Eur. Phys. J. B **57**, 139–146 (2007)
DOI: 10.1140/epjb/e2007-00132-5

THE EUROPEAN PHYSICAL JOURNAL B

Cluster structure of EU-15 countries derived from the correlation matrix analysis of macroeconomic index fluctuations

M. Gligor[1,2,a] and M. Ausloos[1,b]

[1] GRAPES, Université de Liège, B5 Sart-Tilman, 4000 Liège, Belgium
[2] National College 'Roman Voda', Roman-5550, Neamt, Romania

Received 31 August 2006 / Received in final form 28 December 2006
Published online 16 May 2007 –

Abstract. The statistical distances between countries, calculated for various moving average time windows, are mapped into the ultrametric subdominant space as in classical Minimal Spanning Tree methods. The Moving Average Minimal Length Path (MAMLP) algorithm allows a decoupling of fluctuations with respect to the mass center of the system from the movement of the mass center itself. A Hamiltonian representation given by a factor graph is used and plays the role of cost function. The present analysis pertains to 11 macroeconomic (ME) indicators, namely the GDP (x_1), Final Consumption Expenditure (x_2), Gross Capital Formation (x_3), Net Exports (x_4), Consumer Price Index (y_1), Rates of Interest of the Central Banks (y_2), Labour Force (z_1), Unemployment (z_2), GDP/hour worked (z_3), GDP/capita (w_1) and Gini coefficient (w_2). The target group of countries is composed of 15 EU countries, data taken between 1995 and 2004. By two different methods (the Bipartite Factor Graph Analysis and the Correlation Matrix Eigensystem Analysis) it is found that the strongly correlated countries with respect to the macroeconomic indicators fluctuations can be partitioned into stable clusters.

PACS. 89.65.Gh Economics; econophysics, financial markets, business and management – 89.75.Fb Structures and organization in complex systems – 05.45.Tp Time series analysis

1 Introduction

Modelling the dependences between the macroeconomic (ME) variables has to take into account circumstances that differ substantially from those encountered in the natural sciences. First, experimentation is usually not feasible and is replaced by survey research, implying that the explanatory variables cannot be manipulated and fixed by the researcher. Second, the number of possible explanatory variables is often quite large, unlike the small number of carefully chosen treatment variables frequently found in the natural sciences. Third, the ME time series are short and noisy. Most data have a yearly frequency. When social time series have been produced for a very long period, there is usually strong evidence against stationarity.

Some macroeconomic (ME) indicators are monthly and/or quarterly registered, increasing in this way the number of available data points, but some additional noise is naturally enclosed in the time series so generated (seasonal fluctuations, external and internal short range shocks, etc.). This seems to be a solid argument for the fact that the main data sources, at least the ones freely available on the web, tend only to keep the annual averages/rates of growth of the ME indicators.

Let us consider, for example, a time interval of one hundred years, which is mapped onto a graphical plot of 100 data points. From the statistical physics viewpoint, 100 is a quite small number of data points, surely too small for speaking about the so called *thermodynamic limit*. On the other hand, from a socio-economic point of view, we can justifiably wonder if a growth, say, of 2% of any ME indicator has at the present time the same meaning as it had one century ago. One must take into account that during that time, the social, politic and economic environment was drastically changed. Moreover the methodology of data collecting and processing is today different from what it was two generations ago. Indeed, the economic world is created by people and is substantially changing from a generation to another one (sometimes also during one and the same generation). Thus, this way of statistical data aggregation turns to be controversial.

Several papers [1,2] investigated the statistical patterns in GDP annual rates of growth by aggregating (in a "horizontal" way) the data from all countries for which statistical data were reported. Even if all data are supposed to be reliable, and even if the relative rates of growth are investigated (to diminish the actual large difference influences), this way of aggregation, as well as the previous one, supposes a priori a certain degree of homogeneity across countries. A certain GDP rate of growth in an underdeveloped country is certainly based on factors that

[a] e-mail: mgligor@ulg.ac.be
[b] e-mail: Marcel.Ausloos@ulg.ac.be

differ substantially from the ones that generate the same rate of growth in a developed country. Both theoretical and empirical investigations [3,4] reported the evidence of the country partitioning in clusters after their common patterns of evolution. For such subsystems only, the data might be meaningfully aggregated. In the present paper we demonstrate *the clustering emergence in the relatively stable and homogeneous system* composed of the 15 EU countries for data taken between 1994 and 2004, starting from the annual rates of growth of 11 ME indicators, namely the GDP (x_1), Final Consumption Expenditure (x_2), Gross Capital Formation (x_3), Net Exports (x_4), Consumer Price Index (y_1), Rates of Interest of the Central Banks (y_2), Labour Force (z_1), Unemployment (z_2), GDP/hour worked (z_3), GDP/capita (w_1) and Gini coefficient (w_2).

