Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 345 (2005) 196–206
PHYSICA A
www.elsevier.com/locate/physa

Clustering stock market companies via chaotic map synchronization

N. Basalto[a], R. Bellotti[b,c], F. De Carlo[b,*], P. Facchi[b], S. Pascazio[b]

[a]*Institute for Advanced Studies at University of Pavia, via Bassi 6, I-27100 Pavia, Italy*
[b]*Dipartimento di Fisica, Università di Bari, and Istituto Nazionale di Fisica Nucleare, Sezione di Bari, via Amendola 173, I-70126 Bari, Italy*
[c]*TIRES, Center of Innovative Technologies for Signal Detection and Processing,via Amendola 173, I-70126 Bari, Italy*

Received 19 April 2004
Available online 20 August 2004

Abstract

A pairwise clustering approach is applied to the analysis of the Dow Jones index companies, in order to identify similar temporal behavior of the traded stock prices. To this end, the chaotic map clustering algorithm is used, where a map is associated to each company and the correlation coefficients of the financial time series to the coupling strengths between maps. The simulation of a chaotic map dynamics gives rise to a natural partition of the data, as companies belonging to the same industrial branch are often grouped together. The identification of clusters of companies of a given stock market index can be exploited in the portfolio optimization strategies.

PACS: 89.65.Gh; 05.45.Ra; 05.45.Tp

Keywords: Stock index; Clustering algorithms; Chaotic maps

*Corresponding author. Tel.: +39-080-5442364.
E-mail address: francesco.decarlo@ba.infn.it (F. De Carlo).

0378-4371/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2004.07.034

1. Introduction

Stock markets are recently triggering a growing interest in the physicists' community. The objective of this attention is to understand the underlying dynamics which rules the companies' stock prices. In particular, it would be useful to find, inside a given stock market index, groups of companies sharing a similar temporal behavior. To this purpose, a clustering approach to the problem may represent a good strategy. Clustering deals with the partitioning of a set of N elements into K clusters, based on a suitable (and not unique) similarity criterion [1]. Non-parametric methods represent the optimal strategy when no prior knowledge on the clusters to find is available: these methods make few assumptions about the structure of the data, rather they employ local criteria for reconstructing the clusters, e.g. by searching for high density regions in the data space. Moreover, as the number of clusters is not selected a priori, they are particularly suited when a hierarchical structure, rather than a fixed partition, of the data should be obtained: this is the case with stock index dynamics and portfolio optimization strategies [2,3]. Examples of non-parametric methods are the *linkage* (*agglomerative* and *divisive*) algorithms [4], whose output is a *dendrogram* displaying the full hierarchy of the clustering solutions at different scales. The agglomerative approaches merge, at each step, the two clusters with the smallest *distance*, starting from clusters containing only one element. In this article we use a non-parametric clustering approach, named chaotic map clustering (CMC) [5], which relies on the synchronization properties of a chaotic map system [6,7] in order to obtain a hierarchy of classes without any assumptions on the underlying structure of the data.

This paper is organized as follows: in Section 2 we give a brief review of the chaotic map algorithm, suitably modified for pairwise clustering of financial times series. Section 3 deals with the analysis of the companies' stock prices. Finally, some conclusions are drawn in Section 4.

2. Pairwise chaotic map clustering

The chaotic map clustering was originally introduced as a central algorithm, where the elements to cluster are embedded in a D-dimensional feature space. In such a picture, the data-points are viewed as sites of a grid, hosting a chaotic map dynamics: the map variables $x_i \in [-1, 1]$, $i = 1, ..., N$, are assigned to each site of the lattice, and short-range interactions between neighboring maps are introduced as exponential decreasing function of the site distance. In the stationary regime, clusters of synchronized maps appear, corresponding to high density regions in the original data space. The mutual information between maps is used both as a similarity index for building the clusters, and a scale parameter for reconstructing the hierarchical tree [5].

It should be remarked that a pairwise version of the algorithm can be easily implemented if an $N \times N$ matrix of similarities (not necessarily distances in the mathematical sense) is provided instead of the feature vectors for all data.

