A CLUSTERING OF DJA STOCKS - THE APPLICATION IN FINANCE OF A METHOD FIRST USED IN GENE TRAJECTORY STUDY

Moldovan Darie
Babeş-Bolyai University of Cluj-Napoca Business Information Systems Department Str. Theodor Mihali 58-60, 400599, Cluj-Napoca, Romania {Darie.Moldovan, Gheorghe.Silaghi}@econ.ubbcluj.ro
Silaghi Gheorghe Cosmin
Babeş-Bolyai University of Cluj-Napoca Business Information Systems Department Str. Theodor Mihali 58-60, 400599, Cluj-Napoca, Romania {Darie.Moldovan, Gheorghe.Silaghi}@econ.ubbcluj.ro

Previously we employed the Gene Trajectory Clustering methodology to search for different associations of the stocks composing the DJA[612] index [4] with the aim of finding different, logic clusters, supported by economic reasons, preferably different than the classic, "by industry" classification. In this paper we enter the insights of the clustering results from a financial and business perspective, to see if the clustering results are validated by the market knowledge and history.

Keywords: clustering model, data trajectory, cluster analysis

JEL Classification: G10, C61

1. Introduction

The behavior in time of a single stock can't describe the evolution of the entire market, but studied alongside with other ones, weighting their importance, one can tell the main direction of the group. For this reason stock indexes were created. In reverse, it is easier to forecast the price evolution of a single stock, taken away from a group where most of the stocks have a similar behavior.

A stock market index is a method for measuring a section of the market. In the last few decades, indexing has been a strong preoccupation for every fund manager, raising the performance expectations [1]. Created by financial services companies or news providers, the indexes are the first benchmark for the performance of a portfolio.

There are many types of indexes, based on the size, specific sector, type of management or other criteria considered useful by their creators. Indexes are usually built by financial experts or by investment companies and their structure is more or less subjective. The literature consists of several attempts [6, 7 ,8] to automatically obtain the structure of a stock market objectively, without human intervention, only from historical data. In line with this trend, we employed an artificial intelligence approach [4] to obtain groups of stocks, considering only their price evolution during the same period of time. In this paper our scope is to further investigate the results obtained in [4], to see if the stock groups and the associations determined by the clustering methodology are validated by the financial and business knowledge present at that time in the market. We worked with the the 65 companies (traded on the New York Stock Exchange and NASDAQ) composing the DJA index, analyzed between years 2000 and 2007.

The paper is structured as follows: in section 2 we describe the gene trajectory clustering methodology applied in [4] for clustering the DJA index stocks. Section 3 presents and explains the clustering results by an economic point of view and section 4 concludes the paper.

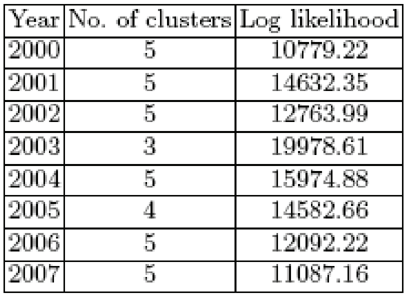
2. Gene Trajectory Clustering for structuring the DJA index components

Gene Trajectory Clustering [2, 3] is a method implemented into GNetXP software and developed to extract the Gene Regulatory Network from gene trajectory data The proficiency of the hybrid algorithm (using a mixture of Multiple Linear Regression models) in clustering was first demonstrated by tests on time series containing hundreds to thousands records. The methodology consists in two steps in clusters determination: first, local centers for the clusters are determined with the help of a Genetic Algorithm and second, by using a local-learning method to refine the initial centers selected. The likelihoods of the solution are then used as objective function for the Genetic Algorithm. Even this approach is time consuming, it relies on temporal information between data and the results are considerably improved, compared to the standard Expectation Maximization algorithm and it is unlikely to be trapped in the local optima.

In [4] we successfully applied the Gene Trajectory Clustering method on financial data. The data analyzed consisted of the daily adjusted closing prices of the DJA stocks for the period 2000-2007. The data was divided into 8 natural periods for detailed analysis. As recommended by the financial investments literature [5], we calculated the daily logarithmic returns. Moreover, for the data to fit the rigors of the GNetXP software, and to obtain a global vision on the price evolutions of each stock, we needed to scale the trajectory of each stock, considering a start of 100 points and applying the daily logarithmic returns computed at the previous step. In this way we obtained a dataset, in appearance very resembling to the gene data originally tested.

