for $t < t_i$ or $t > t_i + \tau_i$. In this way the summation can be taken over all firms in the generic sector s even if they are inactive at time t.

The maxima of these function are the points in which all the partial derivatives simultaneously equal zero. This gives a non-linear system of equations:

$$\frac{\partial \mathcal{L}}{\partial g_s^2} = -\frac{1}{2}\sum_t \left\{ \frac{n_s}{\sigma_s^2 + n_s g_s^2} - \left(\frac{\mu n_s - \sum_{i:s_i=s} x_i}{\sigma_s^2 + n_s g_s^2} \right)^2 \right\} = 0, \tag{2.9}$$

$$\frac{\partial \mathcal{L}}{\partial \mu_s} = -\sum_t \frac{\mu_s n_s - \sum_{i:s_i=s} x_i}{\sigma_s^2 + n_s g_s^2} = 0, \tag{2.10}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \sigma_s^2} = -\frac{1}{2}\sum_t \Bigg\{ & \frac{\sigma_s^4(-\mu_s^2 n_s + n_s\sigma^2 + 2\mu_s \sum x_i - \sum x_i^2)}{\sigma_s^4(g_s^2 n_s + \sigma_s^2)^2} \\ & + \frac{g_s^4 n_s(n_s^2\sigma_s^2 + \sum x_i^2 - n_s(\sigma_s^2 + \sum x_i^2))}{\sigma_s^4(g_s^2 n_s + \sigma_s^2)^2} \\ & + \frac{g_s^2\sigma_s^2(2n_s^2\sigma_s^2 + 2\sum x_i^2 - n_s(\sigma_s^2 + 2\sum x_i^2))}{\sigma_s^4(g_s^2 n_s + \sigma_s^2)^2} \Bigg\} = 0. \end{aligned} \tag{2.11}$$

This system of equation is too difficult to be solved analytically. Anyway, equation (2.10) can be rewritten making explicit the dependence of μ_s on σ_s and g_s:

$$\mu_s = \frac{\sum_t \frac{\sum_{i:s_i=s} x_i(t)}{\sigma_s^2 + n_s(t) g_s^2}}{\sum_{t'} \frac{n_s(t')}{\sigma_s^2 + n_s(t') g_s^2}}. \tag{2.12}$$

Therefore, to complete the task of finding the optimal parameters θ_s^* we have to rely on numerical methods.

2.3 Maximisation over the clustering assignment

The log-likelihood (2.8)can be written in the following form:

$$\mathcal{L}(\mathcal{S}, \Theta) = \sum_{s=1}^{S} l(\theta_s, \vec{X}_s). \tag{2.13}$$

$\vec{X}_s$ indicates the time series of the firms that belong to the s-th sector.

Our algorithm is composed of the following steps:

1. Select a permutation of the firms which gives you the order in which you'll take firms in the next step.

2. Take the next firm in the order given by the permutation. Suppose you take firm i, then move firm i from its sector to all the others $i : s_i = s \to s' \quad \forall s'$ and for each of them compute $\Delta\mathcal{L}(i : s \to s')$, where

$$\Delta\mathcal{L}(i : s \to s') = l(\theta_s, \vec{X}_{s\setminus\{i\}}) + l(\theta_{s'}, \vec{X}_{s'\cup\{i\}}) - l(\theta_s, \vec{X}_s) - l(\theta_{s'}, \vec{X}_{s'}). \tag{2.14}$$

3. Change the cluster to which i belongs in order to choose the move that gives the biggest $\Delta\mathcal{L}$.

4. Exit if you have just considered the last firm in the permutation, otherwise go back to step 2.

Chapter 3

Preliminary results

3.1 The dataset

Compustat is a database of financial, statistical and market information on active and inactive global companies throughout the world. The service began in 1962 and provides a broad range of information products directed at institutional investors, universities, bankers, advisors, analysts, etc. Our dataset is composed of 23,566 firms with annual history available from 1950 to 2013.

Our algorithm needs some initial division into clusters to start. Then, in an iterative way, the program converges to an optimal solution with the firms divided into the best fitting clusters. A reasonable starting point is the one given by a classical division into economic sectors. What we use is the Standard Industrial Classification (SIC), a system for classifying industries by a four-digit code. Established in the United States in 1937, it is used by government agencies to classify industry areas. The SIC codes can be grouped into progressively broader industry classifications: industry group, major group, and division. The first 3 digits of the SIC code indicate the industry group, and the first two digits indicate the major group. In table 3.1 we can finally see that each division encompasses a range of SIC codes.

After looking to the data, we have tried to aggregate sectors with just a few firms respect to the others in orderer to diminish the number of clusters and to start with more or less equally populated sectors. This led us to the division into sectors in table 3.2 which is also our initial clustering. To denote a sector we switch to a notation with just one digit, and we get sectors from 0 to 6 for a total of seven clusters.

The database we use is very rich in information, for each company we have its history from 1950 to 2013 in terms of its market value, net sales, COGS, EBIT, TSR, etc. We decided to use market values for this simulation. In fact, it is very interesting to observe the correlation of firms through their interaction with the financial sector.

The model we use for the fit was presented in chapter 2.2

$$x_i(t) = \mu_s + g_s \eta_s(t) + \sigma_s \epsilon_i(t). \tag{3.1}$$

The quantity $x_i(t)$ is almost directly read from the dataset, in fact it is defined

SIC Codes	Division
0100-0999	Agriculture, Forestry and Fishing
1000-1499	Mining
1500-1799	Construction
1800-1999	not used
2000-3999	Manufacturing
4000-4999	Transportation, Communications, Electric, Gas and Sanitary service
5000-5199	Wholesale Trade
5200-5999	Retail Trade
6000-6799	Finance, Insurance and Real Estate
7000-8999	Services
9100-9729	Public Administration
9900-9999	Nonclassifiable

Table 3.1: To look at a particular example of the hierarchy, SIC code 2024 (ice cream and frozen desserts) belongs to industry group 202 (dairy products), which is part of major group 20 (food and kindred products), which belongs to the division of manufacturing.

Sector #	Economic areas
0	Agriculture, Forestry and Fishing
1	Mining, Construction, not used
2	Manufacturing
3	Transportation, Communications, Electric, Gas and Sanitary service
4	Wholesale Trade, Retail Trade
5	Finance, Insurance and Real Estate
6	Services, Public Administration, Nonclassifiable

Table 3.2: The arbitrary division into 7 sectors of all the firms in our database.

as

$$x_i(t) \equiv \log\left(\frac{p_i(t)}{p_i(t-1)}\right), \tag{3.2}$$

where $p_i(t)$ is the market value of the i-th firm at the year t. A normal distribution cannot be used to model stock prices because it has a negative side and stock prices cannot fall below zero while the log-normal is by definition a positive distribution. Therefore, we assume a gaussian distribution for the logarithm of the returns which is equivalent to assume a log-normal distribution for the returns.

3.2 Overview of the data

Once the simulation has ended running, we have a new clustering for the firms which allows for a more flexible classification of the firms respect to the one in traditional economic sectors. So we expect of being able to reveal new (maybe unexpected) correlation between firm or at least just test the goodness of our model to fit the data. In figure 3.2 we see how the total number of firms changes over time. Figure 3.1 is useful to compare the trends in different sectors.