As far as one deals with clustering temporal patterns $y_i(t)$, the correlation coefficients $c_{ij} \in [-1, 1]$ are a natural measure of similarity:

$$c_{ij} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{(\langle Y_i^2 \rangle - \langle Y_i \rangle^2)(\langle Y_j^2 \rangle - \langle Y_j \rangle^2)}} , \tag{1}$$

where the temporal averages are computed over the whole duration of the time series. In Ref. [8], the correlation coefficients between financial time series are used as entries into the super-paramagnetic clustering (SPC) algorithm [9,10]. The SPC algorithm shares the same philosophy of the CMC approach, the physical system used to partition the data being an inhomogeneous ferromagnetic model: Potts spin s_i are assigned, instead of map variables, to each data-point and short-range interactions between neighboring sites are introduced. The spin–spin correlation function replaces the mutual information as similarity index for clustering data. In the super-paramagnetic regime, domains of aligned spins appear, corresponding to the classes present in the data.

Kullmann et al. [8] generalize the SPC to the case of anti-ferromagnetic couplings by introducing the following spin–spin strength as a function of the correlation coefficients c_{ij}:

$$J_{ij} = \mathrm{sgn}(c_{ij}) \left(1 - \exp\left\{ -\frac{n-1}{n} \left[\frac{c_{ij}}{a} \right]^n \right\} \right) , \tag{2}$$

where the sign function sgn maps positive/negative correlations between companies' stock prices into positive/negative interactions between Potts spins, n is an even positive integer tuning the shape of the interaction function (whose value should be chosen so that a stable non-trivial partition can be obtained inside the hierarchical solution), and a is the average of the largest correlation coefficients for each sequence [8]:

$$a = \frac{1}{N} \sum_{i=1}^{N} \max_j (c_{ij}) . \tag{3}$$

We shall try to follow a similar strategy in our CMC approach. We first observe that, in order to implement a chaotic map dynamics, the correlation coefficients between financial time series should be mapped into positive interactions between maps, ranging in $[0, 1]$. Hence, we are naturally led to adopt the couplings (2) for $c_{ij} \geqslant 0$, while setting $J_{ij} = 0$ for $c_{ij} < 0$. In this way, we build up a partially coupled map lattice with exponential increasing interactions between positively correlated companies. In the case of randomly coupled systems, although exact synchronization and formation of clusters of identical dynamical states are not found as in the globally coupled case [6], yet, clusters of *almost* synchronized maps are still observed, even for a significant fraction (up to 40–45%) of lacking connections [7]. By retaining the interactions only between positively correlated time series, we bias the formation of *almost* synchronized maps to correspond to groups of companies sharing the same temporal behavior, while anti-correlated companies are likely to belong to different

clusters. The chaotic map dynamics reads

$$x_i(\tau+1) = \frac{1}{C_i}\sum_{j\neq i} J_{ij} f(x_j(\tau)) \,, \tag{4}$$

where $f(x) = 1 - 2x^2$ is the logistic map, $C_i = \sum_{j\neq i} J_{ij}$ is a normalization factor, and τ denotes the evolution time of the chaotic map system (not to be confused with the real time t of the financial series). A detailed description of the above-mentioned dynamics for clustering purposes is described elsewhere [5]; roughly speaking, after a certain equilibration time, the dynamics (4) yields a partition of the maps x_i into synchronized clusters, that remain stable during the remaining part of the τ-evolution. Applications of the CMC algorithm cover a number of fields, such as buried land-mines detection by dynamic infrared imaging [11,12], human evolution study with mitochondrial *DNA* sequences [13], and diagnosis of pathological electroencephalographic patterns affected by Huntington's disease [14,15].

3. Application to financial time series

Here we apply the CMC algorithm to cluster the companies of the Dow Jones (DJ) market index, including $N = 30$ stocks, whose names are listed in Appendix A, together with the identifying tickers and the related industrial branches. We first analyze one-year time periods, from 1998 to 2002. For each year, the correlation coefficients (1) are computed for the logarithmic daily price variation time series

$$y_i(t) = \ln(P_i(t+1)) - \ln(P_i(t)) \,, \tag{5}$$

where $P_i(t)$ is the closure price of stock i at day t.

