FULL-LENGTH REPRINT – differs from the original in layout and no. of pages but not in contents
Chmielewski L., Janowicz M., Ochnio L., Orłowski A.,
Clusterization of indices and assets in the stock market.
In K. Jackowski et al., editors, Lecture Notes in Computer Science vol. 9375
of Lecture Notes in Computer Science, pages 541-550
(Proc. Int. Conf. IDEAL 2015, Wrocław, Poland, 14-16 Oct 2015)
The copyright owner of the paper is Springer. The final publication is available at SpringerLink
via doi:10.1007/978-3-319-24834-9_63.

Clusterization of Indices and Assets in the Stock Market

Leszek J Chmielewski, Maciej Janowicz, Luiza Ochnio, and Arkadiusz Orłowski

Faculty of Applied Informatics and Mathematics (WZIM),
Warsaw University of Life Sciences (SGGW), Poland
ul. Nowoursynowska 159, 02-775 Warsaw, Poland
{leszek_chmielewski,maciej_janowicz,luiza_ochnio,
arkadiusz_orlowski}@sggw.pl
http://www.wzim.sggw.pl

Abstract. K-means clustering algorithm has been used to classify major world stock market indices as well as most important assets in the Warsaw stock exchange (GPW). In addition, to obtain information about mutual connections between indices and stocks, the Granger-causality test has been applied and the Pearson R correlation coefficients have been calculated. It has been found that the three procedures applied provide qualitatively different kind of information about the groups of financial data. Not surprisingly, the major world stock market indices appear to be very strictly interconnects from the point of view of both Granger-causality and correlation. Such connections are less transparent in the case of individual stocks. However, the "cluster leaders" can be identified which leads to the possibility of more efficient trading.

Keywords: technical analysis, clustering, K-means, Granger causality, Pearson correlation

1 Introduction

The technical analysis of the stock market assets [1, 2] belongs to the most controversial branches of analysis of financial data series. This is because that the aim of technical analysis is, basically, no less than the approximate predictions of trends and their corrections in the data which appear as realizations of a *random* process.

On one hand, it has been declared a kind pseudoscience, which, because of the incorrectness of its most important principles, cannot lead to any sustainable increase of returns above the market level [3, 4]. On the other hand, it has been being applied rather blindly without any serious knowledge about the market dynamics. More recent publications, e.g. [5, 6, 7, 8] have lead to considerable revision of the ultra-critical stand of the many experts regarding technical analysis,

As a part of a common knowledge about the stock market dynamics let us notice that time series generated by the prices of stocks are *not* random walks,

K. Jackowski et al. (Eds.): IDEAL 2015, LNCS 9375, pp. 541-550, 2015.
DOI: 10.1007/978-3-319-24834-9_63

and at least the short-time correlations *are* present. Whether they can indeed be exploited with the purpose of maximization of returns is an open question. What we investigate here is meant to be a very small contribution to answer it.

One of the many possible strategies of "beating the market" which are close to the spirit of technical analysis consists of identification of groups of stocks with similar historical behavior of prices. Then, it may happen that within such groups certain "leaders" can be found. Such "leaders" should exhibit behavior which is, to some extent, emulated – with some time delay – by some other members of the group. It is exactly that time delay which might possibly lead to forecasting of the price movement of the whole cluster or at least one of its parts.

As a preliminary but useful exercise we have first performed such grouping, or clusterization, of 24 of the world stock market indices: ALL_ORD, AMEX_MAJ, BOVESPA, B-SHARES, BUENOS, BUX, CAC40, DAX, DJIA, DJTA, DJUA, EOE, FTSE100, HANGSENG, INTERNUS, MEXICIPS, NASDAQ, NIKKEI, RUSSEL, SASESLCT, SMI, SP500, TOPIX, TSE300. Among them there have been both the most important ones like S&P500, DJIA, NIKKEI, HANGSENG and DAX, and those related to smaller markets like Budapest's BUX and Amsterdam's EOE. Then, 30 stocks belonging to the "blue chips" of the Warsaw stock market (WIG30 group) have also been classified.