One has to stress here that the problem of studying the patterns of growth across countries is actually a subject of great attention to economists [4,5]. An important reason for the increasing interest in this problem is that persistent disparities in aggregate growth rates across countries have, over time, led to large differences in welfare. On the other hand, the intellectual payoffs are high: various statistical tools might be considerably enriched and extended by applying them to the non-stationary, short and noisy macroeconomic time series.

In the present paper we focus on two recent lines of research, of growing interest in physics, which can bring important contributions to ME time series analysis. On one hand, the recent developments in nonequilibrium networks [6]; on the other hand, the random matrix theory (RMT), initially developed in nuclear physics, also successfully used in the study of canonical correlations between stock changes and portfolio optimization problem [7]. The way in which these methods are adapted to the macroeconomic time series analysis is described in the next section.

The Minimal Spanning Tree (MST) is one of the most usual methods in cluster analysis, and has been largely used so far both by physicists [8] and economists [4]. Nonetheless, both sides [4,7] noted some lack of univocity due to choosing the MST root. Moreover, the MST structure proves to be not stable when a constant size time window is moved over the considered time span. The solution briefly presented in Section 3, namely the Moving Average Minimal Length Path (MAMLP) method comes as a development of some previous methods where some arbitrariness in the root of the tree was underlined considering that an a priori more common root, like the sum of the data, called the "All" country, from which to let the tree grow was permitting a better comparison [9].

The target group of countries is composed of 15 EU countries, data taken between 1994 and 2004. The main sources used for all the above indicators annual rates is the World Bank database [10] and the OECD database [11]. We abbreviate the countries according to the Roots Web Surname List (RSL) which uses 3 letters standardized abbreviations to designate countries and other regional locations (`http://helpdesk.rootsweb.com/codes/`). Inside the tables, for spacing reasons we use the countries two letters abbreviation (`http://www.iso.org`).

The remainder of the paper is organized as follows: in Section 2 the theoretical and methodological tools from the network analysis and matrix theory which we try to adapt to the considered time series are briefly described. The results are largely presented and discussed in Section 3. Some concluding remarks are done in Section 4.

2 Theoretical and methodological framework

As mentioned in Section 1, MST cannot be built in a unique way, whence this becomes a problem when we try to construct a cluster hierarchy for each position of a moving time window. The hierarchical structure proved to be not robust against fluctuations induced by a moving time window. In the MAMLP method described here below we propose to construct the hierarchy starting from a virtual 'average' agent. The method is developed in the following steps:

(i) an 'AVERAGE' agent (AV) is virtually included into the system; the statistical distance matrix is constructed, having the elements: $d_{ij} = [2(1 - C_{ij})]^{1/2}$, where C_{ij} is the correlation coefficient between the ME time series corresponding to the $i - j$ pair of countries in the considered time interval, T. The matrix elements are thereafter set into increasing order (i.e. the decreasing order of correlations);

(ii) the hierarchy is constructed, connecting each agent by its minimal length path (MLP) to AV. Its minimal distance to AV, $\hat{d}_i(t)$, is associated to each agent;

(iii) the procedure is repeated by moving a given and constant time window (in this case a $T = 5$ years time window size) over the investigated time span (in the present analysis: 1994–2004). The agents are sorted through their movement inside the hierarchy. Therefore, a new correlation matrix between country distances to their own mean is constructed. The matrix elements are defined as:

$$\hat{C}_{i,j}(t) = \frac{\langle \hat{d}_i(t)\hat{d}_j(t)\rangle - \langle \hat{d}_i(t)\rangle\langle \hat{d}_j(t)\rangle}{\sqrt{\langle (\hat{d}_i(t))^2 - \langle \hat{d}_i(t)\rangle^2\rangle\langle (\hat{d}_j(t))^2 - \langle \hat{d}_j(t)\rangle^2\rangle}} \quad (1)$$

where $\hat{d}_i(t)$ is the i-country minimal length path (MPL) distance to the AVERAGE. For simplicity, the explicit dependencies on the time window size T are not included in equation (1). The angular brackets in equation (1) represent averages over the different distances (country pairs) obtained as the time window is moved.

As we shall show in Section 3.3, for five of the analysed indicators, namely for $x_1 \equiv$ GDP, $x_2 \equiv$ Consumption, $x_3 \equiv$ Capital Formation, $w_1 \equiv$ GDP/capita and $y_2 \equiv$ Interest Rates, there is some sort of countries collective movement, while for the other six ME indicators there is no such tendency. However, the average rate of growth can *always* be defined by simply arithmetical averaging

the individual rates of EU-15 countries (the average behaviour is often analysed in the economics papers, e.g. in the OECD reports [11]). The algorithm above described is nothing else but the classical Minimal Spanning Tree on the condition that *the root of the tree is the "average" agent instead of one of the strongest correlated agents.* The "average country" plays the role of the mass centre of the system, with respect to which the movements of the other ("real") countries are analysed.

Two points must be also stressed here. Firstly the present study pertains to the *fluctuations* of the ME indicators, not to their actual values (the rough data used are the annual rates of growth). If two countries display a correlated movement with respect to the average (the both going near or the both going far from the average), one may suspect some kind of economic interaction between them. In this sense, $\hat{C}_{ij}$ can be seen as a measure of one country *sensitivity* to the economic fluctuations of another, rather than a direct correlation or anti-correlation derived from the rough ME time series.