It should be remarked that, for each investigated period, the number of pairs of anti-correlated companies $N_{c<0}$ is very small in comparison with the total number of pairs $N(N-1)/2 = 435$, and the mean value of the anticorrelations $\langle c\rangle_{c<0}$ is almost zero (see Table 1). At this point it should be stressed that the very fact that almost all stocks are correlated, and practically lack any anticorrelation, makes not easy any possible clustering procedure.

Table 1
Number of pairs of anti-correlated stocks $N_{c<0}$ and mean value of the anticorrelation $\langle c\rangle_{c<0}$

Year	1998	1999	2000	2001	2002
$N_{c<0}$	0	25	34	11	1
$\langle c\rangle_{c<0}$	0	−0.0453	−0.0494	−0.0495	−0.0071

$N_{c<0}$ and $\langle c\rangle_{c<0}$ must be compared with the total number of pairs $N(N-1)/2 = 435$ and with the mean correlation $\langle c\rangle \simeq 0.28$, respectively.

As a result of the processing, a dendrogram displays the hierarchical structure of the clusters at different values of the mutual information I_{ij} defined as follows:

- extract a bitwise sequence S_i from each map $x_i(t)$, such that

$$S_i = \begin{cases} 1 & \text{if } x_i(t) \geqslant 0, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

- evaluate the probability $P(S_i)$ as the number of times the state S_i occurs along the sequence S_i, normalized to the sequence length; in a similar way, $P(S_i, S_j)$ is the frequency of simultaneous occurrence of the states (S_i, S_j) along the sequences S_i and S_j;
- compute the string entropy H_i and the joint entropy H_{ij} as

$$H_i = -\sum_{S_i=0,1} P(S_i) \ln P(S_i)\,, \tag{7}$$

$$H_{ij} = -\sum_{S_i=0,1} \sum_{S_j=0,1} P(S_i, S_j) \ln P(S_i, S_j)\,, \tag{8}$$

- the mutual information is then given by: $I_{ij} = H_i + H_j - H_{ij}$.

The mutual information is a measure of the correlations between maps [16], ranging between $I_{ij} = 0$, for maps evolving independently, and $I_{ij} = \ln 2$, for exactly synchronized maps. For this reason, I_{ij} can be appropriately adopted as a similarity index for clustering the companies: by cutting the dendrogram at a certain level $I \in [0, \ln 2]$, the clusters thus obtained are made up of companies whose associated maps are characterized by $I_{ij} \geqslant I$. The level I can be suitably chosen by relying on a certain stability criterion of the clustering solution. To this purpose, the cluster entropy $S(I)$ [6] can be used to select the most stable partition among the whole hierarchy yielded by the algorithm, by looking for a plateau in the widest possible range of I values:

$$S(I) = -\sum_{k=1}^{N_I} P_I(k) \ln P_I(k)\,, \tag{9}$$

where $P_I(k)$ is the fraction of elements belonging to cluster k, and N_I is the number of clusters found at level I.

This model depends on one parameter, the positive even integer number n, which tunes the range of the interactions (2). For each period, the optimal value of the parameter n should be chosen according to the stability criterion of the entropy (9), at different cluster partitions. As an example, we consider the processing relative to the year 1999: Fig. 1 displays the entropy S in the plane spanned by I (mutual information) and n, with $n = 2, 4, 6, \ldots, 24$. We choose $n = 8$ to be the optimal value, by looking for the widest range of constant values of S, along the I-direction $(0.4 \lesssim I \lesssim 0.6)$.

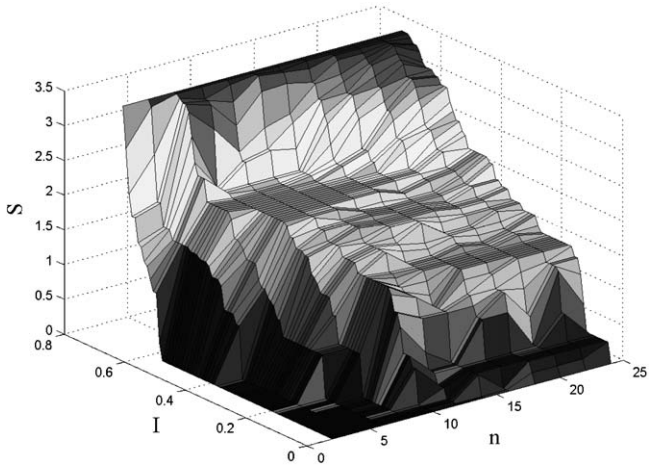

Fig. 1. Cluster entropy S in the plane spanned by the mutual information I and the parameter n. The widest S-plateau along the I -direction (namely, the range of values of I for which S is constant) is $0.4 \lesssim I \lesssim 0.6$ and corresponds to $n = 8$. This analysis refers to year 1999.