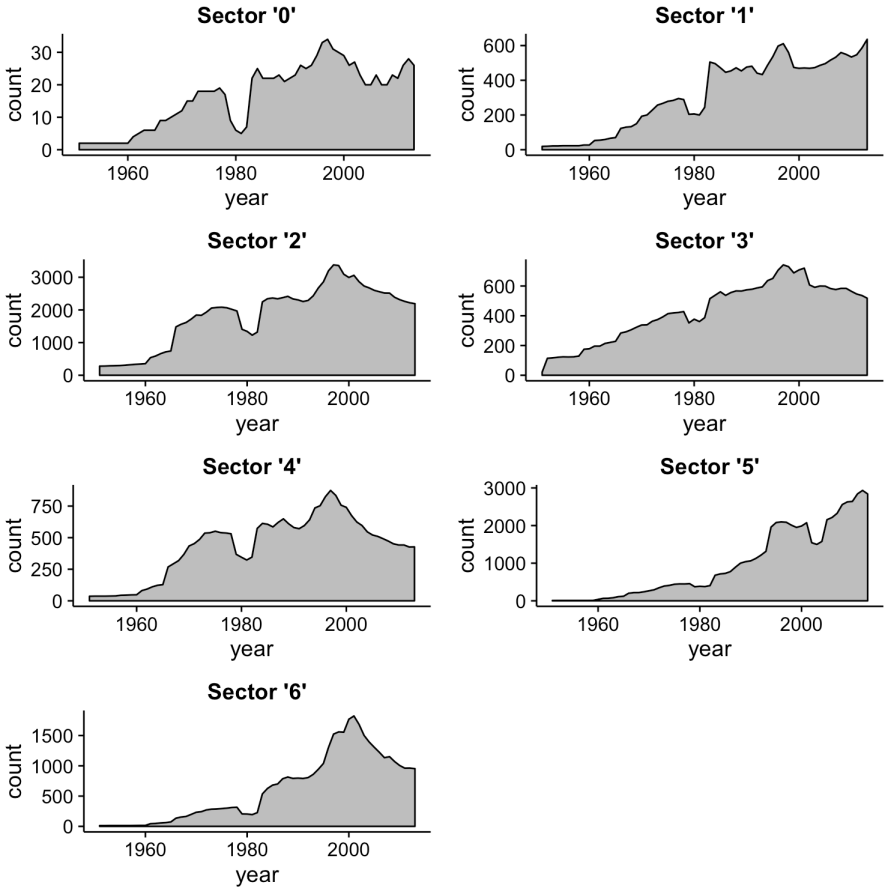

Figure 3.1: Distribution of the active firms in the different economic sectors.

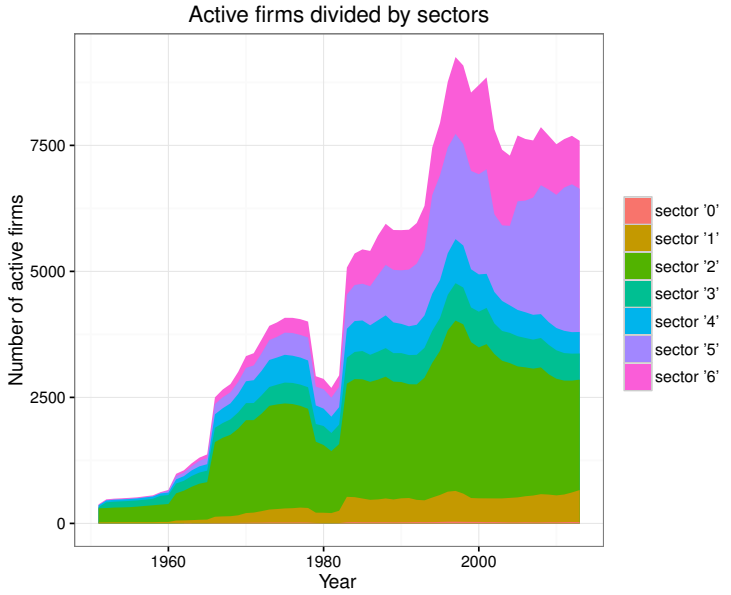

Figure 3.2: Distribution of the active firms firms from 1950 to 2013.

3.3 Distributions into the clusters

Our expectation-maximization algorithm gives back a set of optimal parameters for our model and the optimal clusters. The latter is the quantity of interest of our work. The graph in figure 3.3 is very useful to have a general idea of how different are our clusters from the initial division into classical economic sectors.

Let us start from the left-hand side of figure 3.3 to see what happens to the different sectors of table 3.2. The firm with SIC code (the first two digits) between 10 and 19 are classified in the table under the name "sector 1" and the corresponding economic activities are mining and construction. This sector does not spread very much in other different clusters, in fact the highest density for these sectors in the final clusters coincides again with the "EM cluster" number 1. Notice that also other clusters are populated by the firms that were initially in sector 1, like 3 and 6, even if in a less conspicuous way. Manufacturing corresponds to SIC between 20 and 39. Interestingly this sector splits into two and we have to peaks of population in correspondence of clusters 1 and 2 and in a more tenuous way in clusters 3 and 4. Transportation, Communications, Electric, Gas and Sanitary service correspond to the 3-rd economic sector in our notation with SIC code between 40 and 49. Firms in this sector tends to remain in the same cluster and the same holds in an approximate way for the 4-th sector, Wholesale Trade and Retail Trade that span SIC codes from 50 to 59. The behaviour of the financial sector is instead very different and marked. Firms in the 5-th sector divide between cluster 5 and cluster 0 for the majority. But there is also a non negligible part which occupies the

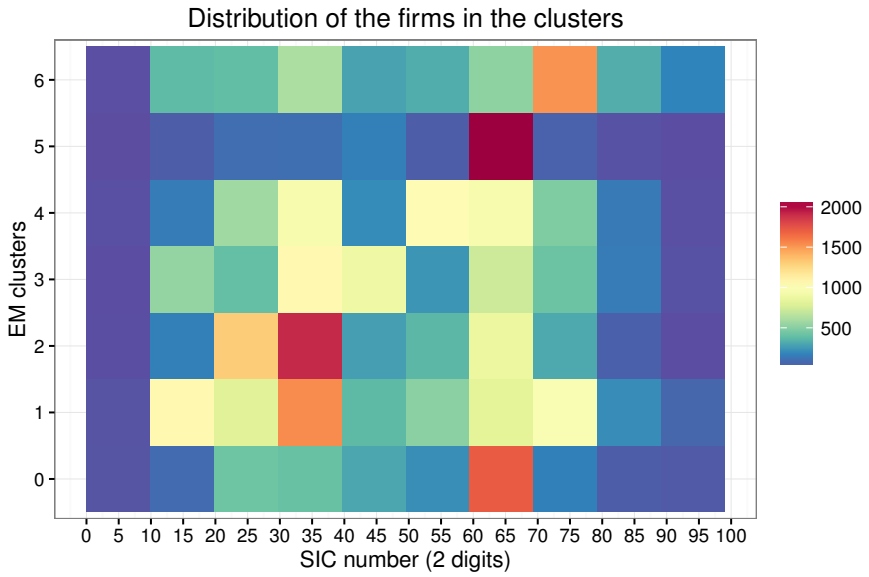

Figure 3.3: The chart shows the concentration of firms classified with the same SIC code in the different clusters obtained through an EM algorithm.

clusters between 1 and 4. Finally, services and the public administration tend to split between cluster 1 and 6.

To sum up, there seems to be a certain degree of similarity between the initial division in classical economic sectors and our more flexible division based on the correlation between firms. Nonetheless, some interesting phenomena appear evident. The financial sector is the one which occupies the greater majority of the other clusters so it seems to be very heterogeneous and very much correlated in its components with all the other sectors.

In this discussion we started from the initial division into sectors. It is really interesting looking also at the composition of the clusters that we have obtained from the EM algorithm. We investigate this relation in the following graphs but we also add a physical dimension in our analysis: time. Therefore, in these charts we can see the changing in the total number of firms that are active at a certain point in time in a certain cluster. Furthermore, for each year we don't just look at the total number of firms in that sector but we also see its composition in term of the sectors to which the firms belong.

The general trend seems to be an increasing one but with lots of different peculiarities.

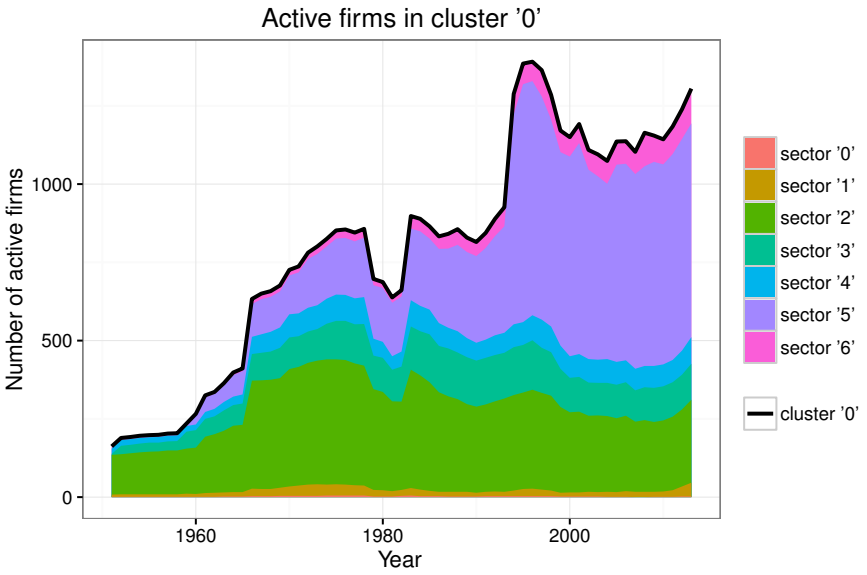

Figure 3.4: Distribution of the active firms in cluster 0 divided by sectors.