Secondly, if one of the time series is *constant* in the investigated time window, then $\hat{C}_{ij}$ becomes undetermined, as a fraction of zero. This problem sometimes arises when the (Pearson's) correlation coefficient is calculated in a finite-size time window. This result, rather than expressing a mathematical limit, simply shows that the correlation coefficient cannot be defined if one of the two time series has no variability. Note that this particular situation is more likely to arise when the classical MST is applied than one applies the MAMPL algorithm. Indeed, while two countries can keep their statistical distance unchanged, it is very unlikely one of them to maintain itself at a constant distance to the average (as long as the average is depending on all the other countries idiosyncratic behaviour).

Let us recall that for systems with discrete degrees of freedom, denoted by s, the statistical mechanical models are generally defined through the Hamiltonian $H = H(s)$, which is typically a sum of terms, each involving a small number of variables. A useful representation is given by the *factor graph* [12]. A factor graph is a bipartite graph made of variable nodes $i, j, \ldots$ one for each variable, and function nodes $a, b, \ldots$ one for each term of the Hamiltonian. In the present approach the variable nodes are the macroeconomic indicators and the function nodes are the countries. An edge joins a variable node i and a function node a if and only if $i \in a$, i.e., the variable s_i appears in H_a - the term of the Hamiltonian associated to a. The Hamiltonian can then be written as:

$$H = \sum_a H_a(s_a), \text{ with } s_a = \{s_i, i \in a\}. \tag{2}$$

In combinatorial optimization problems [12], the Hamiltonian plays the role of a *cost function*. In the low temperature limit $T \to \infty$, one is interested by only minimal energy states (ground states) having a non-vanishing probability.

Usually, a *cluster* k is defined as a subset of the factor graph such that if a function node belongs to k, then all the variable nodes $i \in a$ also belong to k (while the converse needs not to be true, otherwise the only legitimate clusters would be the connected components of the factor graph). Here, this condition will be relaxed by partitioning the function nodes after the criterion if it is connected or not to a certain variable node.

Once the correlation matrix is constructed, it is natural to ask for the interpretation of its eigenvalues and eigenvectors. Note that since the matrix is symmetric, the eigenvalues are all real numbers. We will call $\mathbf{v}_a$ the normalized eigenvector corresponding to eigenvalue λ_a, with $a = 1, 2, \ldots, M$. The vector $\mathbf{v}_a$ is the list of the weights $v_{a,i}$ in this linear combination of the different countries. The variance corresponding to such a combination is thus:

$$\sigma_a^2 = \left\langle \left(\sum_{i=1}^{M} v_{a,i} \hat{d}_i \right)^2 \right\rangle = \sum_{i,j=1}^{M} v_{a,i} v_{a,j} \hat{C}_{i,j} \equiv \mathbf{v}_a \cdot \hat{C} \mathbf{v}_a. \tag{3}$$

Furthermore, using the fact that different eigenvectors are orthogonal, we obtain a set of uncorrelated random fluctuations e_a, which are the elements of the system constructed from the weights $v_{a,i}$:

$$e_a = \sum_{i=1}^{M} v_{a,i} \hat{d}_i, \text{ where } \langle e_a e_b \rangle = \lambda_a \delta_{a,b}. \tag{4}$$

Conversely, one can think of the initial distances as a linear combination of the uncorrelated factors E_a:

$$\hat{d}_i = \sum_{a=1}^{M} v_{a,i} e_a. \tag{5}$$

In this decomposition, usually called "the principal component analysis", the correlated fluctuations of a set of random variables are decomposed in terms of the fluctuations of underlying uncorrelated factors. In the case of the country clustering, the principal components E_a could have an economic interpretation in terms of the macroeconomic indicators.

Since, as generally accepted [7,13], the largest eigenvectors are the ones carrying the useful information, one can try to define clusters on the basis of the structure of these eigenvectors. Often (but not always), the largest one, $\mathbf{v}_1$, has comparable and of the same sign components on all countries, and defines the largest cluster, containing all countries. The second one, $\mathbf{v}_2$, which by construction has to be orthogonal to $\mathbf{v}_1$, may have some of its components positive, and the others negative. This means that a probable move of the countries around the average (global) fluctuations occurs when some countries over-perform the average, and others under-perform it. Therefore, the sign of the components of $\mathbf{v}_2$ can be used to group the countries in two families. Each family can then be divided further, using the relative signs of $\mathbf{v}_3$, $\mathbf{v}_4$, etc.

Table 1. MPL distances to AVERAGE. The moving time window size is 5 years for data taken from 1994 to 2004.

