CORRELATIONS BETWEEN THE MOST DEVELOPED (G7) COUNTRIES. A MOVING AVERAGE WINDOW SIZE OPTIMISATION*

Janusz Miśkiewicz

Institute of Theoretical Physics, Wroclaw University
M. Borna 9, 50-204 Wroclaw, Poland

Marcel Ausloos

SUPRATECS, B5, University of Liège
B-4000 Liège, Euroland

(Received April 20, 2005)

Different distance matrices are defined and applied to look for correlations between the gross domestic product of G7 countries. Results are illustrated through displays obtained from various graph methods. Significant similarities between results are obtained. A procedure for choosing the best distance function is proposed taking into account the size of the window in which correlations are averaged.

PACS numbers: 89.65.Gh, 05.45.Tp, 07.05.Rm

1. Introduction

Various conclusions on correlations depend on the window size in which the averaging technique is performed, *e.g.* one can obtain correlation lengths, Hurst exponents, detrended fluctuation analysis exponents, [1–4] *etc.* There are two main competitive factors: requirements on statistical precision and information loss. If the time window size is large the "quality" of calculated statistical parameters is high (due to the central limit theorem in probability theory [5]) while the information is lost and the results are less sensitive to local features if the window is small. In the case of economy and financial time series it is known that the time series are nonstationary, so not only

* Presented at the First Polish Symposium on Econo- and Sociophysics, Warsaw, Poland, November 19–20, 2004.

the value of considered parameters evolves in time but also their stochastic properties. Therefore, the problem of the time window size is one of the very important factors in such analyses.

There is also another factor for optimising the choice of the time window size: numerical stability requirements. In the case of any numerical calculation every step of the procedure introduces an error. Numerical errors accumulate and in some cases (especially in nonlinear analysis) can quickly influence the results. Therefore, this factor should be taken under consideration when deciding upon the size of the time window. Therefore, the procedure of adjusting the size of the time window is an optimisation problem with two opposite competing factors. In order to find correlations the time window must be moved along the signal, in so doing the subsequent analysis is a "moving average" size window optimisation.

The analysis of the time window optimisation is here-below done on the basis of correlation analyses between G7 countries (France, USA, United Kingdom, Germany[1], Japan, Italy, Canada). The macroeconomy situation is described by their Gross Domestic Product (GDP), since in most countries GDP is considered as an official parameter of the economic situation. GDP is usually defined as a sum of all final goods and services produced in the country, *i.e.* equal to the total consumer, investment and government spending, plus the value of exports, minus the value of imports[2]. Additionally in order to define a reference country an artificial "All" country is constructed. GDP of "All" country is defined as a sum of GDP of all 7 countries. So the GDP increment of "All" can be considered as an averaged level of development.

The GDP values for each of these countries are first normalised to their 1990 value given in US dollars as published by the Groningen Growth and Development Centre on their web page[3]. The data cover the period between 1950 and 2003, *i.e.* 54 points for each country.

2. Distance definition

The equal time t correlation function between A and B is defined as

$$\mathrm{corr}_{(t,T)}(A,B) = \frac{\langle AB\rangle_{(t,T)} - \langle A\rangle_{(t,T)}\langle B\rangle_{(t,T)}}{\sqrt{(\langle A^2\rangle_{(t,T)} - \langle A\rangle^2_{(t,T)})(\langle B^2\rangle_{(t,T)} - \langle B\rangle^2_{(t,T)})}}. \tag{1}$$

The brackets $\langle \ldots \rangle$ denote a mean value over the time window T at time t. In the following, A and B will be the GDP yearly increments of a given

[1] Germany is considered as a one country. To have a record before consolidation the data are constructed as a sum of GDP of both German countries.

[2] http://www.investorwords.com/2153/GDP.html

[3] http://www.ggdc.net/index-dseries.html#top

country, *i.e.*

$$\Delta\mathrm{GDP}(t) = \frac{\mathrm{GDP}(t) - \mathrm{GDP}(t-1)}{\mathrm{GDP}(t-1)}. \tag{2}$$

Since the time series consists of a discrete set of numbers and the time evolution of which is thought to be stochastic the following metrics are used and the results compared.