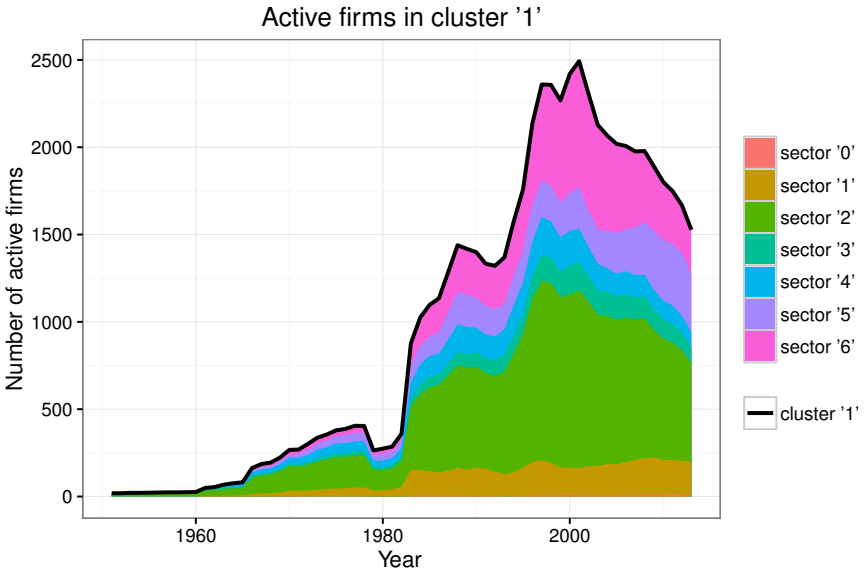

Figure 3.5: Distribution of the active firms in cluster 1 divided by sectors.

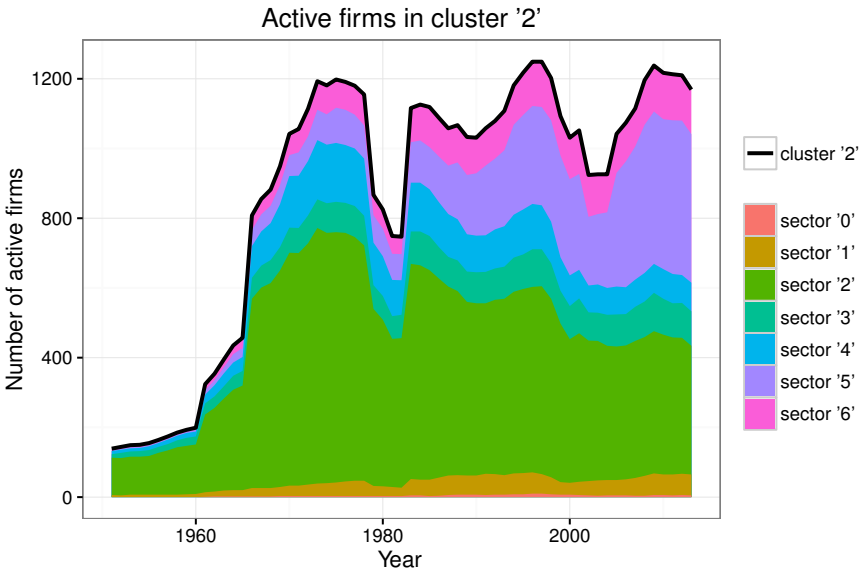

Figure 3.6: Distribution of the active firms in cluster 2 divided by sectors.

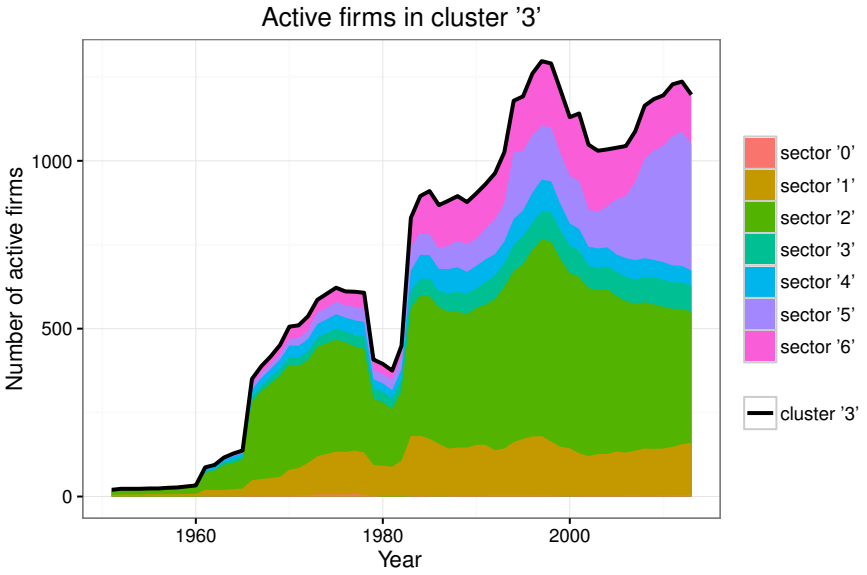

Figure 3.7: Distribution of the active firms in cluster 3 divided by sectors.

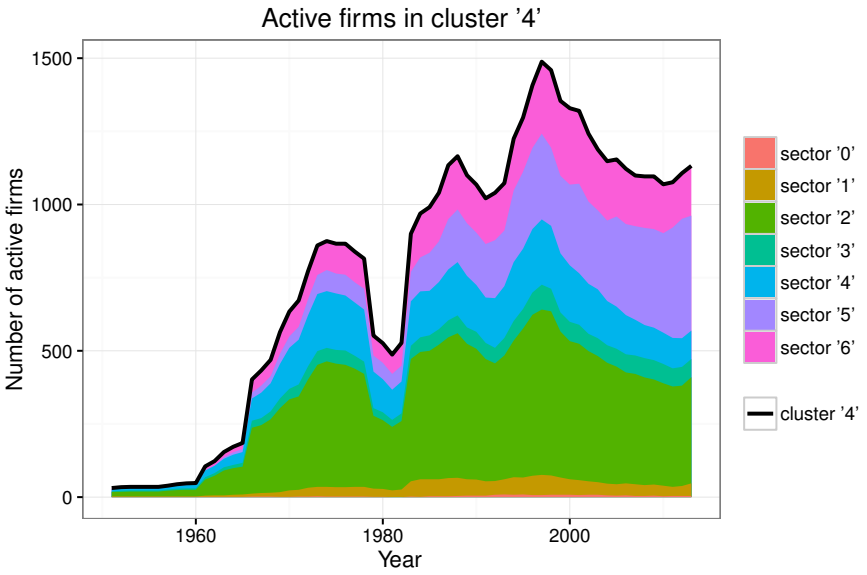

Figure 3.8: Distribution of the active firms in cluster 4 divided by sectors.

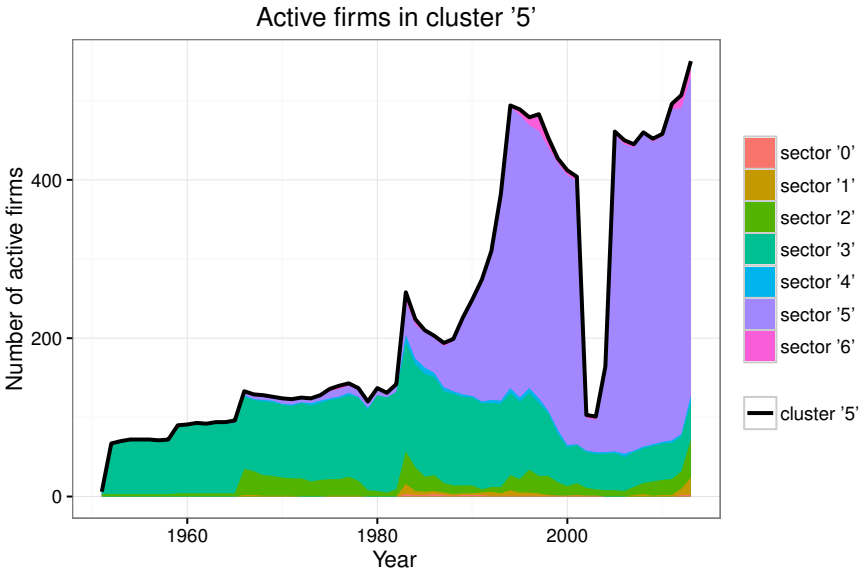

Figure 3.9: Distribution of the active firms in cluster 5 divided by sectors.

