ON CLUSTERS FORMATION IN HIGH TECH STOCK MARKET

M. Bucolo[1], L. Fortuna[1], L. Galvagno[1], G. Tomarchio[1,2]

1 DIEES Universita'degli studi di Catania, e-mail mbucolo@diees.unict.it
2 STMicroelectronics Catania Site, e-mail giusy.tomarchio@st.com

Abstract: Stock Markets, like many other natural systems, are complex systems in a state of unstable equilibrium where future evolution depends on many parameters. Their evolution is characterized by different factors: political and social events, government policies, humour and strategy of the consumer environment. Even if attention has been focused on this system since the early years of the nineteen-century and many efforts are still concentrated on it, no valid and recognized model has been accepted by the scientific community. This paper aims at being a scientific contribution which tries to formalize the main features in the stock market topology. Cluster formation has been investigated in different time windows to identify specific properties. The analysis has been concluded in this case to highlight clusters in industrial sectors linked to the semiconductor industry and it has been focused on a big set of high-tech sectors quoted in the New York stock market.

Keywords: Financial Systems, Time series Classification, Neural Networks

1. INTRODUCTION

Financial and Economic Systems have catalysed an ever-increasing scientific attention. Evolutions of these systems have a direct impact on the human daily life and a new branch of engineering defined as financial engineering has emerged.

Particularly, in the last years market globalisation and high interconnectivity all over the world have required more scientific attention and efforts to find reliable models to monitor and to control financial systems; in fact no valid and recognized model has been accepted by the scientific community.

The structure of the system is close to the structure of generic complex systems constitued of a network of interacting units together with all the elements that could have an impact on the economic system like political and social events, government policies, and performance of quoted entities, humour and strategy (Neil F. Johnson, et al., 2003).

The power law analysis which considers as concerned variable the value returns has been already adopted in literature [] and moreover the research of topology structures in financial market (Mantegna, et al., 2000) is an appealing field of study in relation to the different possible applications; nevertheless these properties have not been explored deeply and exhaustively.

This work is focused on the investigation of topology evolution and it aims at being a contribution to this field of research. The considered case study consists of the stock quotations belonging to industrial firms in the high technology market. Sectors have been selected according to the relationship they have with the semiconductor market.

The data set has been analysed through a clustering strategy in different time windows in order to track how clusters evolve and if a certain pattern is present. To overcome the computational complexity of the clustering phase, GHSOM have been implemented in order to have a robust algorithm with a low level of supervision. The obtained results are definitely satisfactory and offer the possibility to trace new ideas for future studies.

The second Section describes the principles applied to select the data set, the data set itself and the pre-processing procedure. The third Section reports the adopted algorithm and the clustering strategy; the results will be commented in the fourth Section, and finally the main conclusion are dealt in the fifth Section.

2. DATA SET

The purpose of this study is to investigate cluster formation, and as a first step the evolution of a set of different industrial sectors strictly related has been addressed. High technology industries have been chosen since they represent undoubtedly the protagonists in the economic world.

Particularly the data set has been selected focusing on a set of high tech industries quoted at the New York stock exchange; the link with the semiconductor industry has been the main discriminator, since all the sectors are directly linked to the semiconductor industry. Selected sectors are reported in Table 1 together with the number of companies: the classification of the industrial sectors is the one adopted in the Yahoo Financial Web Portal.

The data set is built with a total of 153 companies in the period from 29.11.1996 to 23.02.2004; samples have been collected on a daily basis for a total of 1818 points.

Table 1 Industrial Sectors that constitutes the data set and number of societies belonging to

Sector	#Societies
Computer – Hardware	20
Computer storage device	11
Computer peripherals	19
Audio-Video	6
Communication Equipment	49
Semiconductor	48

The variables collected are:

- Open value (**ap**)
- Close value (**ch**)
- Volumes (**vl**)

Particularly the analysis has been focused on the open value. The data set has been pre-processed through a normalization phase. The time series have been normalized between [0, 1] in order to classify them by trend.