[612] Dow Jones Composite Average

For every year, we obtained a clustering of the stocks, clusters that acknowledge more or less the division of DJA in the 3 sectors: industrial, services and utilities. We found out that running the clustering methodology for 8 successive years, the clusters remains quite stable, with a kernel of stock classified in the same cluster during all years. This sign was a first theoretical indication that the clustering methodology was successful from the algorithmic point of view and some useful insights might be obtained out of the clustering procedure. Table 1 presents the number of clusters obtained for each successive year.



Year	No. of clusters	Log likelihood
2000	5	10779.22
2001	5	14632.35
2002	5	12763.99
2003	3	19978.61
2004	5	15974.88
2005	4	14582.66
2006	5	12092.22
2007	5	11087.16

Table 1. Number of clusters obtained for each year using the hybrid GTC algorithm [4]

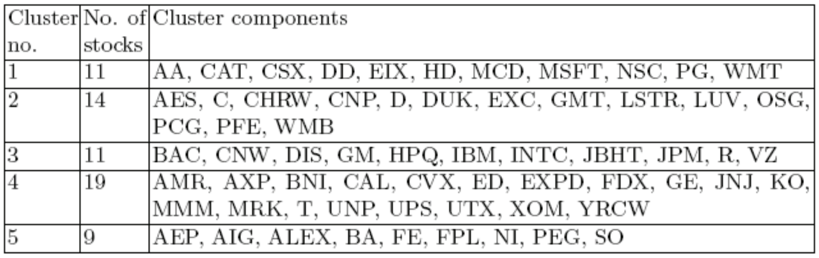
3. Results

Since approximations were used in cluster formation, a refined analysis is welcomed. After obtaining the clusters for every data set, we measured the intra-cluster distance between its components. This is important, for being able to observe the correlation between stocks in a cluster, especially useful when finding unnatural associations, apparently hard to correlate. If a small distance is shown, one can be sure the stock is not misclassified, and will look for a logic explanation. The method proposed for this computation is the measure of the Euclidean distance (L2-norm), proposed also by [6],[7]. For each year we drew a dendrogram showing the internal adhesion inside clusters. An example of such dendrogram is shown in Figure 2, based on year 2000 clusters: on the (X) axis we have the cluster components while on the (Y) axis are shown the Euclidean distances between stocks.

Next, we present the economic cluster analysis, describing the most interesting correlations found between components for the above mentioned period of time.