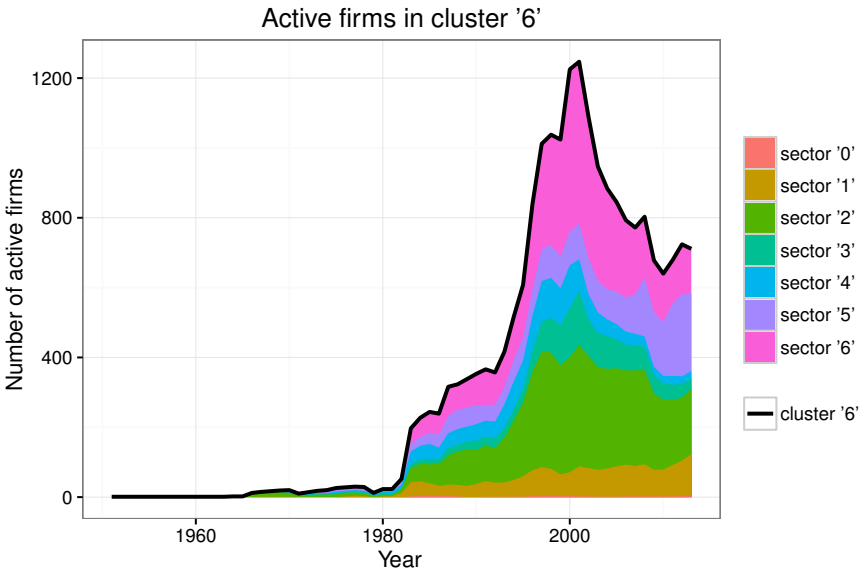

Figure 3.10: Distribution of the active firms in cluster 6 divided by sectors.

3.4 Stability of the algorithm

In our preceding analysis we assumed that our division into cluster is a "small perturbation" respect to the traditional division into economic sectors. For this reason, we started from the economic division and we have made our algorithm start from that. Actually, we have no guarantee that our algorithm is converging to the global maximum of the likelihood. Many problems can occur if the likelihood is particularly flat in one direction due to the fact that in a computational algorithm truncations and approximations occur. The best way to check the validity of the preceding results is to run another simulation starting in this case from an initial clustering which is initialised randomly. If our algorithm is efficient and stable we expect to find, also in this case, approximately the same result of the preceding chapter.

Anyway, we have to keep in mind that there are firms whose time series are completely empty. These firms are invisible to the likelihood, that means that they do not contribute to make it bigger nor smaller. For this reason once we assign the initial cluster to which these firms belong, they stay in that sector till the end of the simulation. Therefore, when we start from the division provided by the SIC there is a quantity of firms that belongs, before and after, to the same cluster. The same holds when we initialise the sectors randomly. We highlight that these kind of time series are a negligible amount respect to the total number of the timeseries. Still, this represent a source of noise when we compare the charts in section 3.3 with the graphs of this section.

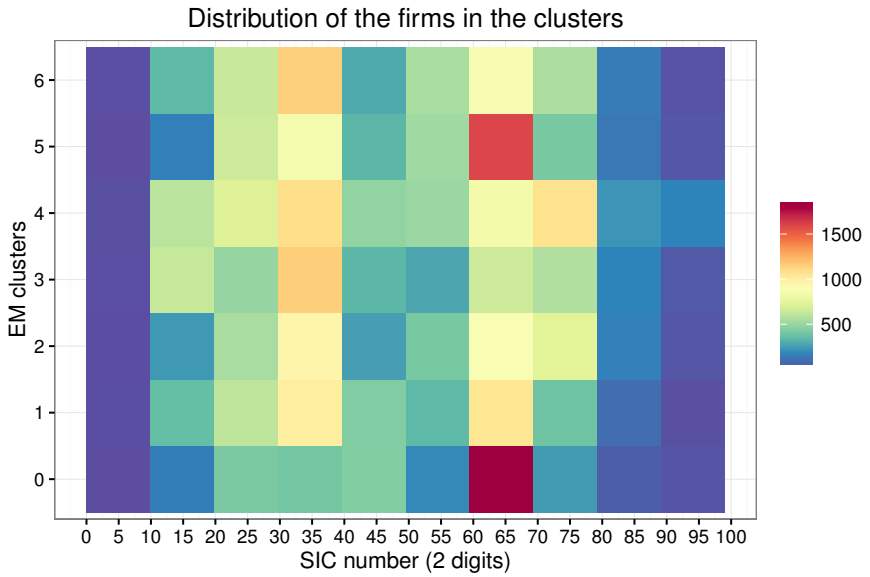

Figure 3.11: The chart shows the concentration of of firms classified with the same SIC code in the different clusters obtained through an EM algorithm and starting with a random initial clustering.

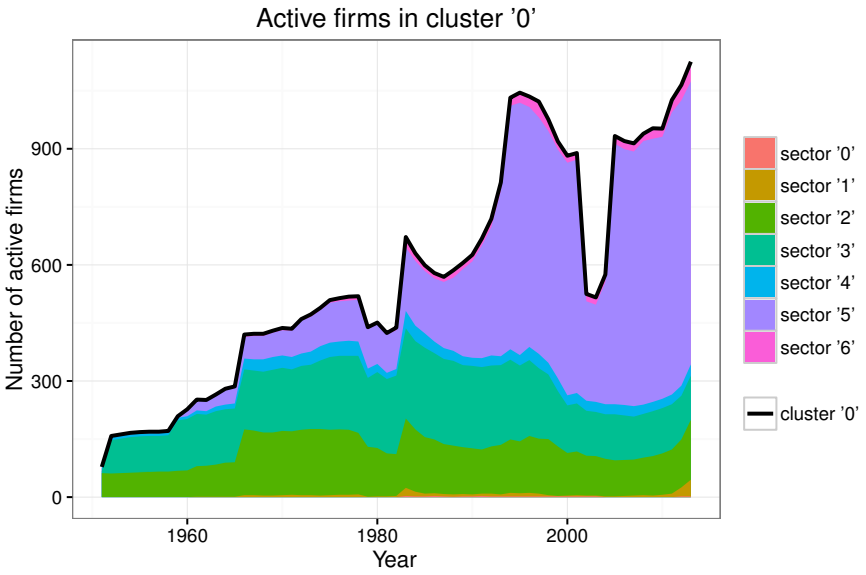

Figure 3.12: Distribution of the active firms in cluster 0 divided by sectors. The initial clustering here is a random one.

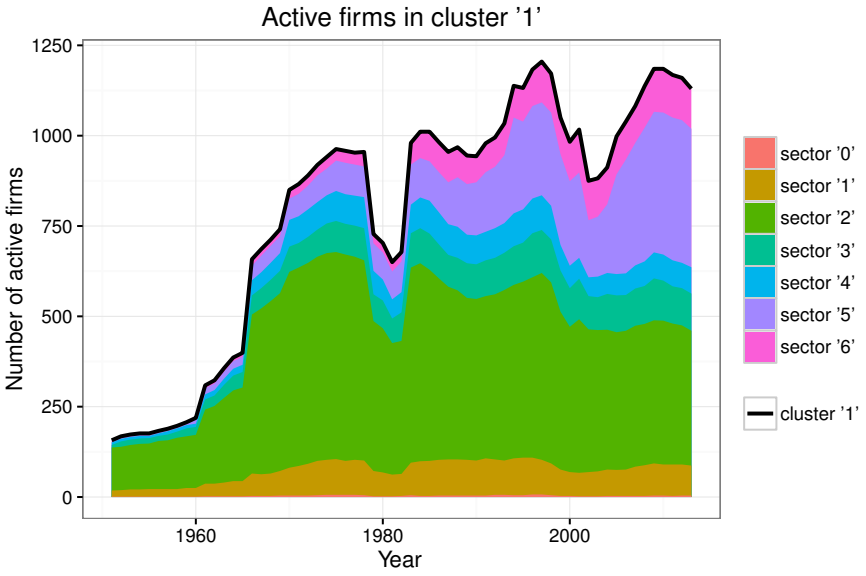

Figure 3.13: Distribution of the active firms in cluster 1 divided by sectors. The initial clustering here is a random one.