Once this parameter has been adjusted, the full hierarchy of clusters can be displayed by a dendrogram: Fig. 2 shows the result obtained for the year 1999. The dendrogram has been cut in the region of stable partitions at $I \simeq 0.6$. For low value of mutual information, all pairs of companies are linked together in one single cluster, which splits into two big clusters at $I = 0.16$: on one side, we clearly recognize companies dealing mainly with capital goods (BA, CAT, HON) and basic materials (AA, DD, IP). On the other side, we find a cluster of strongly correlated companies represented by the branch marked by a star. This cluster, which gradually breaks as the mutual information approaches its maximum value $I = \ln 2$, groups together different industrial branches: financial (C, AXP, JPM), services (DIS, HD, MCD, SBC, T, WMT), healthcare (JNJ, MRK), conglomerates (GE, UTX), consumers non-cyclical (GM, KO, MO, PG). Besides this cluster, it should be remarked the formation of technological cores (IBM and HPQ, INTC and MSFT).

For comparative purposes, the financial time series have been clustered by means of an alternative simpler method: the agglomerative single linkage (SL) algorithm. In this procedure, one first introduces the distance between the time series i and j [17,18]:

$$d_{ij} = \sqrt{2(1 - c_{ij})}\,, \tag{10}$$

where c_{ij} is the correlation coefficient. The SL algorithm consists in merging at each step the two closest clusters, relying on the following similarity index between clusters $\mathscr{C}$ and $\mathscr{C}'$:

$$d(\mathscr{C}, \mathscr{C}') = \min_{i \in \mathscr{C}, j \in \mathscr{C}'} d_{ij}\,. \tag{11}$$

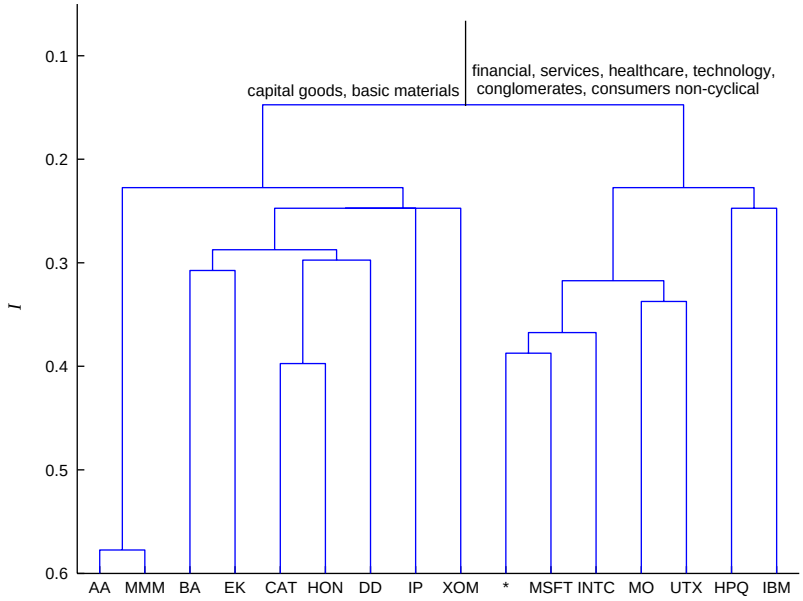

Fig. 2. Dendrogram obtained for the year 1999 ($n = 8$), cut in the region of stable partitions at $I \simeq 0.6$. The branch marked by a star (not explicitly shown) groups together different industrial sub-classes: financial (C, AXP, JPM), services (DIS, HD, MCD, SBC, T, WMT), healthcare (JNJ, MRK), conglomerates (GE, UTX), consumers non-cyclical (GM, KO, MO, PG).