	AU	BE	DE	DK	ES	FI	FR	UK	GR	IE	IT	LU	NL	PT	SE
94–98	0.67	0.86	0.86	0.86	0.40	0.40	0.67	0.86	0.40	0.86	0.86	0.40	0.40	0.86	0.86
95–99	0.60	0.65	0.52	0.71	0.21	0.77	0.45	0.77	0.37	0.65	0.90	0.37	0.23	0.83	0.52
96–00	0.58	0.32	0.46	0.61	0.34	0.81	0.46	0.32	0.32	0.53	0.32	0.20	0.60	0.60	0.46
97–01	0.48	0.30	0.48	0.30	0.28	0.42	0.48	0.44	0.68	0.38	0.68	0.14	0.28	0.28	0.48
98–02	0.43	0.26	0.19	0.19	0.21	0.43	0.19	0.19	1.04	0.29	0.44	0.12	0.21	0.21	0.29
99–03	0.25	0.23	0.19	0.19	0.29	0.26	0.19	0.37	1.15	0.26	0.37	0.23	0.19	0.19	0.28
00–04	0.27	0.27	0.17	0.26	0.28	0.27	0.21	0.27	0.53	0.50	0.28	0.27	0.21	0.21	0.27

Table 2. The correlation matrix of EU-15 country movements inside the hierarchy. Indicator: GDP. The moving time window size is 5 years for data taken from 1994 to 2004.

	AU	BE	DE	DK	ES	FI	FR	UK	GR	IE	IT	LU	NL	PT	SE
AU	1	0.77	**0.88**	**0.88**	0.33	0.69	**0.88**	0.69	−0.69	0.75	0.71	0.42	0.61	**0.89**	**0.85**
BE		1	**0.88**	**0.90**	0.41	0.27	0.80	**0.94**	−0.59	**0.92**	**0.83**	**0.85**	0.23	**0.90**	**0.91**
DE			1	**0.90**	0.61	0.35	**0.98**	**0.86**	−0.65	**0.85**	0.78	0.61	0.52	**0.86**	**0.99**
DK				1	0.50	0.58	**0.87**	**0.84**	−0.80	**0.93**	0.67	0.77	0.58	**0.99**	**0.88**
ES					1	−0.10	0.61	0.34	−0.38	0.55	0.05	0.36	0.66	0.37	0.64
FI						1	0.42	0.25	−0.62	0.34	0.27	0.14	0.60	0.64	0.26
FR							1	0.79	−0.71	0.81	0.73	0.52	0.60	0.82	**0.95**
UK								1	−0.52	0.82	**0.90**	**0.85**	0.12	**0.86**	**0.86**
GR									1	−0.82	−0.38	−0.56	−0.62	−0.76	−0.60
IE										1	0.63	**0.85**	0.43	**0.89**	**0.87**
IT											1	0.59	−0.05	0.73	0.77
LU												1	0.06	0.77	0.65
NL													1	0.50	0.47
PT														1	**0.84**
SE															1

3 Results

3.1 The statistics of the correlation coefficients

In order to exemplify the MAMPL method, the corresponding steps for x_1 = GDP are explicitly shown below. Firstly, the virtual 'AVERAGE' country is introduced in the system. The statistical distances corresponding to the fixed 5 years moving time window are set in increasing order and the minimal length path (MPL) connections to the AVERAGE are established for each country in every time interval (Tab. 1).

The resulting hierarchy is found to be changing from a time interval to another. Therefore, corresponding correlation matrix is built, this time for the country movements inside the hierarchy (Tab. 2). The above procedure is repeated for each macroeconomic indicator. Thus, the MAMPL method leads us to a set of $M = 11$ correlation matrices, having size $N \times N$, where $N = 15$ is the number of countries under consideration.

Firstly, we analyse the whole set of correlation coefficients. A correlation coefficient $\hat{C}_{i,j}$ will be taken into account as representing a strong connection if and only if $|\hat{C}_{i,j}| > C_{thr}$, where C_{thr} is a certain a priori chosen threshold value. For small values of the C_{thr}, all 15 countries have at least one strong connection, i.e. the graph is fully connected. Increasing the C_{thr}, the number of the connections decreases. In Figure 1 the relative number of links (the ratio between the number of actual links and the number of all possible links) is plotted versus the threshold value. One can observe that the data is well fitted by a low order polynomial. In Figure 1 the cumulative distribution of the correlation coefficients is also plotted (now, the values are the cumulative frequencies and the abscissas are the corresponding correlation coefficients). For comparison, the cumulative uniform distribution is also plotted. The high value of the square of the Pearson product moment correlation coefficient, $R^2 > 0.99$, indicates a good fit of both distributions.