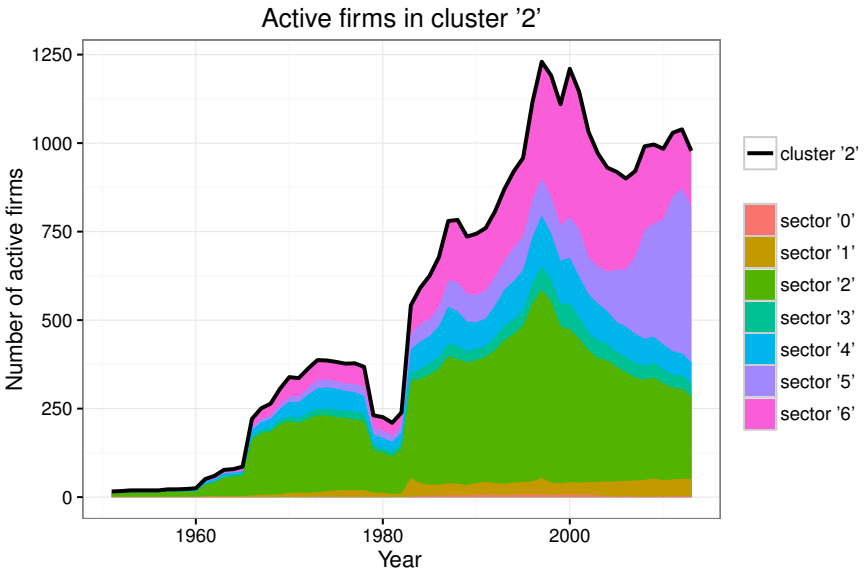

Figure 3.14: Distribution of the active firms in cluster 2 divided by sectors. The initial clustering here is a random one.

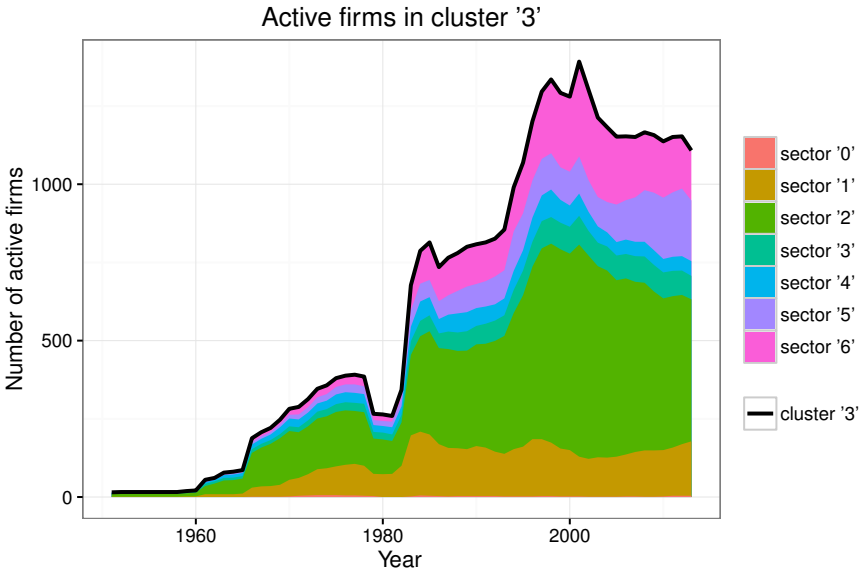

Figure 3.15: Distribution of the active firms in cluster 3 divided by sectors. The initial clustering here is a random one.

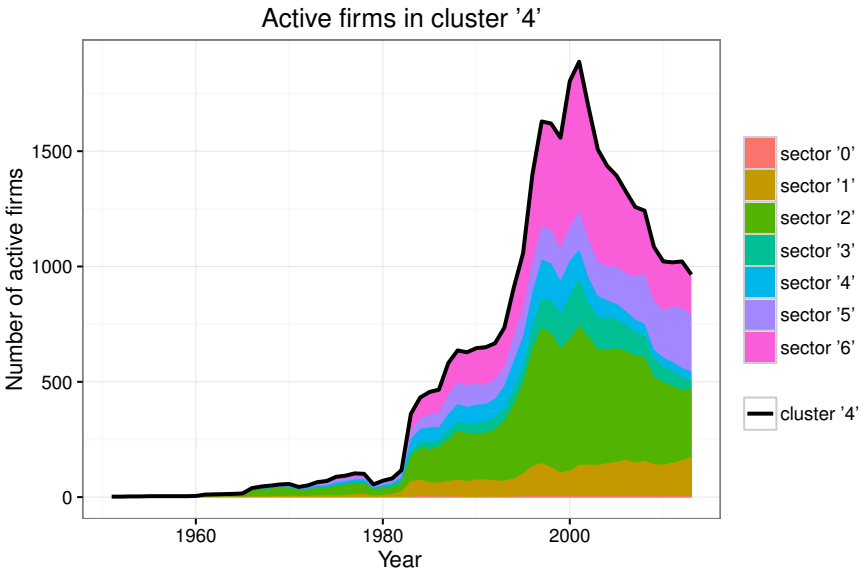

Figure 3.16: Distribution of the active firms in cluster 4 divided by sectors. The initial clustering here is a random one.

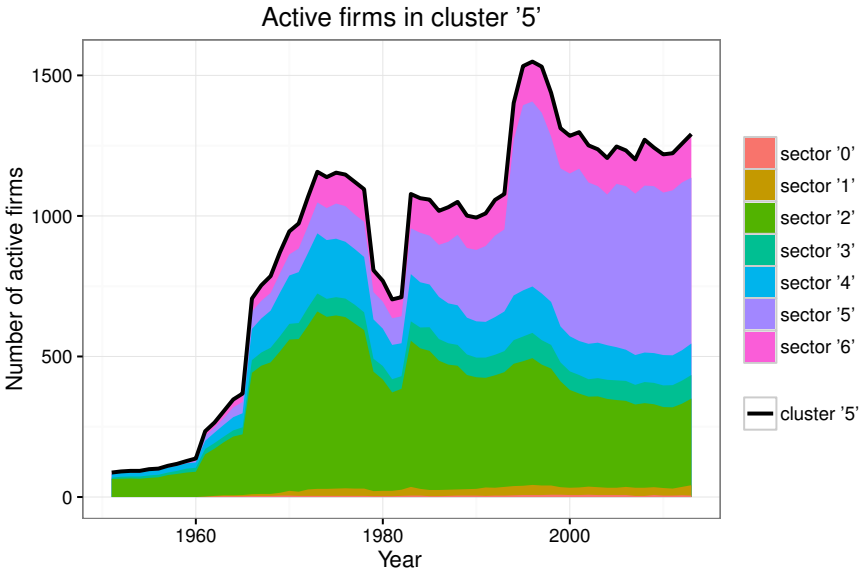

Figure 3.17: Distribution of the active firms in cluster 5 divided by sectors. The initial clustering here is a random one.

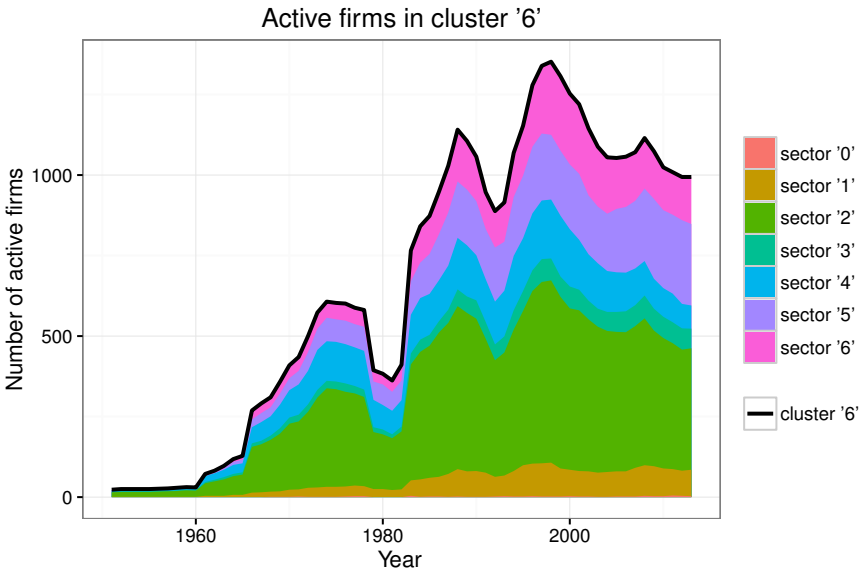

Figure 3.18: Distribution of the active firms in cluster 6 divided by sectors. The initial clustering here is a random one.