The dendrogram obtained as a result of the SL clustering is displayed in Fig. 3: the so called *chaining effect*, giving rise to elongated clusters, can be clearly observed. Notice that this drawback is completely absent in our procedure, as can be seen in Fig. 2.

The analysis has been carried out for each of the 5 years considered (1998–2002). In the following, we report the main clusters found for different years, together with the values chosen for the parameter n, and the values of the mutual information at which the dendrogram has been cut. Sub-clusters of companies belonging to the same industrial branch have been underbraced:

- *Year 1998*, $n = 16$, $I = 0.62$
  (1) DIS MCD T WMT KO MO PG JNJ MRK
  (2) AXP C JPM GM
- *Year 1999*, $n = 8$, $I = 0.24$
  (1) DIS HD MCD SBC T WMT KO MO PG AXP C JPM JNJ MRK INTC MSFT GE UTX GM
  (2) BA CAT HON DD IP EK XOM
- *Year 2000*, $n = 18$, $I = 0.26$
  (1) BA CAT HON AA DD IP KO PG MMM UTX EK MCD
  (2) AXP C JPM SBC T GE GM

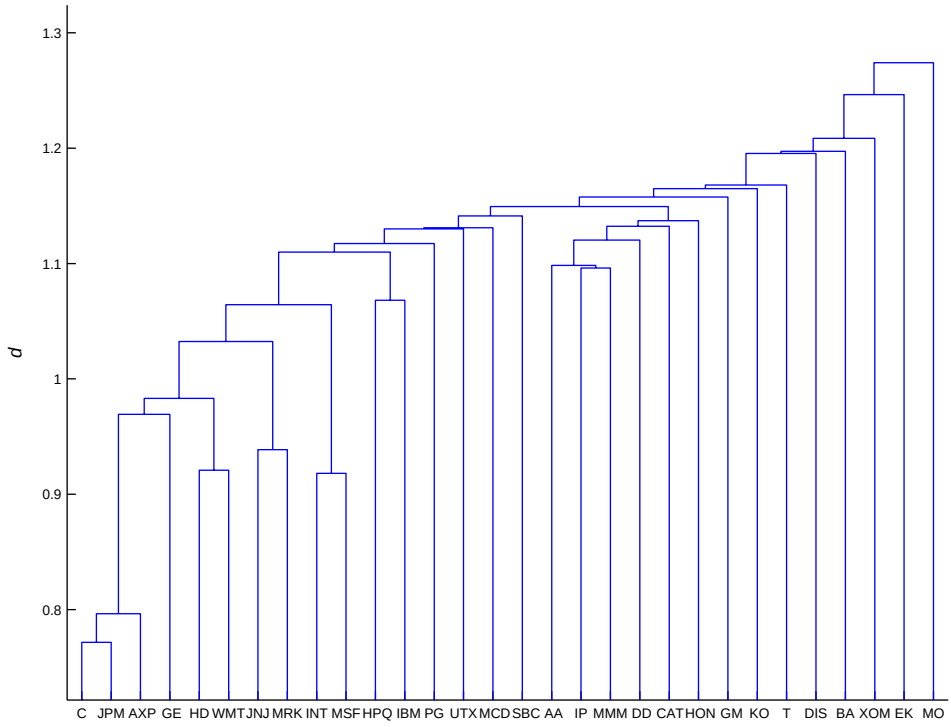

Fig. 3. Dendrogram obtained for the year 1999 using the agglomerative single linkage algorithm. The similarity index d is defined in Eq. (11). Notice the "chaining effect", that tends to yield elongated clusters.

- *Year 2001*, $n = 20$, $I = 0.15$
  (1) DIS HD MCD SBC T WMT BA CAT HON AXP C JPM AA DD IP GE MMM UTX EK GM MO XOM
  (2) HPQ IBM INTC MSFT
- *Year 2002*, $n = 16$, $I = 0.62$
  (1) AA DD IP CAT HON MMM UTX GM MCD XOM;
  (2) AXP C JPM DIS SBC EK GE MRK
  (3) HPQ IBM MSFT HD.