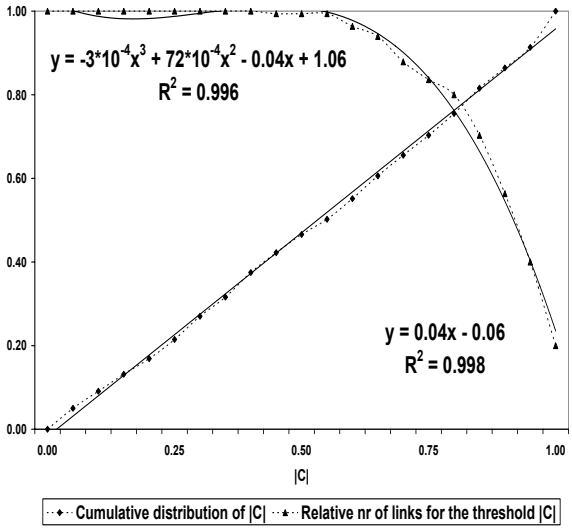

Fig. 1. The cumulative distribution of the correlation coefficients and the relative number of connections versus the $|\hat{C}_{i,j}| \equiv |C|$ (respectively $C_{thr} \equiv |C|$).

Nevertheless, performing the χ^2 test over the whole set of correlation coefficients we must reject the null hypothesis of the fitting $|C|$ distribution by the uniform in the confidence interval of 99%. Investigating by sight the data set one remarks an anomalous large number of correlation coefficients ($N_{20} = 100$) in the range 0.95–1.00, while the mean of the distribution is 57.75 and the standard deviation is $\sigma = 7.45$. According to Chebyshev's theorem [14], an interval of ± 4 standard deviations ensures that at least 94% of the data (of an arbitrary distribution) falls inside this interval. Thus, the last point of the distribution can be treated as an outlier, and, performing the χ^2 test for the remainder points we can accept the hypothesis of the same distribution in a confidence interval of over 75%. We must note here that the same conclusion is supported by t-Student's test in a confidence interval of 100%, the two distributions having *exactly* the same mean. Joining together the results of the statistical tests, we can conclude that the correlation coefficients distribution is a uniform distribution.

3.2 The bipartite factor graph analysis

As it has been already shown, the factor graph structure is strongly dependent on the threshold value C_{thr}. In order to establish the most appropriate C_{thr}, a two tailed t-test of statistical significance is performed over the correlation matrix elements [14]. The null hypothesis (a correlation coefficient of zero) assumes that there is no linear relationship between the two variable sets. In order to test the significance of the correlation coefficients we use the test statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{6}$$

where $r \equiv \hat{C}_{i,j}$ and $n = 2$ is the number of degrees of freedom. The correlation coefficient is considered to be statistically significant if the computed t value is greater than the critical value t_C of a t-Student's distribution with a level of significance of α. From equation (6) one derives:

$$r_C = \frac{t_C}{\sqrt{t_C^2 + n - 2}}. \tag{7}$$

Taking $n = 7$ (the number of statistical distances used for computing each correlation coefficient, from the t-Student distribution tables we find the critical value $t_C = 3.365$ for a reasonable level of significance $\alpha = 0.02$ (or, equivalently, 98% confidence interval). From equation (7) we get $r_C \equiv C_{thr} = 0.83$ i.e. the null hypothesis can only be rejected for the correlation coefficients greater or at least equal to this value. The significant correlation coefficients are emphasized in bold in Table 2.

It is interesting to remark that the two plots from Figure 1 do intersect at the abscissa 0.83 which is equal to the r_C above found. The intersection point seems to correspond to an *optimal* choosing of C_{thr}, under the constrain of the competition between link removing and the remainder correlations to be taken into account.

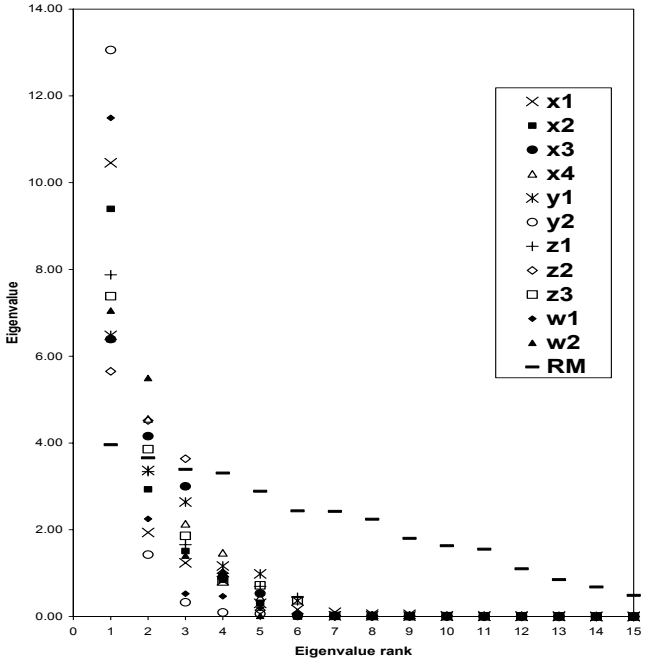

Fig. 2. The eigenvalue spectrum of the correlation matrices between EU-15 country movements with respect to AVERAGE, for each ME indicator (inset). RM: the eigenvalue spectrum of the random matrix.

One can easily see that not all 15 countries (function nodes) are connected through the variable node x_1 (GDP fluctuations), but only 11 of them. Their contributions to the Hamiltonian include the variable x_1.