1. The statistical distance $d(A,B)_{(t,T)}$ is

$$d(A,B)_{(t,T)} = \sqrt{\tfrac{1}{2}(1 - \mathrm{corr}_{(t,T)}(A,B))}, \tag{3}$$

where t and T are the final point and the size of the time window over which an average is taken, respectively.

2. The discrete distance $d_{L_q}(A,B)$ is defined as the sum of absolute values between time series, *i.e.*

$$d_{L_q}(A,B) = \left(\sum_{i=1}^{n} |a_i - b_i|^q\right)^{1/q}, \tag{4}$$

where A, B are time series: $A = (a_1, a_2, \ldots, a_n)$, $B = (b_1, b_2, \ldots, b_n)$.

3. The distribution distance $d_{\mathcal{L}_q}(A,B)$ the distance defined between distribution functions. As an initial step a distribution function should be chosen on the basis of statistical tests, then the considered distribution functions have to be fitted (or appropriate parameters calculated). Since the statistical parameters describing GDP increments are very close to the normal distribution [6] it is hereby assumed that the GDP increments are truly described by the normal distribution. The distance is taken as the metrics of $\mathcal{L}_q$ in Hilbert space [7], *i.e.*

$$d_{\mathcal{L}_q}(A,B) = \left[\int_{-\infty}^{+\infty} |p_A(r) - p_B(r)|^q dr\right]^{1/q}, \tag{5}$$

where $p_A(r)$ and $p_B(r)$ are the appropriate distribution functions fitted to the data.

For the sake of result clarity the analysis is restricted to the case of $q = 1$. The properties of distances measured with $q > 1$ will be discussed elsewhere.

There are different advantages and inconveniences to those distance functions. Eq. (3), a statistical distance, is specially sensitive to observing linear

correlations. The discrete Hilbert space L_q distance, Eq. (4), can be applied to any data and does not require any special properties of the data, thereby seems to be very useful for comparing various sets of data. The distribution distance, Eq. (5), is the most sophisticated one since it requires a knowledge of the data distribution function, but it allows to compare the statistical properties of the data. The main disadvantage of this method is its sensitivity to the size of the data set, since it compares the (assumed) distribution functions.

3. Network definition

In order to obtain the information about the correlations between countries the graph analysis of distance matrices is performed. The described below graphs (LMST, BMLP and UMLP) are built as a function of time and for moving time windows of various sizes (from 5 years (yrs) up to 52 yrs). The size of the time window is constant during the displacement. The mean distance between countries is calculated and averaged over number of generated graphs. (The number of generated graphs is equal to the difference of increments data points (here 53) and the size of the time window.) Finally statistical properties (mean value, standard deviation, skewness and kurtosis) of the twice averaged distance between countries are calculated and discussed. Within the paper the mean value of the distance between countries is understood as a mean value averaged over the number of links on a graph and the number of calculated graphs.

The graph analysis of the distance matrices are based on three types of graph structures:

LMST The Locally Minimal Spanning Tree is the version of a Minimal Spanning Tree (MST) under the assumption that the root of MST is the pair of closest neighbours.

For the sake of simplicity the minimal length path algorithm (MLP) [6] is used. It is a 1-D modification of the MST algorithm. This algorithm emphasises the strongest correlation between entities with the constraint that the item is attached only once to the network. This results in a lack of loops on the "tree". Two different graphs: the unidirectionally growing and the bidirectionally growing minimal length paths (UMLP and BMLP, respectively) are constructed. The UMLP and BMLP algorithms are defined as follows:

UMLP The algorithm begins with choosing an initial point of the chain. Here the initial point is the "All" country. Next the shortest connection between the initial point and the other possible neighbours (in terms of

the distance definition — Eq. (4), (3) or (5) is looked for. The closest possible one is selected and the country attached to the initial point. One searches next for the entity closest to the previously attached one, and repeats the process.

BMLP The algorithm begins with searching for the pair of countries which has the shortest distance between them. Then these countries become the root of a chain. In the next step the closest country for both ends of the chain is searched. Being selected it is attached to the appropriate end. Next a search is made for the closest neighbour of the new ends of the chain. Being selected, the entity is attached, a.s.o.

4. Distance and network analysis

The results are presented for every distance measured, *i.e.* application of statistical distance in Fig. 1, discrete distance in Fig. 2 and finally the distribution distance in Fig. 3.