For the purpose of that classification we have employed: (i) a state-of-the-art clustering algorithm called K-means [9, 10, 11, 12, 13]; (ii) the Granger causality test [14] with lags equal to 1, 2, and 3 trading sessions; (iii) Pearson correlation coefficient [15] with lags; that is, we have computed the Pearson correlation coefficients between shifted closing prices of asset A ($P_{n-k}^{(A)}$) and unshifted closing prices of asset B ($P_n^{(B)}$), where k has been equal to 1, 2, or 3.

The main body of this work is organized as follows. In Section 2 we provide our preliminary results for the 24 stock market indices. Section 3 is devoted to analogous results for the stocks in the Warsaw stock market. Finally, Section 4 comprises some concluding remarks.

2 Classification of the major world stock market indices

From every index a we have obtained the time series $x_n^{(a)}$ in the following way. For each trading session we obtained the sequence of opening (O), maximum (X), minimum (N) and closing (C) values:

$$(O_k^{(a)}, X_k^{(a)}, N_k^{(a)}, C_k^{(a)}).$$

where k enumerates the trading sessions, $k = 1, 2, ..., N$. Then the series has been normalized by subtracting the mean (over k) value of the closing values and dividing by the standard deviation of those values. This gave us the sequence:

$$\left(\dots (\bar{O}_k^{(a)}, \bar{X}_k^{(a)}, \bar{N}_k^{(a)}, \bar{C}_k^{(a)}) \dots \right),$$

where the overbar denotes normalization in the above sense.

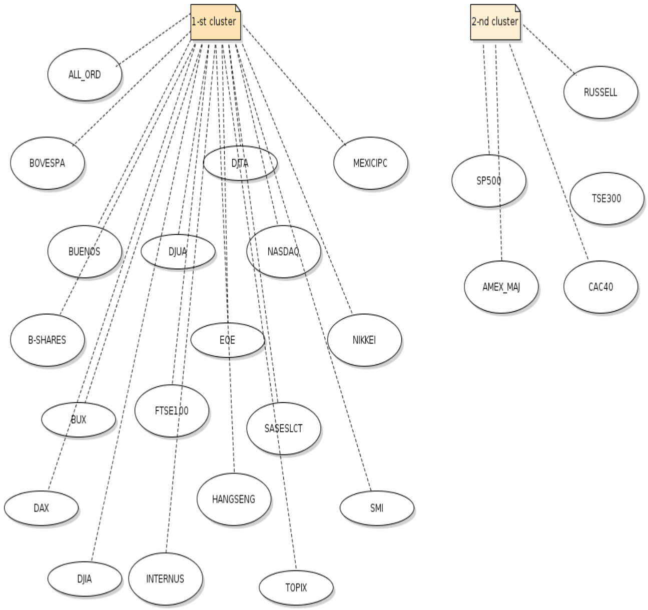

Fig. 1. Two clusters obtained from the K-means algorithm as applied to 24 world stock market indices.

The series $x_n^{(a)}$ is obtained by flattening of the above sequence.

The K-means algorithm which has been applied to series generated in this way requires the number of clusters as an input parameter. We have chosen two, four, and six clusters.

The results are displayed in Figs. 1-3.

We have been somewhat astonished by the fact that the clusters visible in Figs. 1-3 do not appear to follow any pattern associated with geographical locations of the market to which an index corresponds. Also, our hopes to see, say, Euroamerican indices contrasted with the East Asian ones have obviously failed. In fact, we have had some difficulties to provide any reasonable qualitative explanation of the clusterization results. For instance, it is not easy to understand why the French CAC40 index is in the same cluster as Australian ALL ORDINARIES index given completely different methods employed to calculate those indices and rather doubtful resemblance of the French and Australian markets.

To obtain better insight into the connection between the stock market indices we have turned to the Granger causality test. This time we have taken into account only the *normalized closing values* of the indices. It is employed for determining whether one time series is useful in forecasting another. We have obtained all cases in which the null hypothesis that there is *no* causality relationships between two indices has been rejected with p-value lower that 0.01. It has turned out, however, that it is actually quite difficult two find an index which *does not* have any causal relationships (in the Granger sense) with some other.