The above procedure is repeated for each ME variable and leads us to the Hamiltonian (or cost function) having the form: $H = AUT(x_1, x_2, x_3, x_4, y_2, z_1, z_2, z_3, w_1, w_2) +$
$BEL(x_1, x_2, x_3, y_1, y_2, z_1, z_3, w_1, w_2) +$
$DEU(x_1, x_2, x_4, y_1, y_2, z_1, z_2, z_3, w_1, w_2) +$
$DNK(x_1, x_3, x_4, y_2, z_1, z_2, w_1, w_2) +$
$ESP(x_2, x_3, y_2, z_1, z_2, w_1, w_2) +$
$FIN(x_3, x_4, y_1, y_2, z_2, z_3, w_1, w_2) +$
$FRA(x_1, x_3, x_4, y_1, y_2, z_2, z_3, w_1, w_2) +$
$GBR(x_1, x_2, x_3, x_4, y_1, y_2, z_1, z_2, z_3, w_1, w_2) +$
$GRC(x_4, y_1, z_2, w_1, w_2) +$
$IRL(x_1, x_2, x_3, x_4, y_1, y_2, z_2, w_1, w_2) +$
$ITA(x_1, x_4, y_1, y_2, z_1, z_2, w_1, w_2) +$
$LUX(x_1, x_4, y_1, y_2, z_1, z_2, z_3, w_1, w_2) +$
$NLD(x_2, x_4, y_2, z_2, w_1, w_2) +$
$PRT(x_1, x_2, x_3, x_4, y_1, y_2, z_1, z_2, z_3, w_1, w_2) +$
$SWE(x_1, x_2, x_3, x_4, y_1, y_2, z_2, w_1, w_2).$

3.3 The correlation matrix analysis

From the result of the bipartite graph analysis, some countries binary partition in respect to each ME variable can be already seen: a country is connected or not to the respective variable node. Nonetheless, a complete solution to this problem can only be obtained by analyzing the correlation matrix eigensystems. A parallel to similar results from the stock market investigation [7,13] can be also drawn.

Table 3. The first eigenvector components.

	GDP	CONS	CAPF	NEXP	CPI	INTR	LABF	UNEMP	GDPH	GDPC	GINI
AU	−0.276	−0.300	0.373	−0.328	−0.109	−0.274	0.239	0.305	−0.294	−0.289	−0.261
BE	−0.287	−0.325	0.357	0.189	0.003	−0.271	0.308	0.229	−0.351	−0.259	−0.371
DE	−0.296	−0.304	0.257	−0.371	−0.334	−0.274	−0.343	0.299	−0.284	−0.261	−0.122
DK	−0.303	−0.097	0.281	0.111	−0.003	−0.276	−0.293	−0.250	−0.161	−0.287	−0.131
ES	−0.167	−0.325	0.356	−0.171	−0.260	−0.276	0.331	−0.271	0.244	−0.275	0.360
FI	−0.155	−0.159	0.277	0.077	0.342	−0.268	−0.199	−0.322	−0.343	−0.213	−0.047
FR	−0.288	−0.188	0.356	0.282	0.368	−0.272	0.100	0.372	−0.320	−0.229	0.317
UK	−0.274	−0.321	0.088	0.244	0.003	−0.234	0.328	−0.322	−0.352	−0.250	−0.310
GR	−0.239	−0.103	0.132	0.048	−0.266	−0.189	0.152	0.230	0.130	0.257	0.360
IE	−0.290	−0.325	0.274	0.351	0.300	−0.276	−0.163	−0.322	0.068	−0.282	0.188
IT	−0.236	0.001	−0.053	−0.354	−0.363	−0.276	−0.308	0.105	0.045	−0.222	0.216
LU	−0.231	0.026	−0.140	0.077	−0.266	−0.201	0.299	−0.140	−0.210	−0.251	−0.107
NL	−0.165	−0.325	0.059	0.056	0.110	−0.274	0.151	−0.194	−0.207	−0.272	−0.345
PT	−0.297	−0.325	−0.030	−0.387	−0.341	−0.276	−0.277	−0.029	−0.320	−0.254	0.262
SE	−0.293	−0.325	0.361	0.351	−0.254	−0.208	0.209	0.239	0.258	−0.257	−0.154

Table 4. The second eigenvector components.