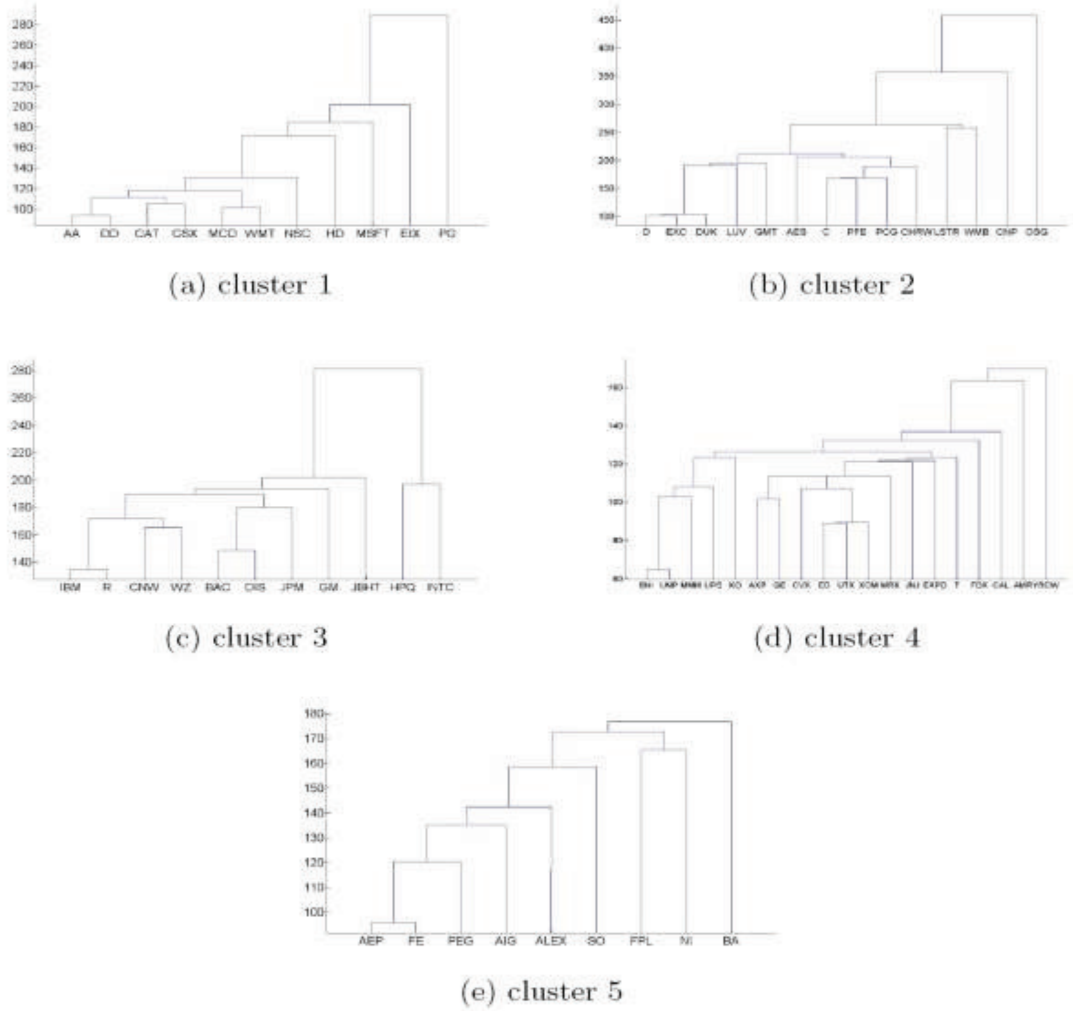
Table 2 presents the clustering results for year 2000. Year 2000 was the end of the largest economic boom in US history. Stocks were volatile, with Centerpoint Energy (Utilities) doubling it’s market value, but Microsoft (Technology) dropping 55%. The first cluster obtained contains the stocks(11) who dropped most during the analyzed period, a strong correlation being shown between Alcoa and Du Pont(both from Basic Materials industry), McDonalds and Wallmart(both from Services sector) and CSX(Services-Railroads) and Caterpillar (Industrial goods). The only IT representative of this cluster was Microsoft.

Cluster number two incorporates the companies (14) with a volatile evolution, but all of them rose during the whole year. Most closed were Dominion Resources, Duke Energy, Exelon (Utilities), GATX (Services-Rental) and Landstar System (Services-Trucking). An interesting association is created between Citigroup (the only representative of the financial sector in the cluster) and Pfizer (the only Healthcare stock in the cluster)



Cluster no.	No. of stocks	Cluster components
1	11	AA, CAT, CSX, DD, EIX, HD, MCD, MSFT, NSC, PG, WMT
2	14	AES, C, CHRW, CNP, D, DUK, EXC, GMT, LSTR, LUV, OSG, PCG, PFE, WMB
3	11	BAC, CNW, DIS, GM, HPQ, IBM, INTC, JBHT, JPM, R, VZ
4	19	AMR, AXP, BNI, CAL, CVX, ED, EXPD, FDX, GE, JNJ, KO, MMM, MRK, T, UNP, UPS, UTX, XOM, YRCW
5	9	AEP, AIG, ALEX, BA, FE, FPL, NI, PEG, SO