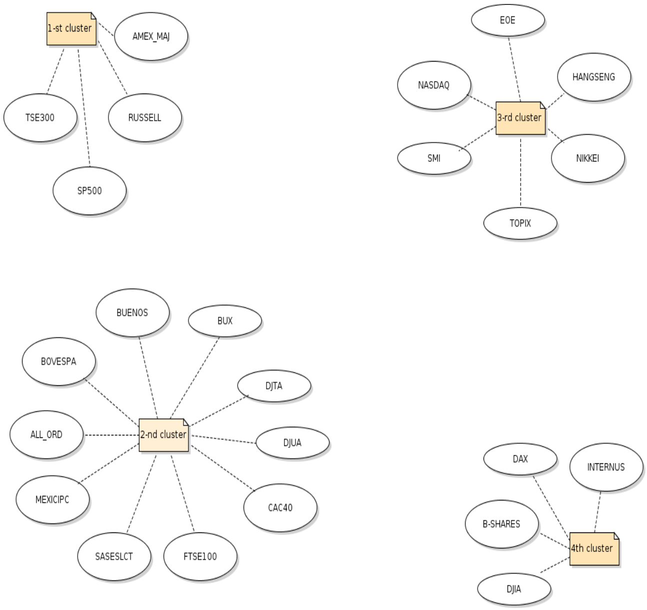

Fig. 2. The same as in Fig. 1 but with four clusters.

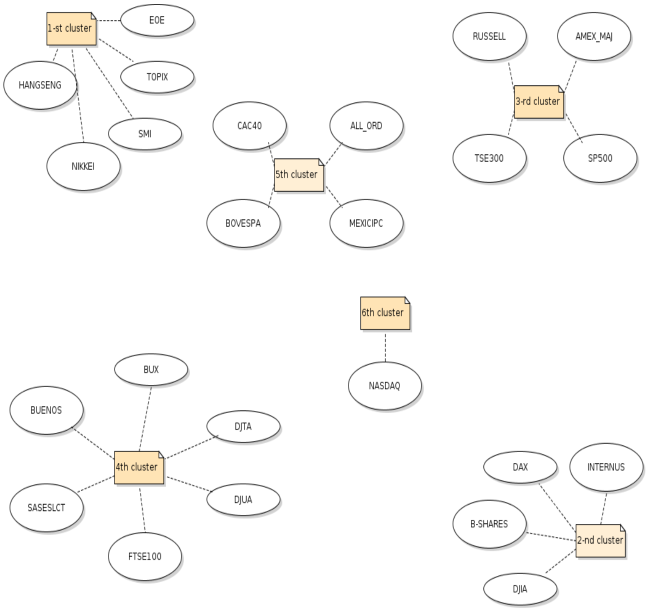

Fig. 3. The same as in Fig. 2 but with six clusters

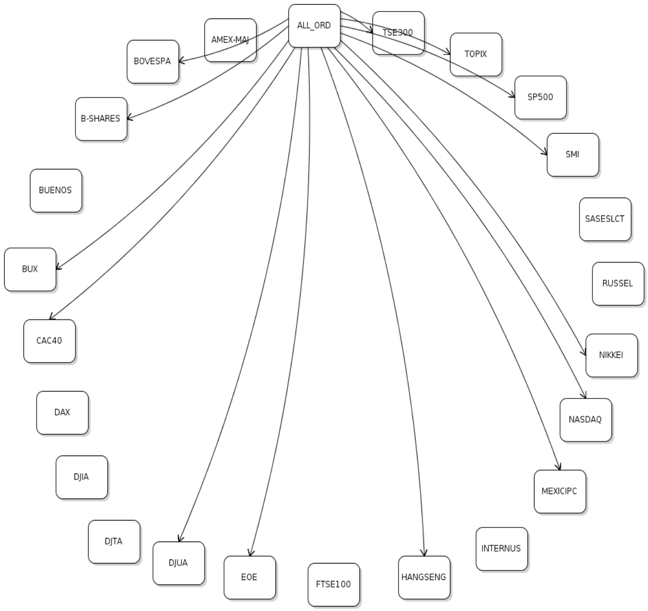

Fig. 4. Example of the Granger-causality relation: the arrows indicate that the Granger causality test rejected the null hypothesis (with p-value < 0.01) that there is no causality relation between the values of (normalized) ALL ORDINARIES index and other stock market indices.