	GDP	CONS	CAPF	NEXP	CPI	INTR	LABF	UNEMP	GDPH	GDPC	GINI
AU	0.014	−0.155	0.043	−0.030	−0.285	−0.079	0.393	0.268	−0.204	−0.078	0.121
BE	−0.236	−0.042	−0.124	0.279	−0.179	−0.074	−0.026	−0.060	−0.086	0.224	0.051
DE	0.013	−0.141	0.204	−0.110	−0.162	−0.046	0.009	0.273	0.174	0.295	0.339
DK	0.052	0.335	−0.315	−0.433	0.387	0.003	−0.238	0.335	0.276	−0.099	−0.397
ES	0.247	−0.033	0.146	−0.094	−0.234	−0.032	−0.040	−0.197	−0.192	−0.232	−0.083
FI	0.404	0.427	−0.306	−0.423	−0.164	−0.114	0.359	−0.054	0.006	−0.424	−0.385
FR	0.079	0.142	0.146	0.012	−0.149	−0.086	−0.256	−0.012	0.194	0.268	0.190
UK	−0.309	0.039	−0.420	−0.191	0.085	0.314	−0.110	−0.061	−0.011	0.092	0.103
GR	0.238	0.332	0.266	−0.356	0.241	−0.605	−0.399	−0.358	0.340	0.283	−0.083
IE	−0.055	−0.042	−0.075	0.156	−0.343	−0.020	−0.385	−0.196	0.429	−0.108	0.295
IT	−0.323	−0.456	−0.417	0.040	0.051	−0.032	−0.172	0.000	0.306	0.402	0.340
LU	−0.306	0.560	−0.090	−0.423	−0.309	0.471	−0.113	0.424	0.392	0.199	0.300
NL	0.576	−0.033	−0.264	−0.372	−0.448	−0.079	−0.355	0.381	−0.352	−0.186	0.109
PT	0.007	−0.033	−0.438	0.052	−0.094	−0.032	0.129	0.443	0.126	−0.323	−0.241
SE	−0.062	−0.033	0.094	0.156	−0.342	0.519	0.296	0.061	0.286	0.318	−0.372

The eigenvalue spectrum for the empirical correlation matrices is plotted in Figure 2 for all the ME variables. The results are compared with those of a random uncorrelated matrix (RM), having the same size (15×15), constructed by generating random numbers.

In stock market analysis the largest eigenvalue, often called "market effect", is supposed to describe the collective movement of stock prices, because the corresponding eigenvector components have the same sign and approximately the same size. Looking at the first and second eigenvector components (Tabs. 3 and 4) one can easily see that, for the ME correlation matrices, the above interpretation is only partially valid, for $x_1 \equiv$ GDP, $x_2 \equiv$ Consumption, $x_3 \equiv$ Capital Formation, $w_1 \equiv$ GDP/capita and $y_2 \equiv$ Interest Rates. The fluctuations of these indicators seem to reflect a global similarity, as a result of the so-called "globalization trend". The same result was also found in [15] for the first four indicators, by another method, namely measuring the mean statistical distances between countries. The fifth indicator analyzed in [15] was the Net Exports, for which *no occurrence* of this effect was reported – in perfect agreement with the actual results.

3.4 Clustering method and results

The clustering scheme can be next elaborated as follows: firstly, the so-called first order clusters are selected using the bipartite factor graph, i.e. meaning the clusters of countries having at least one connection to the respective variable node. The countries are further partitioned after *the sign* and *the magnitude* of eigenvector components, using Table 4 (for x_1, x_2, x_3, y_2 and w_1) and Table 3 (for the others). For several indicators (x_1, x_2 and z_3) we also selected some groups that can be called second-order clusters, including some countries which are not tied in the factor graph, but have important contributions to the eigenvector structure i.e. large size components. These clusters are written into parentheses in Table 5.

Looking at the development indicators (x_1, x_2, x_3, x_4 and w_1), we find approximately the same clustering scheme as reported in [15] but more extended. There is some agreement with the results reported by Chen in [5] regarding the co-movement between real activity and prices during the period 1992–1997 i.e. the partition of FRA-DEU and ITA into different clusters with respect to the Consumer Price Index fluctuations.

Table 5. The EU-15 clustering. The second column displays the eigenvector whose components are used for building the classification scheme. The groups into parentheses are the second-order clusters.

INDICATOR	EVC	CLUSTERS
GDP	$\mathbf{v_2}$	BEL-GBR-ITA-LUX AUT-DEU-DNK-FRA-PRT (ESP-FIN-NLD)
Final Consumption Expenditure	$\mathbf{v_2}$	AUT-DEU (DNK-FIN-FRA-GRC-LUX)
Gross Capital Formation	$\mathbf{v_2}$	BEL-DNK-FIN-GBR-PRT ESP-FRA
Net Exports	$\mathbf{v_1}$	AUT-DEU-ITA-PRT DNK-FRA-GBR-IRL-SWE
Consumer Price Index	$\mathbf{v_1}$	DEU-ITA-GRC-LUX FIN-FRA-IRL
Rate of Interest	$\mathbf{v_2}$	GBR-LUX-SWE All the others, except for GRC
Labour Force	$\mathbf{v_1}$	AUT-BEL-ESP-GBR-LUX DEU-DNK-ITA-PRT
Unemployment	$\mathbf{v_1}$	AUT-DEU-FRA-GRC-ITA-SWE DNK-ESP-FIN-GBR-IRL-LUX-NLD
GDP per hour worked	$\mathbf{v_1}$	DEU-FRA-LUX-PRT (ESP-GRC-SWE)
GDP per capita	$\mathbf{v_2}$	BEL-DEU-FRA-GRC-ITA-LUX-SWE ESP-FIN-IRL-NLD-PRT
Gini coefficient	$\mathbf{v_1}$	AUT-BEL-DEU-DNK-GBR-LUX-NLD-SWE ESP-FRA-GRC-IRL-ITA-PRT

Moreover there is agreement with the MST constructed in [4] for 1996 i.e. the strong connections BEL-DEU-FRA-LUX, IRE-FIN and ESP-PRT with respect to the GDP/capita.