Table 4. Year 2000 cluster components



(a) cluster 1

(b) cluster 2

(c) cluster 3

(d) cluster 4

(e) cluster 5

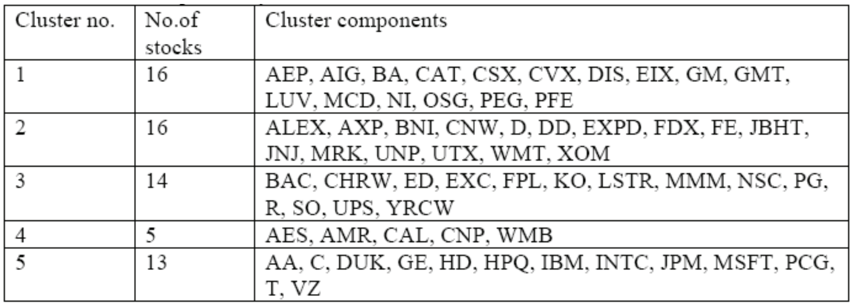
Table 3 presents the results for year 2001, dominated by main event of the terrorist attacks over World Trade Center and Pentagon, which caused a major drop of the indexes in September. Also, Microsoft was accused of the violation of antitrust laws.The first cluster contains 20 stocks who had a descending trend during the year and decent rebounds. We can find alongside AIG, Boeing and other five utilities companies, proving the association between them in the previous year was not by mistake. A closer evolution had the energy companies Chevron, American Electric Power, Edison, First Energy and, more surprisingly with Johnson&Johnson(Healthcare). Caterpillar and General Motors evolve together in this cluster, too. Companies (16) in the fourth cluster, even they rose in the last quarter of the year, didn't have the chance of a positive year-end close. As in the previous situations, the utilities companies are among those who perform closely. The diversified technology company 3M is in this cluster, together with Wal-Mart, P&G, Pfizer and Home Depot.

Cluster no.	No.of stocks	Cluster components
1	20	AIG, AXP, BA, CNP, CNW, DIS, EXC, FPL, GMT, HPQ, INTC, JPM, KO, LUV, MCD, MRK, NI, T, UTX, WMB
2	14	AA, AEP, CAT, CVX, ED, EXPD, FDX, FE, GM, JBHT, JNJ, UNP, VZ, YRCW
3	5	AES, AMR, CAL, EIX, PCG
4	16	ALEX, BNI, C, CHRW, D, DD, DUK, GE, HD, MMM, PEG, PFE, PG, UPS, WMT, XOM
5	9	BAC, CSX, IBM, LSTR, MSFT, NSC, OSG, R, SO

Table 5. Year 2001 cluster components

Figure 8. Dendrograms for the clusters, showing the adhesion between the individuals

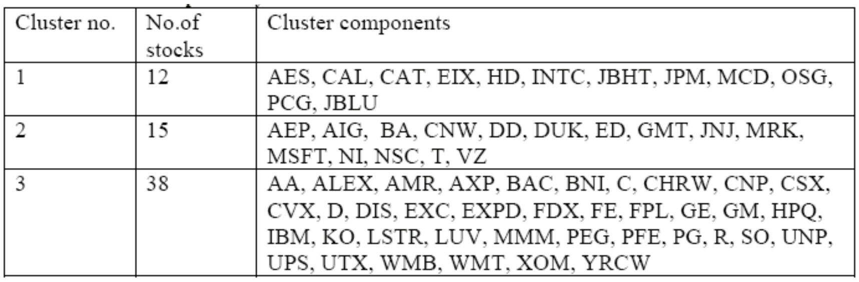
At the beginning of 2002 (results presented in table 4) the Justice Department in US launched it's investigation of Enron and WorldCom filled for bankruptcy. The market dropped and didn't recover till the end of the year. The stocks in the first cluster (16) started to drop in the second quarter and at the end of the year the losses were in the [-12%,-38%] interval. Strongly correlated were AIG and Pfizer. The car manufacturers GM and Caterpillar are in the same cluster, along with Boeing, but also energy and shipping companies. The fourth cluster belongs to the companies who spectacularly dropped, from 60% to 90%: AES Corporation (Utilities), AMR Airlines, Continental Airlines, Centerpoint Energy and Williams (Oil and Gas).