Also, very often the Granger-causality relationships has been two-sided. This is quite intuitive: the world stock markets are interconnected very strongly indeed. As an example, in Fig. 4 we have shown an example of Granger-causality relation with the lag equal to 2 between the Australian ALL ORDINARIES index and other indices. The arrows in Fig. 4 mean that the knowledge of the closing values of ALL ORDINARIES on the day $n - 2$ can be helpful to forecast the closing values of the indices displayed at the end of the arrow on the day n.

Our third classification procedure involved calculations of the Pearson correlation coefficient (PCC) between pairs of closing values of indices a and b with time delay. That is, we define the series $\bar{D}_n^{(a)} = \bar{C}_{n-m}^{(a)}$ and compute $PCC = PCC(m)$ as:

$$PCC(m) = \frac{\mathrm{cov}(\bar{D}^{(a)})\bar{C}^{(b)}}{\sigma_D \sigma C}$$

where cov denotes covariance and σ – standard deviation.

That is, covariances divided by the product of standard deviations have been computed for the series $\bar{C}_{n-m}^{(a)}$ and $\bar{C}_n^{(b)}$ where m has been equal to 1, 2, or 3.

Not surprisingly, we have found very strong correlation among all pairs of indices. In Fig. 5 we have displayed only those arrows which correspond to PCC larger than 0.98 with p-value smaller than 0.001. At last, we have found finally

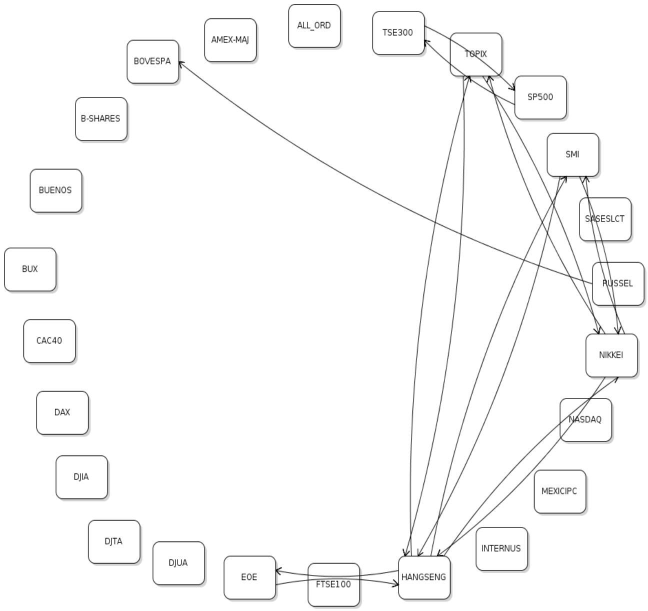

Fig. 5. Example of high correlations among the world indices: the arrows indicate that the Pearson correlation coefficient between the lagged ($m = 3$) and unlagged indices are larger than 0.98 with p-value smaller than 0.01. The beginning of each arrow is at the lagged value of an index.

some connection between the geography and strong correlations. What is perhaps a bit amazing (and amusing) is the strong, one-sided correlation between the RUSSEL (i.e. Russian RTS) and BOVESPA indices.

3 Classification of the most important stocks in Warsaw stock market

We have also employed three procedures described in Section 2 to those Polish stocks of which the WIG30 index has been composed in the beginning of June, 2015: ALIOR, ASSECOPOL, BOGDANKA, BORYSZEW, BZWBK, CCC, CYFRPLSAT, ENEA, ENERGA, EUROCASH, GRUPAAZOTY, GTC, HANDLOWY, INGBSK, JSW, KERNEL, KGHM, LOTOS, LPP, MBANK, ORANGEPL, PEKAO, PGE, PGNIG, PKNORLEN, PKOBP, PZU, SYNTHOS, TAURONPE and TVN.

All the prices have been normalized in the same way as in Section 2.

Firstly, the K-means algorithm has been employed to obtain the clustering for 2, 4, and 6 clusters. The results are shown in Figs. 6-8.