Comparing these with the previous graphs we find a good resemblance. In fact, we are able to identify what we think are the corresponding clusters. For some clusters this is very easy, for others the task is more difficult and the correspondence is more uncertain.

Initial Clustering	
Random	SIC
4	6
5	0
0	5
1	3
2	1
3	2
6	4

Table 3.3: On a qualitative basis we can establish a correspondence between the graphs generated by EM algorithm starting from a random initial assignment of the sectors and the one provided by SIC. We argue that the correspondence is the good enough to assert that both running of the program converge approximately to the same cluster configuration.

For example, the similarity between the trend in figure 3.12 and the one in figure 3.9 is absolutely evident. These two clusters have a very similar trend and almost equally populated. Their composition is basically the same, in fact both clusters have a great percentage of firms belonging to the financial industry (sector 5 in our classification) after the 80's while before that they

are dominated by sector 3 which comprises transportations, communications and electric, gas and sanitary services. Take now cluster 4 in figure 3.16 and compare it with cluster 6 in figure 3.10. In both cases there is a tiny number of firms active in these clusters before 1980, but after that the number grows very fast and finally goes back again after the first years of 2000s.

We could speculate a lot on the kind of firms having similar behaviours and we could also check them one by one to see the clusters in which they ended up. Anyway, at this stage we want to maintain a broader view on the landscape of firms composing the clusters. This brings us to ask ourselves the following questions: how much are firms representative of their sectors? How much do they fit that sector? How much noise is in the data and how much information? We try to answer these important questions in the next section.

3.5 Relevance of the classification

In this section we want to understand how well firms in the same cluster fit together. The way to do it is through the likelihood function. In the notation of section 2.2, the likelihood $\mathcal{L}\{\mathcal{S}, \Theta | \hat{X}\}$ is a function of the parameters Θ and of the clusters $\mathcal{S}$ given the data $\hat{X}$. Now, we want to evaluate this function for the optimal parameters and the optimal clusters given by the EM algorithm and we want to evaluate it when the firm i-th is present and when it is removed from the list. What we mean is that, for each firm i we evaluate the likelihood with and without the i-th firm in the computation. In general, when we remove one firm from the computation of the likelihood three things can happen: the likelihood stays equal, it increases or it decreases. In each case we evaluate how much it does. So, when in the following we talk about likelihood significance what we mean is the difference of the likelihood computed with firm i and the likelihood computed without:

$$\text{"Likelihood significance for the i-th firm"} \equiv \mathcal{L} - \mathcal{L}(\text{"i-th firm removed"}).$$

The result is very interesting for our analysis and for a future continuance of our work and it is presented in the following graphs. Starting from figure 3.19, first of all we notice there is a great number of firms that take negative likelihood significance values. With our division in seven sectors and the assumption of model 2.2 for the underlying process, lots of firms appear to be just noise, they don't fit very well the cluster to which they are assigned. In fact, for our clustering would be better not to have these data. This is an indication that a consistent part of the time series is just noise. Another possible reason is that the number of cluster that we have chosen is not large enough to contain all the types of time series. The economic environment could be very heterogenous with more peculiarities than our division in just seven clusters can account for. Too many species are forced to occupy the same ecological niche by the lack of other possible classifications.

Let us ignore this issue for the moment and proceed with our analysis. For each value of likelihood significance we can find firms belonging to all sectors. In figure 3.20 and 3.21 we see that even if we restrict to the most significant

firms in term of likelihood, these firms span basically all sectors. Therefore we cannot say one sector is described better by our model than another one. But this was actually expected. Let us turn now to figure 3.22 and successively to figure 3.23 and 3.24. In this case the situation is different because we can tell which are the clusters that end up containing significant firms from the ones that are not doing it. Basically, by looking at figure 3.23 we see that among the most significant 5000 firms just cluster 1, 3, 4 and 6 appear but in very different proportions. If we directly look at figure 3.24 we understand that the clusters which contain firms that are very well correlated between themselves are cluster 1 and 6.

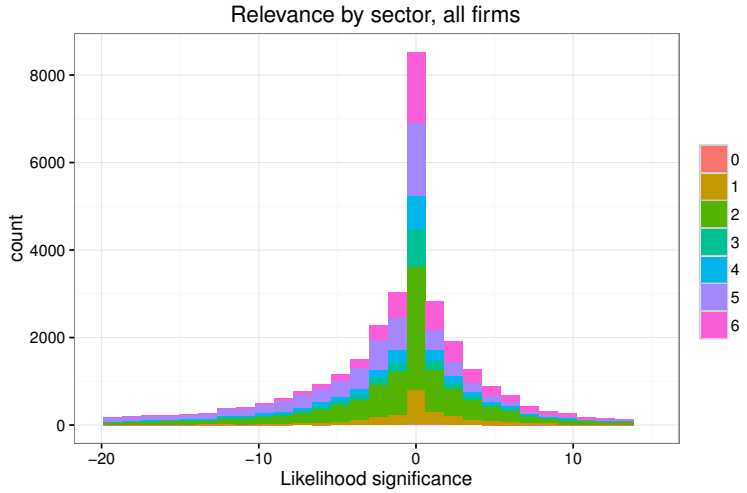

Figure 3.19: The histogram takes into account all the firms in our dataset in relation to how significant they are in term of the total likelihood. The division that is made is respect to the initial sector to which firms belong.

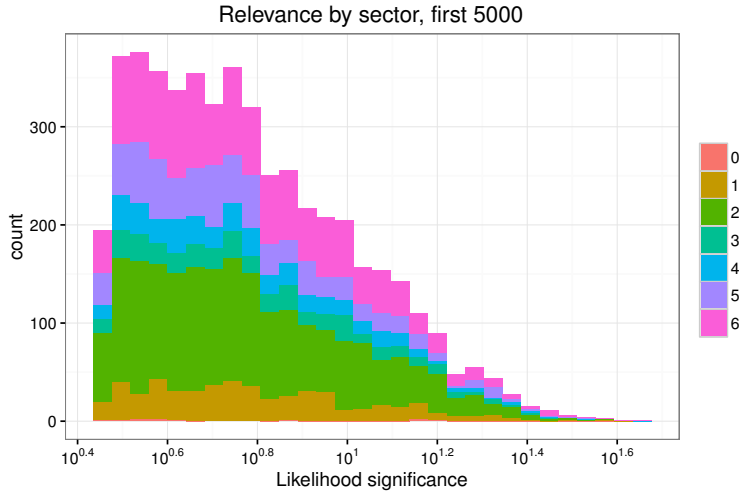

Figure 3.20: The histogram takes into account just the most significant 5000 firms in term of their likelihood. The division is again made depending on the initial sectors.

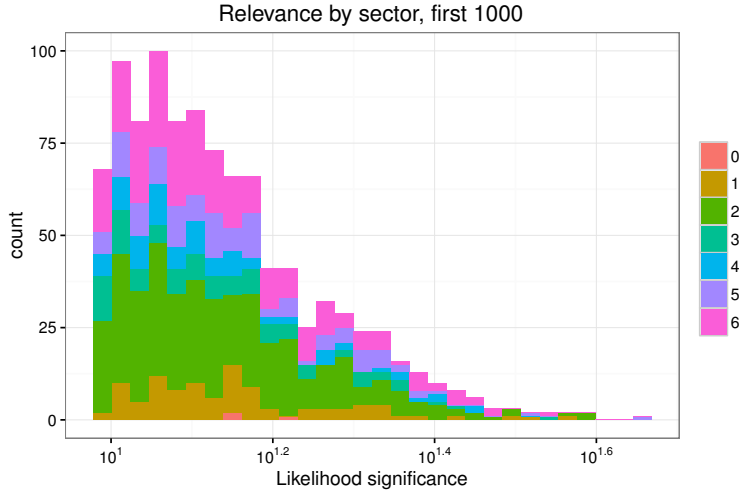

Figure 3.21: The histogram takes into account just the most significant 1000 firms in term of their likelihood.The division is again made depending on the initial sectors.