It is worth stressing the presence of some cores of companies which remain strongly linked together over periods longer than 1 year: financial companies (AXP, C, JPM, 98–02), services (DIS, MCD, T, WMT, 98–99, 01), consumers non-cyclical

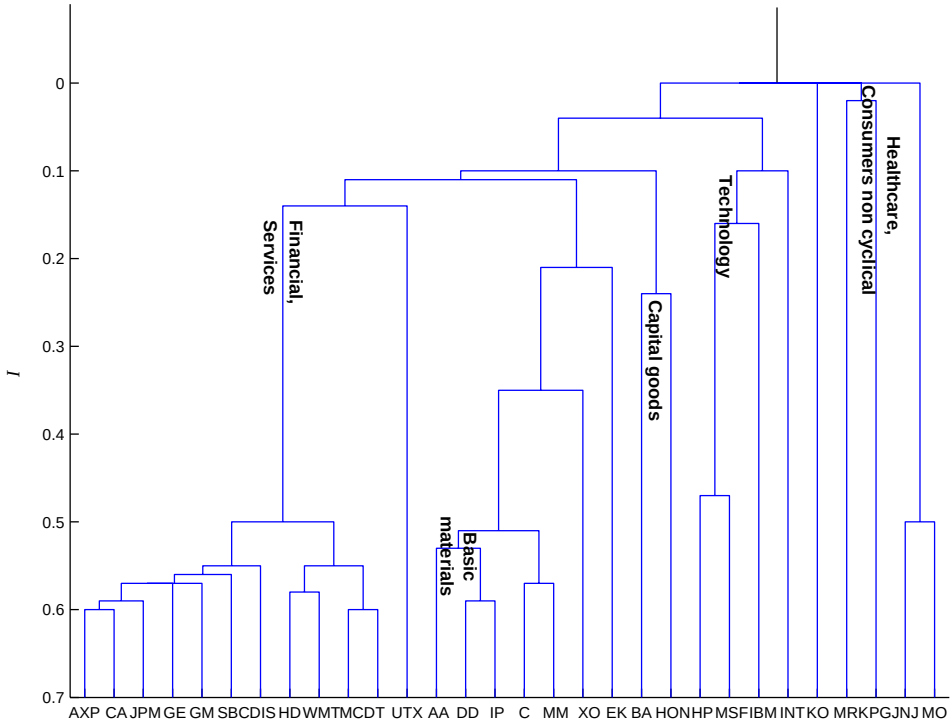

Fig. 4. Dendrogram found from the whole 5-year time period 1998–2002, with $n = 18$. The main branches have been marked by the industrial areas of the companies they are made of.

(KO, MO, PG, 98–99), basic materials (AA, DD, IP, 00–02), capital goods (BA, CAT, HON, 99–01), technology (HPQ, IBM, MSFT, 01–02), healthcare (JNJ, MRK, 98–99), conglomerates (MMM, UTX, 00–02).

Once a partition of companies has been obtained, an efficient portfolio could be made of one "representative" stock per cluster, thus ensuring a diversification of the investment. The choice of the period length for computing the correlation coefficients should be related to the flexibility of the portfolio. From this point of view, an analysis covering the whole 5-year period should be based on more stable correlation coefficients, thus leading to more stable partitions (i.e., less hazardous investment), at the cost of a less flexible portfolio. In Fig. 4, we report the full hierarchy of clusters found from the whole 5-year length time period ($n = 18$). We want to remark that no anticorrelations have been found for such period. The main branches of the dendrogram have been marked by the industrial areas of the companies they are made of.

4. Conclusions

In the present work, a pairwise version of the chaotic map algorithm has been applied to the analysis of the companies' stocks belonging to the Dow Jones market index. The correlation coefficients between financial time series have been used as similarity measures to cluster the temporal patterns. Once the coupling interactions between maps are taken to be functions of these coefficients, the dynamics of such a system leads to the formation of clusters of companies that can often be identified as different industrial branches. The clustering output can be exploited to optimize the portfolio composition.