Cluster no.	No.of stocks	Cluster components
1	16	AEP, AIG, BA, CAT, CSX, CVX, DIS, EIX, GM, GMT, LUV, MCD, NI, OSG, PEG, PFE
2	16	ALEX, AXP, BNI, CNW, D, DD, EXPD, FDX, FE, JBHT, JNJ, MRK, UNP, UTX, WMT, XOM
3	14	BAC, CHRW, ED, EXC, FPL, KO, LSTR, MMM, NSC, PG, R, SO, UPS, YRCW
4	5	AES, AMR, CAL, CNP, WMB
5	13	AA, C, DUK, GE, HD, HPQ, IBM, INTC, JPM, MSFT, PCG, T, VZ

Table 6. Year 2002 cluster components

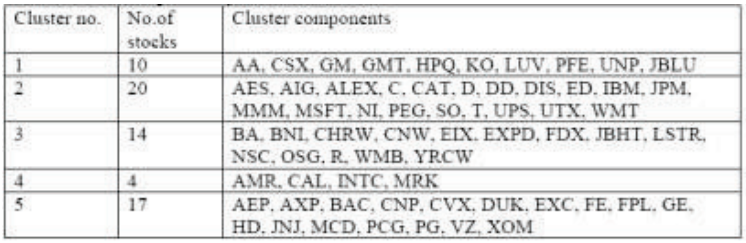
Our algorithm fit best for three clusters in 2003 (Table 5). First cluster incorporates 12 stocks, most of them on a strong ascending trend, who began in the second quarter. Best correlation was achieved between Caterpillar and Edison and PG&E(utilities) and JB Hunt Transport. We can also find McDonalds alongside JP Morgan, Intel and Home Depot. Second cluster's stocks dropped at the beginning of the year, and after recovering evolved almost flat for the rest of the year. The year end performances are at best +20% and at worst -20%. Good correlations were found between Du Pont and AIG. Merck and J&J, the drug manufacturers are in this cluster, but Microsoft and Verizon too.



Cluster no.	No.of stocks	Cluster components
1	12	AES, CAL, CAT, EIX, HD, INTC, JBHT, JPM, MCD, OSG, PCG, JBLU
2	15	AEP, AIG, BA, CNW, DD, DUK, ED, GMT, JNJ, MRK, MSFT, NI, NSC, T, VZ
3	38	AA, ALEX, AMR, AXP, BAC, BNI, C, CHRW, CNP, CSX, CVX, D, DIS, EXC, EXPD, FDX, FE, FPL, GE, GM, HPQ, IBM, KO, LSTR, LUV, MMM, PEG, PFE, PG, R, SO, UNP, UPS, UTX, WMB, WMT, XOM, YRCW

Table 7. Year 2003 cluster components

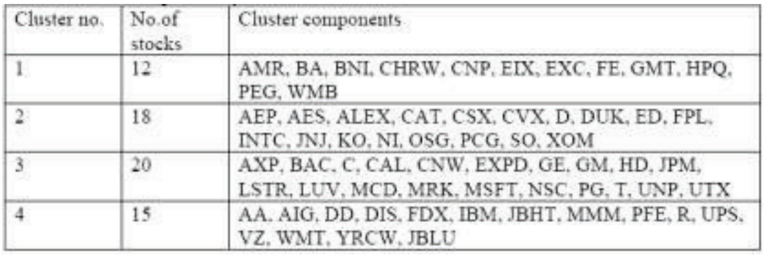
First cluster of 2004(Table 6) includes stocks that were in the negative zone for the most of the year. Strong correlation was found between Pfizer and Coca-Cola. General Motors and HP are here, too. In the second cluster, the stocks(20) are not very volatile, evolved flat until the last quarter, when most of them ended the year higher. As expected from previous observations, the utilities companies are among the strongest correlated: Consolidated Edison, Nisource, Dominion Resources, Southern Company, AES, Public Service Enterprise. Technology companies Microsoft, IBM, 3M and AT&T are incorporated in the cluster, too. Citigroup, AIG, JP Morgan are the representatives of the financial sector. The third cluster presents stocks with outstanding performance during the year, with price increases between 20% and 80%. CH Robinson Worldwide, Expeditors International, FedEx, Overseas Shipholding (delivery services) and Lanstar Systems and YRC(trucking) are strongly correlated. Companies in the last cluster rose smoothly, all of them ending the year in the positive zone, with 10-30%. American Expres and Bank of America are correlated, but also Chevron and Exxon Mobil. Procter&Gamble, Johnson&Johnson, Home Depot, McDonalds are also found here.