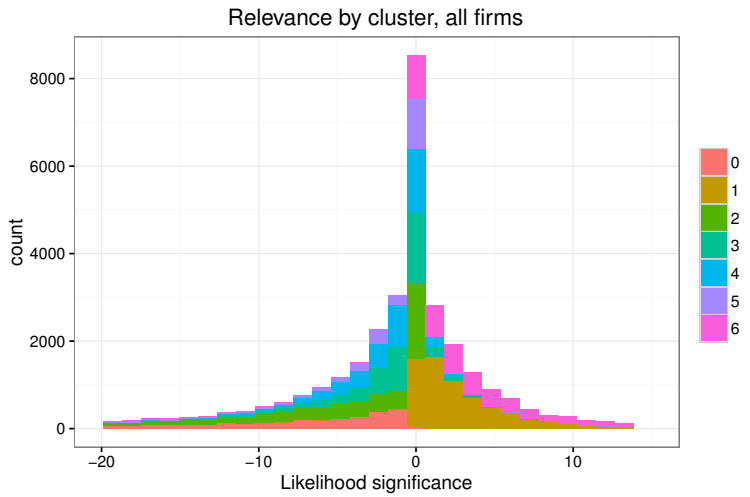

Figure 3.22: The histogram takes into account all the firms in our dataset in relation to how significant they are in term of the total likelihood. The division that is made is respect to the final clusters to which firms belong.

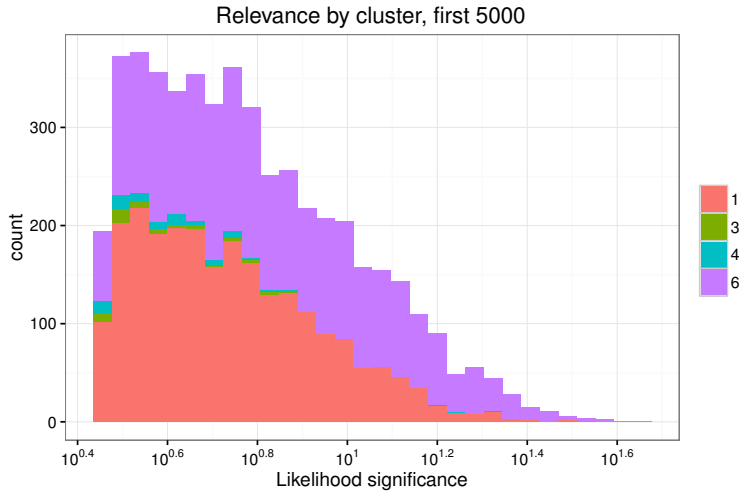

Figure 3.23: The histogram takes into account just the most significant 5000 firms in term of their likelihood. The division that is made is respect to the final clusters to which firms belong.

Figure 3.24: The histogram takes into account just the most significant 1000 firms in term of their likelihood.The division that is made is respect to the final clusters to which firms belong.

3.6 Extracting prototypes

It would be really interesting if we were able to identify the prototypical firms living in each cluster. A prototypical firm must be representative of its cluster in some sense. Proceeding on our previous resolution of looking at the most likelihood-significant firms we end up identifying such individuals for each of our cluster. Therefore, in the graphs that follow, on the right of the graphs we can read the numbers, in the Compustat classification, which correspond to the time series.

The result are presented in the following. For our scopes, we'll take the first firms in each cluster respect to their likelihood significance as the "most prototypical" ones. We expect good results for clusters 1 and 6 since in this case the likelihood is positive and firms should be correlated inside these two, while on the opposite we don't expect very good results for firms inside cluster 0 which appears to be very little significant.

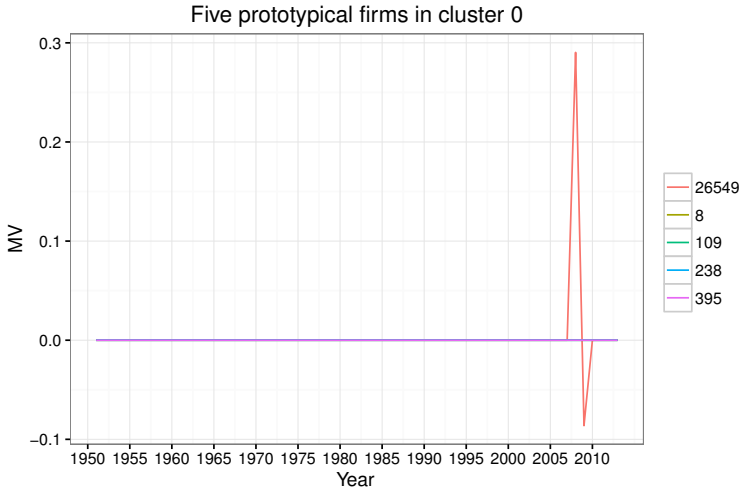

Figure 3.25: The graph presents the trends of the Market Value for the five most significant firms in cluster 0. As expected 4 out of five have likelihood significance 0.

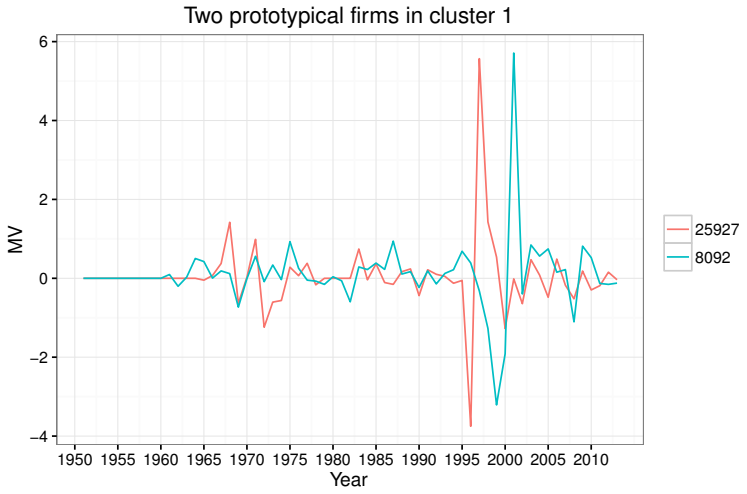

Figure 3.26: The graph presents the trends of the Market Value for the two most significant firms in cluster 1. As expected, we notice a good correlation of the two time series with maybe some outlayers between the years 1995 and 2005

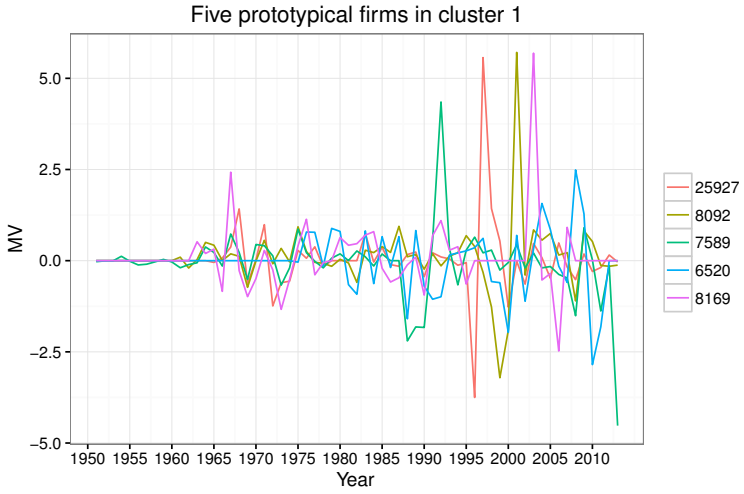

Figure 3.27: The graph presents the trends of the Market Value for the five most significant firms in cluster 1. Looking at five firms the correlation is even more evident and highlights the goodness of our fitting model.

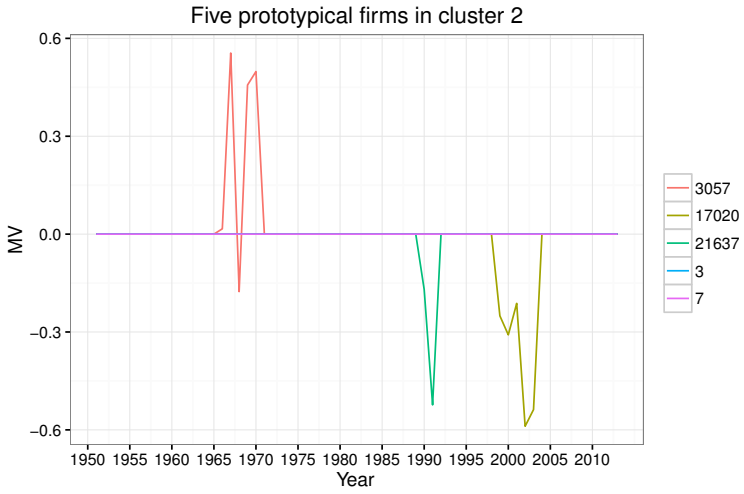

Figure 3.28: The graph presents the trends of the Market Value for the five most significant firms in cluster 2. The graph shows again a pathological case in which the firms in the cluster have a likelihood significance that is zero (flat time series) or negative.