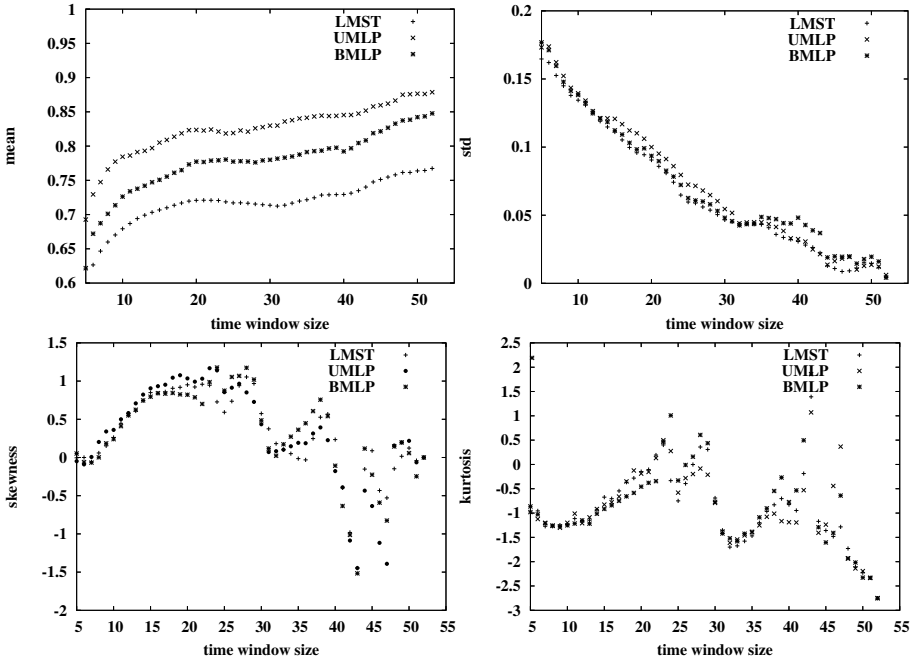

Fig. 1. Statistical analysis of the graph properties obtained by application of the statistical distance. The plots present mean value, standard deviation, skewness and kurtosis of the distance between countries in the case of LMST, UMLP and BMLP averaged over all G7 countries and the considered time interval as a function of the time window size.

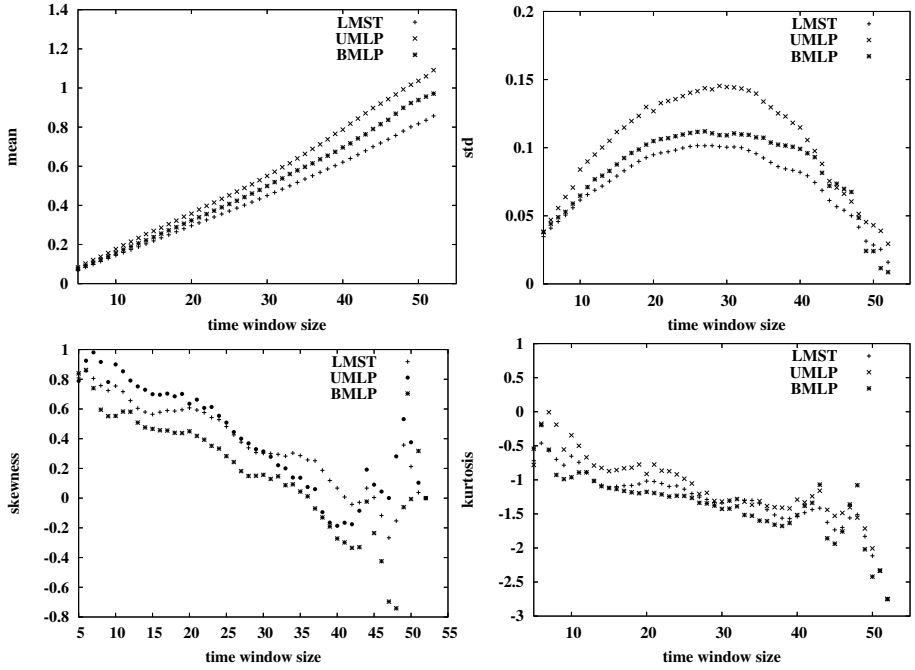

Fig. 2. Statistical analysis of the graph properties obtained by application of the L_1 distance. The plots present mean value, standard deviation, skewness and kurtosis of the distance between countries in the case of LMST, UMLP and BMLP averaged over all G7 countries and the considered time interval as a function of the time window size.