3. CLUSTERING ANALYSIS

The dimension of the data set is one of the problems that arise in the clustering analysis; the complete set is made up of 153 time series with 1818 samples: the processing algorithm and the visualization of the results are the main issues to be solved. The first point has found an optimal solution in GHSOM that stands for Growing Hierarchical Self Organizing Maps that is well suited for unsupervised clustering of big data sets.

3.1 Growing Hierarchical Self Organizing Maps

The GHSOM algorithm is based upon a growing structure of Self-Organizing Maps (SOM) that grows the number of units increases until a quality criterion of the classification has not been reached (Dittenbach, et al., 2000). The growing process builds a hierarchical structure composed by SOM in which lower layers classify data with more granularity, as shown in Fig. 1.
The first layer of the hierarchy is constituted by only one SOM meanwhile in the next layer a new SOM can be connected to any unit of the first map. This link is repeated with each map of each layer versus their previous layers, until a performance target is reached..

3.1 Self Organizing Map

The fundamental element of a GHSOM structure is the Self-Organizing Map (SOM) that is a structure composed of mathematical units, called neurons that are organized in a grid and connected to inputs. The number of neurons may vary from a few dozens to thousands of elements. Each nonlinear unit is represented by a weight vector, or prototype vector, $m = [m_1 \ldots m_n]$, where n is the length of the input vectors. All these units are connected to their neighbours through neighbourhood relations that state the topological structure of the map, which could be rectangular or hexagonal.

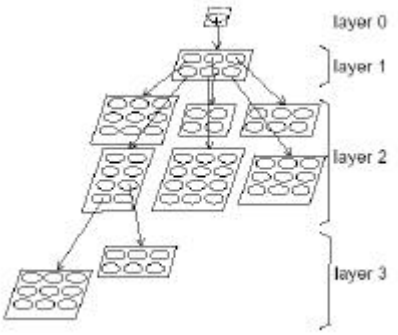

Fig. 1- Hierarchical structure of a GHSOM.

By using the SOM two phases can be distinguished, a *learning phase* in which the SOM is adapted to the input vectors, and a *recognizing phase*.

3.2 GHSOM Algorithm

The learning algorithm of a GHOSM structure starts with a *layer* 0 constituted by an SOM with only one neuron. The weight vector of this neuron is initialised at the mean of all vectors of the classifying data set. The learning process goes on with an SOM map in *layer* 1 of 2x2 neurons that organizes itself through the SOM learning algorithm, by associating the input vectors to the different map regions.
This training process is repeated for a number *n* of iterations that represents the number of the input vectors. After n iterations the neuron with the highest deviation between its weight vector and the input vectors related to this region, is chosen as an *error unit*. A new row or a new column of neurons is inserted between the *error unit* and its most dissimilar neighbour, as shown in Fig. 2. The weight vectors of these new units are assumed to be equal to the neighbours'mean.

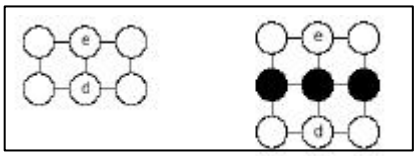

Fig.2- The insertion of a new row between error unit (e) and its most dissimilar neighbour (d).

An error functional guides the growing of the map together with the setting of two parameters. The breath and depth respectively drive the growing in horizontal and vertical dimension.
An error function guides the growing of the map together with the setting of two parameters. The breadth and depth of the map respectively drive the growth of the map in a horizontal and a vertical dimension.
The initial GHSOM architecture is formed by only one SOM, this architecture grows with another layer because dissimilar input data can not be mapped to the same neurons. These neurons are characterized by a high quantization error that is higher than a particular threshold value. This threshold basically indicates the depth level of the hierarchical structure,

which represents the data, and is a fraction of the quantization error in *0*.
GHSOM Maps have been tested in a first phase in order to evaluate the best values of t_1 and q_i allowing a good clustering granularity.. Several tests have been carried out by visual inspection; the best results have been obtained for values of 0.8 and 1. The parameters value of breadth and depth indicate that the growing of the map has been addressed only in the horizontal dimension.