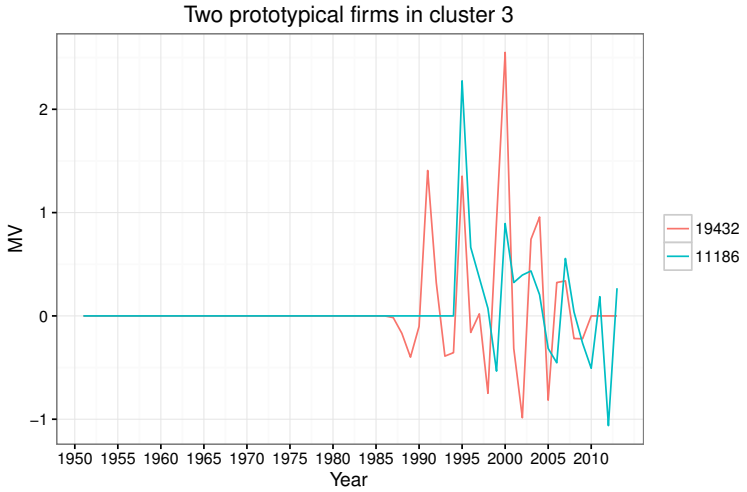

Figure 3.29: The graph presents the trends of the Market Value for the two most significant firms in cluster 3. Again, we notice some correlation as expected from the previous analysis even though it is not as marked as for cluster 1.

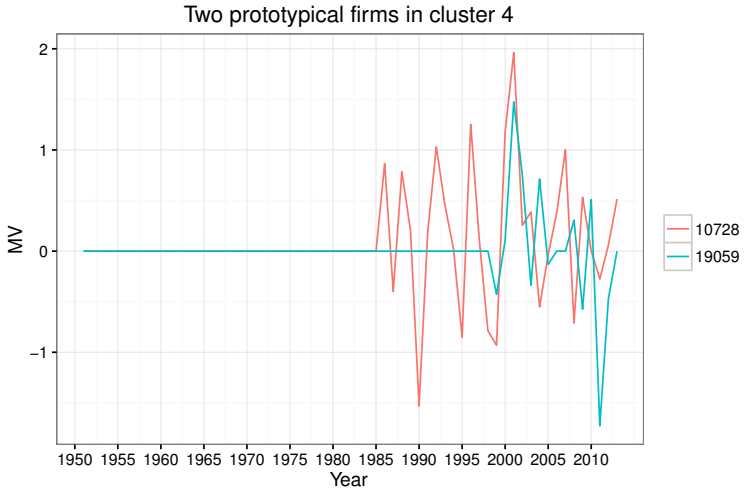

Figure 3.30: The graph presents the trends of the Market Value for the two most significant firms in cluster 4. Still, there are some traces of correlation.

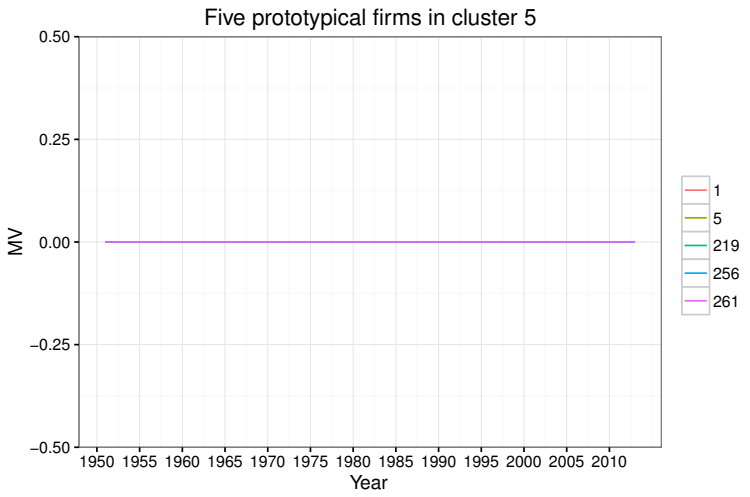

Figure 3.31: The graph presents the trends of the Market Value for the five most significant firms in cluster 5. The five time series are all equal to zero meaning a zero likelihood significance. Cluster five is another pathological cluster.

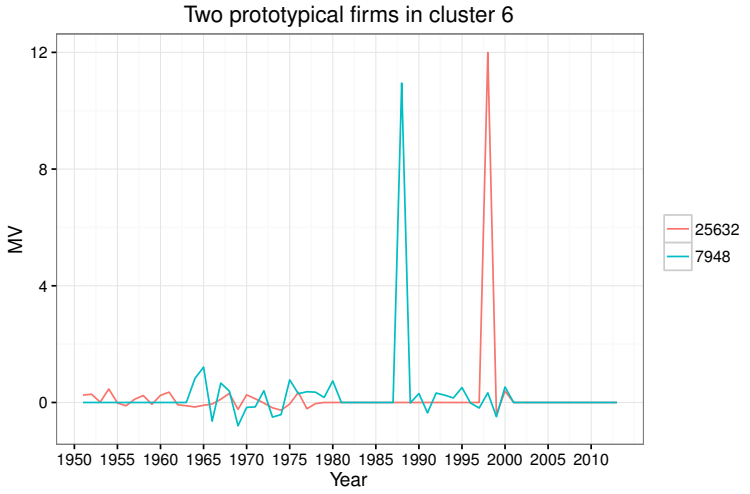

Figure 3.32: The graph presents the trends of the Market Value for the two most significant firms in cluster 6. We notice a good correlation as expected.

Chapter 4

Conclusions and future outlook

From what we have seen, our model for firms' evolution seems to be able to capture the evolution of firms in the global economy even though the classification algorithm can be improved in future works. Furthermore, the model can easily be adjusted to incorporate more than one time series for each firm. One idea is to use time series concerning both the interaction of the firms through the real economy, using for example the sales, and their interaction through the financial sector, using the market value like we have done.

Anyway, our analysis indicates that the model is good for a part of the firms but some of them are not fitted very well by the algorithm. This is what section 3.5 shows. Some firms seem to be forced to stay in the same cluster by the algorithm and not by a real resemblance with the other time series. This is a signal that the dataset we are using is noisy and some time series should be weighted less than other to improve the analysis.

There are many ways to confront the problem of separating information from pure noise. First of all, an idea for future analysis is to increase the number of the clusters in order to provide a greater flexibility in the classification. Also in this case the stability of the algorithm should be checked.

Another possibility, that deserves more investigation, is to introduce another "fictitious" cluster. What we mean is that firms should be assigned to this cluster if the likelihood significance increases by eliminating the corresponding time series from the dataset. Firms inside this cluster are like absent from the dataset but in the reassignment step can be reassigned to one of the other "normal" clusters if this leads to an increase in the likelihood. Our idea is to separate in such a way the noisy time-series from the analysis.

Bibliography

[1] Davide Fiaschi et al. "The interrupted power law and the size of shadow banking". In: *PloS one* 9.4 (2014), e94237.

[2] Lorenzo Giada and Matteo Marsili. "Algorithms of maximum likelihood data clustering with applications". In: *Physica A: Statistical Mechanics and its Applications* 315.3 (2002), pp. 650–664.

[3] Lorenzo Giada and Matteo Marsili. "Data clustering and noise undressing of correlation matrices". In: *Physical Review E* 63.6 (2001), p. 061101.

[4] JA Hartigan. *Clustering Algoritms, J*. 1975.

[5] John A Hartigan and Manchek A Wong. "Algorithm AS 136: A k-means clustering algorithm". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 100–108.

[6] Steffen L Lauritzen. "The EM algorithm for graphical association models with missing data". In: *Computational Statistics & Data Analysis* 19.2 (1995), pp. 191–201.

[7] Simon Levin. *Fragile dominion*. Basic Books, 2007.

[8] Martin Reeves, Simon Levin, and Daichi Ueda. "The Biology of Corporate Survival". In: *Harvard Business Review* 94.1 (2016), pp. 46–55.