Appendix A. Dow Jones stock market companies

AA	Alcoa Inc.—Basic Materials
AXP	American Express Co.—Financial
BA	Boeing—Capital Goods
C	Citigroup—Financial
CAT	Caterpillar—Capital Goods
DD	DuPont—Basic Materials
DIS	Walt Disney—Services
EK	Eastman Kodak—Consumer Cyclical
GE	General Electrics—Conglomerates
GM	General Motors—Consumer Cyclical
HD	Home Depot—Services
HON	Honeywell International—Capital Goods
HPQ	Hewlett-Packard—Technology
IBM	International Business Machine—Technology
INTC	Intel Corporation—Technology
IP	International Paper—Basic Materials
JNJ	Johnson & Johnson—Healthcare
JPM	JP Morgan Chase—Financial
KO	Coca Cola Inc.—Consumer Non-Cyclical
MCD	McDonalds Corp.—Services
MMM	Minnesota Mining—Conglomerates
MO	Philip Morris—Consumer Non-Cyclical
MRK	Merck & Co.—Healthcare
MSFT	Microsoft—Technology
PG	Procter & Gamble—Consumer Non-Cyclical
SBC	SBC Communications—Services
T	AT&T Gamble—Services
UTX	United Technology—Conglomerates
WMT	Wal-Mart Stores—Services
XOM	Exxon Mobil—Energy

References

[1] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, San Diego, 1990.
[2] E.J. Elton, M.J. Gruber, Modern Portfolio Theory and Investment Analysis, Wiley, New York, 1995.
[3] J.-P. Bouchaud, M. Potters, Theory of Financial Risks, Cambridge University Press, Cambridge, 1999.
[4] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, New York, 1988.
[5] L. Angelini, F. De Carlo, C. Marangi, M. Pellicoro, S. Stramaglia, Phys. Rev. Lett. 85 (2000) 554.
[6] K. Kaneko, Phys. Rev. Lett. 63 (1989) 219;
K. Kaneko, Physica D 41 (1990) 137;
K. Kaneko, Physica D 75 (1994) 55.
[7] S.C. Manrubia, A.S. Mikhailov, Phys. Rev. E 60 (1999) 1579.
[8] L. Kullmann, J. Kertesz, R.N. Mantegna, Physica A 287 (2000) 412.
[9] M. Blatt, S. Wiseman, E. Domany, Phys. Rev. Lett. 76 (1996) 3251;
M. Blatt, S. Wiseman, E. Domany, Neural Inform. Proc. Syst. 8 (1996) 416;
M. Blatt, S. Wiseman, E. Domany, Neural Comput. 9 (1997) 1805;
M. Blatt, S. Wiseman, E. Domany, Phys. Rev. E 57 (1998) 3767.
[10] E. Domany, Physica A 263 (1999) 158.
[11] C. Marangi, L. Angelini, F. De Carlo, G. Nardulli, M. Pellicoro, S. Stramaglia, in: S.B. Serpico (Ed.), Proceedings of the EoS/Spie Symposium on Remote Sensing: image and Signal Processing VI, vol. 4170, Barcellona, Spain, 2000, pp. 122–132.
[12] L. Angelini, F. De Carlo, C. Marangi, M. Mannarelli, G. Nardulli, M. Pellicoro, G. Satalino, S. Stramaglia, Opt. Eng. 40 (2001) 2878.
[13] C. Marangi, L. Angelini, M. Mannarelli, M. Pellicoro, S. Stramaglia, M. Attimonelli, M. De Robertis, L. Nitti, G. Pesole, C. Saccone, M. Tommaseo, in: G. Nardulli, S. Stramaglia (Eds.), Proceedings of the International Workshop Modelling Biomedical Signals, Bari, Italy, 2001, pp. 196–208.
[14] R. Bellotti, F. De Carlo, S. Stramaglia, Physica A 334 (2004) 222–232.
[15] R. Bellotti, M. Castellano, F. De Carlo, IEEE Trans. Nucl. Sci. 51 (3) (2004).
[16] R.V. Solè, S.C. Manrubia, J. Bascompte, J. Delgado, B. Luque, Complexity 2 (1996) 13.
[17] R.N. Mantegna, Eur. Phys. J. B 11 (1999) 193–197.
[18] R.N. Mantegna, H.E. Stanley, Introduction to Econophysics, Cambridge University Press, Cambridge, UK, 2000.