In the case of the WIG30 stocks classification it is again difficult to find any regularities. For instance, it is not true that the financial intitutions appear together in a single cluster or two clusters. Nor can it be said about the

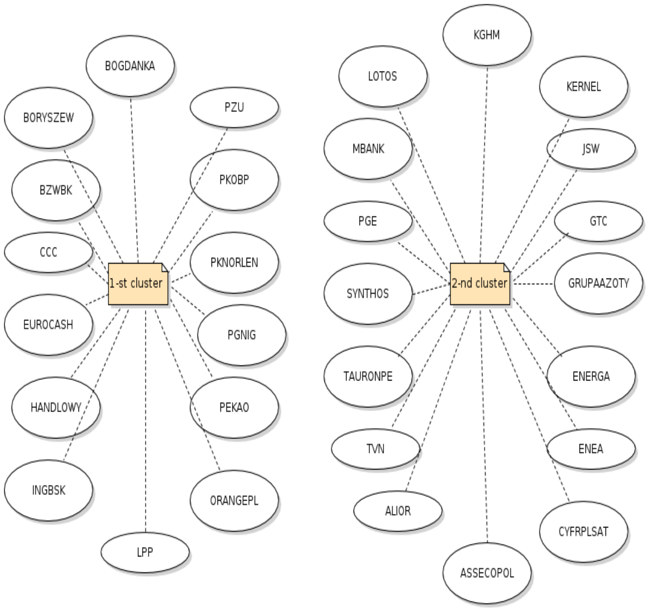

Fig. 6. Two clusters obtained from the K-means algorithm as applied to 30 stocks belonging to WIG30 index.

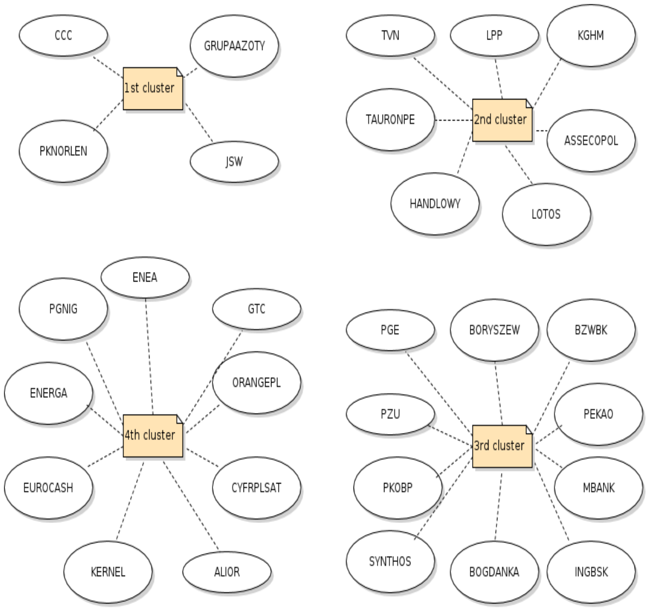

Fig. 7. The same as in Fig. 6 but with four clusters.

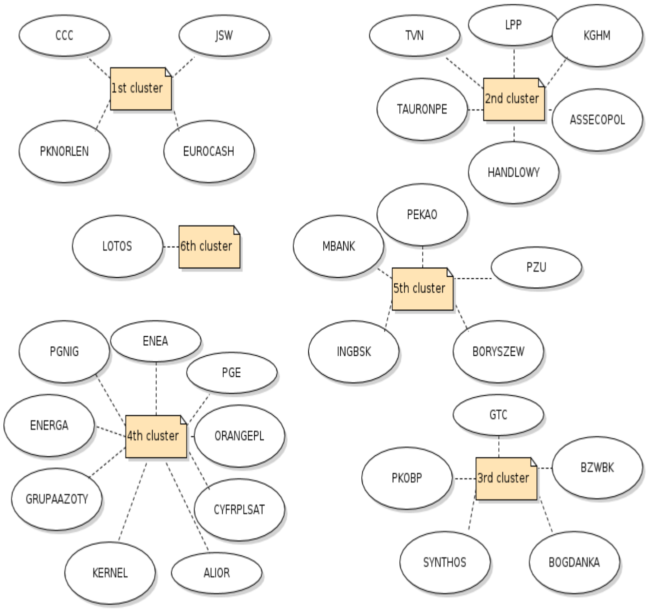

Fig. 8. The same as in Fig. 6 but with six clusters

oil companies or those associated with production and distribution of energy. Our attempts to associate the results of classification with fundamental analysis have also failed. It appears that the clusterization algorithms simply provide knowledge of the type quite different from intuitive correlations.