Cluster no.	No.of stocks	Cluster components
1	10	AA, CSX, GM, GMT, HPQ, KO, LUV, PFE, UNP, JBLU
2	20	AES, AIG, ALEX, C, CAT, D, DD, DIS, ED, IBM, JPM, MMM, MSFT, NI, PEG, SO, T, UPS, UTX, WMT
3	14	BA, BNI, CHRW, CNW, EIX, EXPD, FDX, JBHT, LSTR, NSC, OSG, R, WMB, YRCW
4	4	AMR, CAL, INTC, MRK
5	17	AEP, AXP, BAC, CNP, CVX, DUK, EXC, FE, FPL, GE, HD, JNJ, MCD, PCG, PG, VZ, XOM

Table 8. Year 2004 cluster components

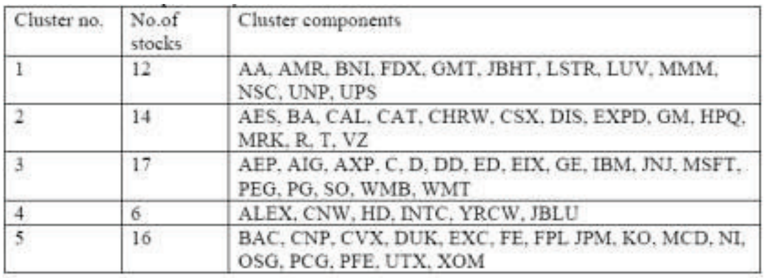
In 2005(Table 7) the US stock markets were dominated by concerns regarding Iraq war, New Orleans floods and rising interest rates, all of these keeping the indices bellow 5%. We found a strong correlation between the financials Bank of America, Citigroup, JP Morgan, General Electric, American Express and Home Depot stores and P&G. The utility companies go together, as well, but in a different cluster, but with a decent performance. Stocks in the fourth cluster poorly performed until the last quarter of year, when a rebound came. Stocks in multiple domains are found here, most correlated being the shipping companies Fedex and JB Hunt Transport, and the chemical company Du Pont and the conglomerate 3M.



Cluster no.	No.of stocks	Cluster components
1	12	AMR, BA, BNI, CHRW, CNP, EIX, EXC, FE, GMT, HPQ, PEG, WMB
2	18	AEP, AES, ALEX, CAT, CSX, CVX, D, DUK, ED, FPL, INTC, JNJ, KO, NI, OSG, PCG, SO, XOM
3	20	AXP, BAC, C, CAL, CNW, EXPD, GE, GM, HD, JPM, LSTR, LUV, MCD, MRK, MSFT, NSC, PG, T, UNP, UTX
4	15	AA, AIG, DD, DIS, FDX, IBM, JBHT, MMM, PFE, R, UPS, VZ, WMT, YRCW, JBLU

Table 9. Year 2005 cluster components

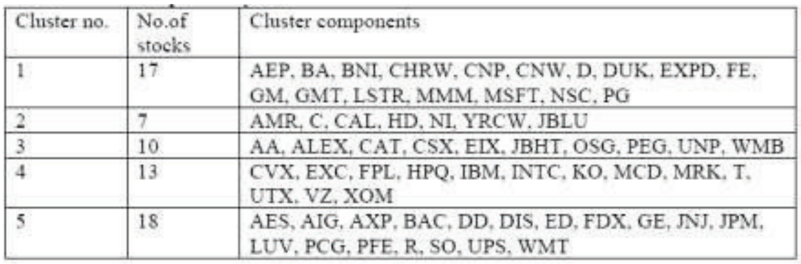
The results of the companies in 2006 restored the confidence in the markets. The benchmark indicator for the US markets, the Dow Jones Industrial Average rose by 16%. The performers of 2006 were the stocks in cluster2 (Table 8), strongly correlated being Boeing, General Motors and Caterpillar and the technology companies HP, Verizon and AT&T. The semiconductor maker Intel was between the stocks with the poorest evolution, alongside with transportation companies ConWay, Jet Blu, Alexander &Baldwin, YRC Worldwide. The situation wasn't better for the companies in cluster 3, who began the year by a small drop, than recovered and ended the year in the positive zone. Here are included, with strong correlation, some major financial stocks (American Expres, Citigroup, AIG), but also the technology companies Microsoft and IBM.