In the case of the statistical distance (Eq. (3)) and the discrete metrics (Eq. (4)) (Fig. 1 and Fig. 2, respectively) the mean distances (understood as it is defined in Section 2) between considered countries increase with the time window size.

The results obtained by application of the statistical distance, Eq. (3), are presented in Fig. 1. The averaged distances between countries are very similar in all considered graph methods and almost parallel to each other. In the case of other basic statistical properties the similarities are even stronger. For other considered statistical parameters *i.e.* standard deviation, skewness and kurtosis plots are almost identical. The standard deviation is decreasing (except few points). This suggests that linear correlations between countries are better seen for longer window size and the co-operation between those countries has a stable form best seen in the long time scale. Of course, there are problems with the size of considered data. While increasing the time window size the amount of data is decreasing which results in significant changes in skewness and kurtosis value for a time window size longer than 40 yrs.

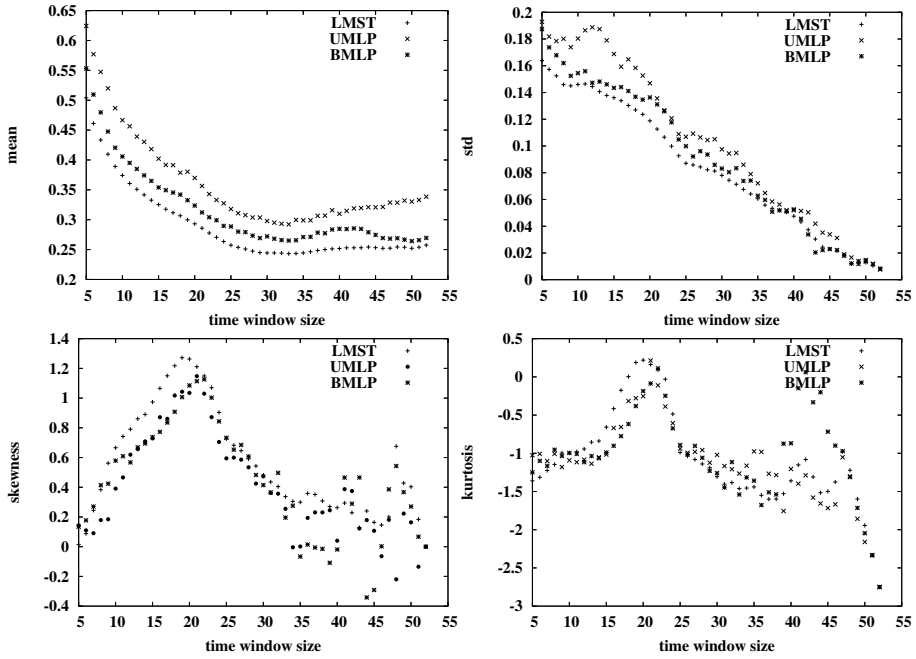

Fig. 3. Statistical analysis of the graph properties obtained by application of the distribution distance. The plots present mean value, standard deviation, skewness and kurtosis of the distance between countries in the case of LMST, UMLP and BMLP averaged over all G7 countries and the considered time interval as a function of the time window size.

In the case of the discrete distance function (Eq. (4)) the mean value of distances between countries increases linearly with the time window size (Fig. 2). This observation is also supported by the value of the linear correlation coefficient (Eq. (1)) which is very close to one (Table I) for all of the investigated structures. This relationship is caused by the properties

TABLE I

Linear correlation coefficients between the time window size and the mean distance between countries in the case of LMST, UMLP and BMLP.

	corr
LMST	0.99854
UMLP	0.99717
BMLP	0.99712

of the applied distance function Eq. (4), which accumulates the differences between the considered time series. Therefore, it is not suggested to use it in order to compare properties of different time window sizes, unless properly normalised, but it may be useful in an analysis of the evolution of a system within a given time window size.