4 Concluding remarks

Here above we have shown that short and noisy macroeconomic time series can be efficiently investigated by moving a constant size time window with a constant step over the time span of interest. The statistical distances between countries, which are calculated using the linear correlations between the datasets for each time interval, can be used for computing the ultrametrical distance from each country to a virtual introduced one, called "Average". This method, called Moving-Average-Minimal-Length-Path, results in a new set of correlation matrices between country distances to their own mean. The new correlation coefficients describe as well as possible the cross-country similarities between the macroeconomic indicator fluctuations around the average common trend.

The distribution of the absolute values of the correlation coefficients is the uniform distribution. This can be an effect due to the relative small number of data used for computing them (see Tab. 1), but can be also seen as reflecting the diversity resulted from the large number of particular factors underling the time evolution of each ME indicator. As well as in the biological systems, the existence of some common patterns does not exclude the idiosyncratic diversity.

The Bipartite Factor Graph connects in the simplest possible way all the countries by means of corresponding variable nodes assimilated here to the ME indicators. In spite of its simplicity, the method requires an appropriate choosing of the threshold value for the correlation coefficients. One way of evaluating the threshold value can be the t-Student's test of statistical significance, as it has been done in the previous section. We have found the threshold value near 0.83, in a confidence interval of 98% of the correlation coefficients statistical significance.

The Bipartite Factor Graph leads to a clustering scheme in which *all* the countries are involved (a country can only be tied or not tied to the respective variable). For a reliable clustering scheme, more investigation is required, particularly concerning the tied countries. This investigation was performed in the previous section by analyzing the correlation matrix eigensystems.

As compared with the similar investigation of stock prices clustering, there are some similarities, but also important differences. The Random Matrix Theory could only be partially used here, except for those results valid in the limit of infinite matrices: the finite size effects are much stronger here than in the stock market they are. For finding the so-called noise band [7], we had to construct the $N \times N$ ($N = 15$) random matrix having all its rows and columns uncorrelated. Its eigenvalue spectrum was plotted in Figure 2.

The first two eigenvalues (the largest) are far outside the noise band, thus the so called *chance* or *noise correlation* hypothesis can be rejected. Unlike the result obtained for stocks, here the largest eigenvalues does not reflect always a collective mode of the system. The few indicators

for which this propriety holds, are the ones more sensitive to the globalization phenomena.

Finally, as regards the clustering structure, some overlapping with similar results reported in the economic literature was found. However, the clusters composition is most likely a variable from a time span to another. What is important is the *existence* of the clusters themselves, as this hierarchical structure emerged in a period in which the globalization tendencies were strong and the European common policy was generally oriented to extension and cohesion. In spite of all convergent economic policies, the emergence of the clustering structure seems to be inherent to EU-15 system, as well as it is inherent, perhaps, to any human community.

References

1. D. Canning, L.A.N. Amaral, Y. Lee, M. Meyer, H.E. Stanley, Econ. Lett. **60**, 335 (1998)
2. L.A.N. Amaral, P. Gopikrishnan, V. Plerou, H.E. Stanley, Physica A **299**, 127 (2001)
3. S.N. Durlauf, D.T. Quah, in *Handbook of Macroeconomics*, edited by J.B. Taylor, M. Woodford (Elsevier Science, North-Holland, 1999), p. 231
4. J.R. Hill, *Linking Countries and Regions using Chaining Methods and Spanning Trees*, paper presented at the Joint World Bank - OECD Seminar on Purchasing Power Parities, in Washington D.C., 30 Jan.–2 Feb. 2001
5. N. Chen, Eur. Econ. Rev. **48**, 1257 (2004)
6. S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ. Press, Oxford, 2003)
7. J.P. Bouchaud, M. Potters, *Theory of Financial Risk and Derivative Pricing*, 2nd edn. (Cambridge Univ. Press, Cambridge, 2003)
8. T. Di Matteo, T. Aste, R.N. Mantegna, Physica A **339**, 181 (2004)
9. J. Miskiewicz, M. Ausloos, Int. J. Mod. Phys. C **17**, 317 (2006)
10. `http://devdata.worldbank.org/query/default.htm`
11. `http://www.oecd.org/about/0,2337,en_2649_201185_1_1_1_1_1,00.html`
12. A. Pelizzola, J. Phys. A **38**, R309 (2005)
13. V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, T. Guhr, H.E. Stanley, Phys. Rev. E. **65**, 066126 (2002)
14. G.E.P. Box, W.G. Hunter, J.S. Hunter, *Statistics for experimenters: An introduction to design, data analysis, and model building* (Wiley, New York, 1978)
15. M. Gligor, M. Ausloos, e-print `arXiv:physics/0606203`