3.3 Clustering Strategy

Time series have been clustered by dividing the whole observation horizon in time frames of different lengths, the aim is to investigate the different dynamics and the creation of different topologies; particularly in this work, time frames of one year, six months and three months have been considered.
Classes have been found by grouping time series with the same trend and, the results, reported in Figs. 3-4, show how this analysis of the clusters in different time frames could highlight different company associations. The classes reported have the same trend in the first part, while become different in the second part of their evolution

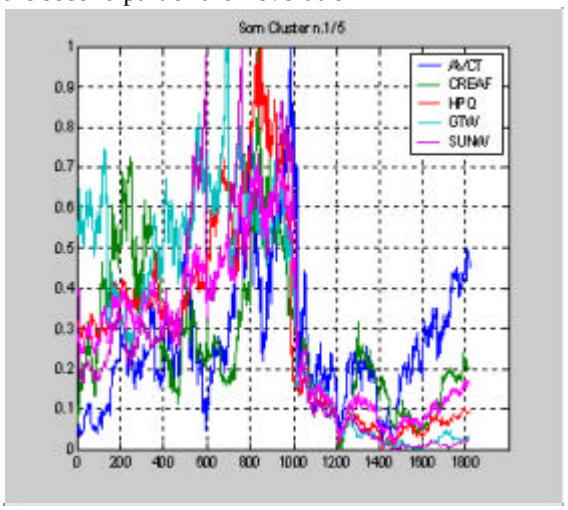

Fig.3. One of the cluster obtained of 5 elements.

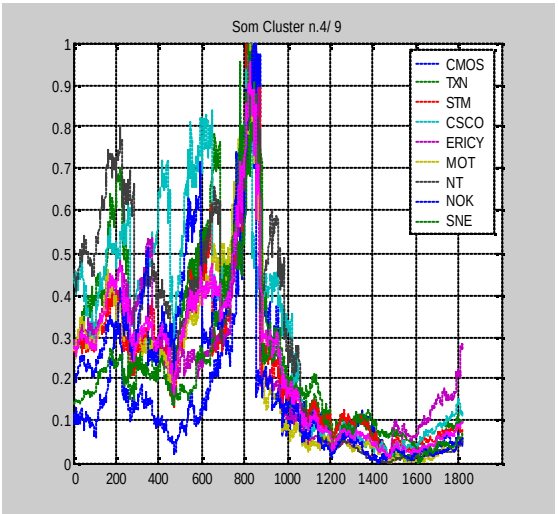

Fig.4. One of the clusters obtained of 9 elements.

4. RESULTS

The clustering analysis on time frames has made it possible to verify how time series are grouped in a set of classes that change the number of elements by varying the time frame.
In Fig.5 the trend of the number of classes in the yearly buckets has been reported. The analysis, therefore, reports 7 points related to each year from 1997 to 2003.
It is possible to highlight how the number of classes swings in the range [12, 15] between 1997 and 2001 while after 2001 the number reaches the minimum, 9 classes, and the maximum, 24, of the distribution.

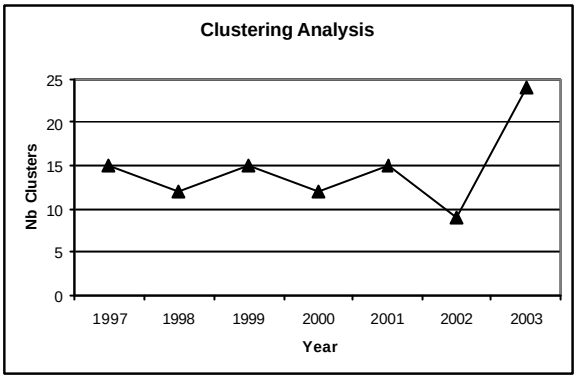

Fig.5. Number of clusters for time frame of one year

This oscillatory behaviour is clearly visible also in the clustering analysis over the six months time frame reported in Fig. 6. The minimum value of 9 classes is reached in the first half of 2002, coherently with the yearly analysis, while the maximum value of 24 classes is never reached here.