In the case of the statistical distance there is no significant differences between standard deviation, skewness, kurtosis for LMST, UMLP and BMLP (Fig. (2)). The standard deviation has a maximum at 30 yrs time window. It means that within this time window there is the largest spread of distances between considered countries. These results may be the most interesting ones for analysis, because the time evolution may reveal significant changes or a nontrivial evolution of the distances between countries. From an information contents point of view the LMST method gives the highest amount of information because the standard deviation of mean distances is the highest for this graph algorithm.

In the case of the distribution distance defined by Eq. (5), the mean distances between countries are decreasing monotonically up to 30 yrs time window and stabilising or slowly increasing (with respect to the behaviour for the time window size < 30 yrs) for longer window size. The other statistical properties (standard deviation, skewness and kurtosis) are as in the previously considered cases (statistical distance and discrete distance) almost identical for all considered graph methods. The standard deviation is decreasing monotonically within the considered period while skewness and kurtosis have a maximum for the time window size equal to 20 yrs.

For all considered graph methods the LMST algorithm gives the lowest mean distance value (Figs. 1, 2, 3). It is also worth noticing that despite different distance metrics the graph methods give always the same order of functions. The lowest value is taken by the mean distance in the case of LMST, the second is BMLP and the highest value is received by application of the UMLP algorithm. This order is caused by an "optimisation level". In the case of the LMST a new point on the graph can be added at every graph node. In the BMLP on one of both ends and in UMLP only on one end so that the variety of possibilities is decreasing thereby resulting in less densely connected graph. However all functions are very similar to each other. Since chain algorithms are significantly simpler numerically and the received information is similar it may be useful to apply one of the chain algorithms instead of building the MST or LMST tree.

5. Conclusions

The distance analysis consists in two steps. One is the choice of the distance metrics, the second is the graph analysis. It has been shown that the results do not depend significantly on the choice of the graph analysis, but rather on the distance function. The mean distance between G7 countries is increasing with the time window size in the case of the statistical and discrete distance, while the application of the distribution distance results in the opposite behaviour — the mean distance between countries is decreasing with the time window size. In this situation it is extremely important to choose properly the distance function, because there are different advantages to each of the used distance functions. The first method Eq. (3), a statistical distance, is specially sensitive to linear correlations. The discrete Hilbert space L_q distance Eq. (4) can be applied to any data and does not require any special properties of the data so this method seems to be very useful for comparing various sets of data. However, the window size should not change in the analysis since it may influence significantly the results. The third method (Eq. (5)) is the most sophisticated one since it requires a knowledge of the data distribution function, but then points out to similarities among data statistical properties. The main disadvantage of the last method is that it is sensitive to the size of the data set, since it requires fitting a distribution function. It is worth noticing that the results do not depend significantly on the choice of the graph analysis. Of course, results may differ in details, but at the first stage it is useful to apply one of the chain methods (BMLP or UMLP) since they are much simpler then the LMST, especially from the numerical point of view. This may help in the distance metrics choice. The optimal window size may be chosen on the analysis of the statistical properties of the appropriate structure [6].

Various distance metrics have been investigated and new methods based on different distance metrics choice have been proposed in order to investigate the correlations between G7 countries. These methods of mean distance analyses could be also applied to stock market analysis.

This work is partially financially supported by FNRS convention FRFC 2.4590.01. J.M. would like also to thank SUPRATECS for the welcome and hospitality.

REFERENCES

[1] M. Couillard, M. Davison, *Physica A* **348**, 404 (2005).

[2] A. Carbone, G. Castelli, H.E. Stanley, *Physica A* **344**, 267 (2004).

[3] E.A. Maharaj, *Computational Statistics and Data Analysis* **40**, 131 (2002).

[4] Z. Chen, P.Ch. Ivanov, K. Hu, H.E. Stanley, *Phys. Rev.* **E65**, 041107 (2002).

[5] M.A. Goldberg, *An Introduction to Probability Theory with Statistical Applications*, Plenum Press, New York 1984.

[6] M. Ausloos, J. Miskiewicz, An Attempt to Observe Economy Globalization: the Cross Correlation Distance Clustering of the Top 19 GDP Countries, submitted for publication.

[7] K. Maurin, *Analiza*, PWN, Warszawa 1991, in Polish.