Cluster no.	No.of stocks	Cluster components
1	12	AA, AMR, BNI, FDX, GMT, JBHT, LSTR, LUV, MMM, NSC, UNP, UPS
2	14	AES, BA, CAL, CAT, CHRW, CSX, DIS, EXPD, GM, HPQ, MRK, R, T, VZ
3	17	AEP, AIG, AXP, C, D, DD, ED, EIX, GE, IBM, JNJ, MSFT, PEG, PG, SO, WMB, WMT
4	6	ALEX, CNW, HD, INTC, YRCW, JBLU
5	16	BAC, CNP, CVX, DUK, EXC, FE, FPL JPM, KO, MCD, NI, OSG, PCG, PFE, UTX, XOM

Table 10. Year 2006 cluster components

Even the market ended the year with modest gains, some quarterly losses reported by the banks in the fall were the first signs that the economy is shrinking. The first cluster of the year 2007(see Table 9) contains very volatile stocks, the majority being from energy and transportation domains. The airlines companies from cluster 2 had a difficult year, ending the year in the negative zone, alongside Citigroup, one of the first banks affected by the subprime crisis, who lost almost 50% of its value. Most of the technology companies (cluster 4) rose,, offering a solid performance during the entire year. The energy and transportation companies were again correlated, even they had volatile prices (cluster 1) or a fair positive evolution (cluster 3).



Cluster no.	No.of stocks	Cluster components
1	17	AEP, BA, BNI, CHRW, CNP, CNW, D, DUK, EXPD, FE, GM, GMT, LSTR, MMM, MSFT, NSC, PG
2	7	AMR, C, CAL, HD, NI, YRCW, JBLU
3	10	AA, ALEX, CAT, CSX, EIX, JBHT, OSG, PEG, UNP, WMB
4	13	CVX, EXC, FPL, HPQ, IBM, INTC, KO, MCD, MRK, T, UTX, VZ, XOM
5	18	AES, AIG, AXP, BAC, DD, DIS, ED, FDX, GE, JNJ, JPM, LUV, PCG, PFE, R, SO, UPS, WMT

Table 11. Year 2007 cluster components

4. Conclusion

Our aim in this paper was to investigate whether the results obtained by applying the Gene Trajectory Clustering methodology to cluster financial data are worth from the financial and the business perspective. More specifically, we wanted to determine whether the clusters obtained have logical and economic importance, besides the mathematical values of the performance indicators and to determine if an alternative grouping of the stocks is welcomed.

Considering the business cluster analysis, we conclude that the GTC algorithm applied in [4] was appropriate for clustering the financial data and that there are many cases when the natural division of the stocks by the company profile is not a solution for grouping them, finding the technology companies uncorrelated with each other, and the banks correlated only in the last three years.

Refrences

1. Barry B. Burr "Essential book of indexing. Pensions & Investments", New York. 2005/01/10

2. Chan, Z.S.H., Kasabov, N.K.: Gene trajectory clustering with a hybrid genetic algorithm and expectation maximization method. In: Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. Volume 3., IEEE Computer Society (2004) 1669-1674

3. Chan, Z.S.H., Kasabov, N.K., Collins, L.: A hybrid genetic algorithm and expectation maximization method for global gene trajectory clustering. J. Bioinformatics and Computational Biology 3(5) (2005) 1227-1242

4. Moldovan, D., Silaghi, G.C. : Gene Trajectory Clustering for Learning the Stock Market Sectors. Proceedings of 9th International conference on adaptive and natural computing, ICANNGA 2009, Kuopio, Finland, to appear in Lecture Notes in Artificial Intelligence, Springer-Verlag

5. Elton, E.J., Gruber, M.J., Brown, S.J., Goetzmann, W.N.: Modern Portfolio Theory and Investment Analysis. Wiley (2006)

6. Jeroen Boets, K. De Cock, M. Espinoza, B. De Moor: Clustering time series, subspace identification and central distances. Communications in Information and Systems, International Press, 2005

7. M. Gavrilov, D. Anguelov, P. Indyk, R. Motwani: Mining the Stock Market: Which Measure is Best? Proc. of the KDD, 2000.

8. Doherty, K.A., Adams, R.G., Davey, N., Pensuwon, W.: Hierarchical topological clustering learns stock market sectors. In: Computational Intelligence Methods and Applications, 2005 ICSC Congress on, IEEE Computer Society (2005) 6