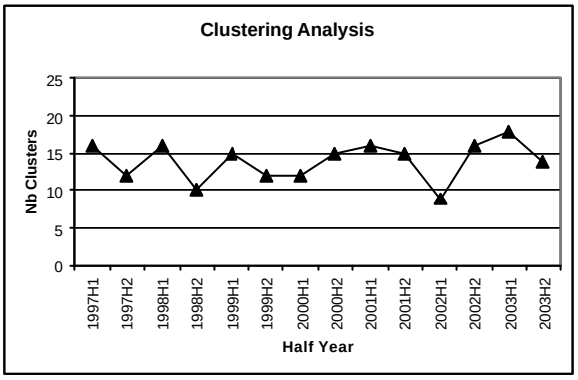

Fig.6. Number of clusters for time frame of six months

The clustering analysis by quarterly time frames, Fig. 7, shows similar characteristics: the number of classes is always limited between [9, 20] and it has a swing trend; a minimum value of 9 classes is reached in the third quarter of 2001.

5 CONCLUSIONS

Financial Systems have been catalysing a growing scientific interest since the early years of the nineteenth century; despite the great interest and efforts of the scientific community no universal

model has been accepted. Particular features like the Power Law Behaviour and the existence of topology are well known characteristics of this system but no exhaustive investigation has been performed on the structural nature of the market even if actually the analogy with other complex systems based on networks has been proposed by the scientific community.

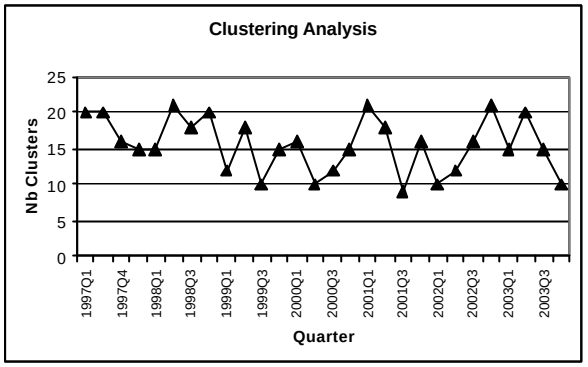

Fig.7. Number of clusters for time frame of three months

In this paper the evolution of Topology has been explored focusing on a set of high-technology industrial sectors particularly linked to the semiconductor market. A data set constituted by 153 time series has been processed to find clusters. To face the computational complexity, cluster processing has been performed through GHSOM that give the right trade off between speed, low level of supervision and clustering granularity. The time window has been divided in time frames of one year, six months and 3 months to look for the evolution of the clusters and resemblance in time scale.

The results have made quantitative and qualitative evidences possible: a repetitive behaviour characterizes the variation in each time frame.

The number of classes remains in any case limited between a minimum and a maximum in the range of [10, 25] and a behaviour close to time invariance has been highlighted even therefore there was no possibility of investigating the data deeply since samples were daily.

The minimum value of the clusters has been found around the third quarter of 2001 and the first quarter of 2002; in that period the world went through one of the most impressive terrorist attacks.

Results lead to the assumption that more aspects have to be investigated and often possibilities have to be made available to enhance the study by looking at a "master" time series or at the synchronization effects related to exceptional events that could find explanations in the collective behaviour of traders.

REFERENCES

Neil F. Johnson, Paul Jeffries, Pak Ming Hui, Paul Jefferies, 'Financial Market Complexity (Economics & Finance)' Oxford University Press Sep 2003

Rosario N. Mantegna, H. Eugene Stanley, 'An Introduction to Econophysics: Correlations and Complexity in Finance' Cambridge University Press 2000

Michael Dittenbach, Dieter Merkl, Andreas Rauber, 'The Growing Hierarchical Self-Organizing Map', *'Proc. of the International Joint Conference on Neural Networks'*, pp. 15 - 19, Lug. 2000.

Michael Dittenbach, Dieter Merkl, Andreas Rauber, 'Organizing and Exploring High-Dimensional Data with the Growing Hierarchical Self-Organizing Map', *'Proc. of the 1st International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2002)'*, pp. 626-630, Nov. 2002.