The results of the Granger causality tests are displayed in Fig. 9. With the help of arrows we have shown the pairs of stocks which have passed the test with p-value smaller than 0.01.

Perhaps the most striking feature of the diagram shown in Fig. 9 is the very special status of the ENERGA stocks which seems to "influenced" (in the sense of Granger causality; there is of course no material innfluence) by eight other stocks. What we believe is also quite interesting is the the fact the prices of INGBSK is in the Granger-causality relation with the prices of two other large banks, PKOBP and ALIOR. We believe that, with sufficient care, the diagram in Fig. 9 can be used to attempt forecasting the behavior of prices of stocks.
The same may be true about the stocks for which strong correlation exists. In Fig. 10 we have displayed pairs for which PCC (computed from retarded ($m = 1$) and unretarded values of the closing prices) has been larger than 0.8 (with significance level 1%).

As we can see, the strong correlations are often two-sided. That is, not only the sequence $\bar{C}_{n-1}^{(a)}$ is strongly correlated with $\bar{C}_n^{(b)}$ for, e.g., $a =$ EUROCASH, and $b =$ PZU, but the opposite is also true. Let us notice here that none of the time series analyzed in this work passes the augmented Dickey-Fuller test for the absence of the unit root, and, most likely, none of them is stationary.

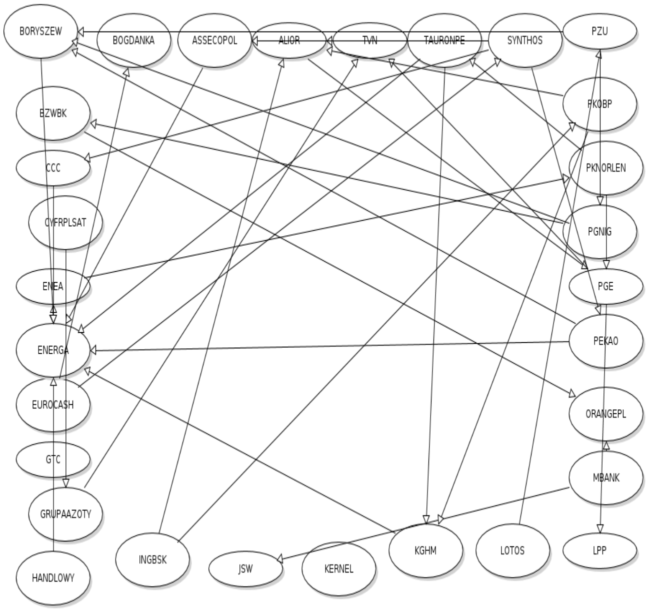

Fig. 9. Example of the Granger-causality relation: the arrows indicate that the Granger causality test rejected the null hypothesis (with p-value < 0.01) that there is no causality relation between the stocks belonging to WIG30 index.

4 Concluding remarks

In this work we have performed classification of a group of important world stock market indices and major stocks in the Warsaw stock market using standard K-means clustering algorithm, Granger causality test, and Pearson correlation coefficients. We have found pairs of indices as well as pairs of stocks in the Warsaw stock market which are either very strongly Granger-causality related or strongly correlated in the Pearson's sense. Let us notice that the correlations obtained with the help of Spearman rank coefficient have not differed qualitatively from those obtained from the Pearson correlation coefficient. We believe that, with sufficient care, the diagrams similar to those obtained here may be used in practice, possibly even to enhance the predictive power of technical indicators for trading purposes.

It is to be noticed, however, that the above results are very preliminary and require careful reexamination. In particular, we have not yet provided any tests for the forecasting powers based on the above classifications. We hope to report results of such improved analysis in a forthcoming publication.

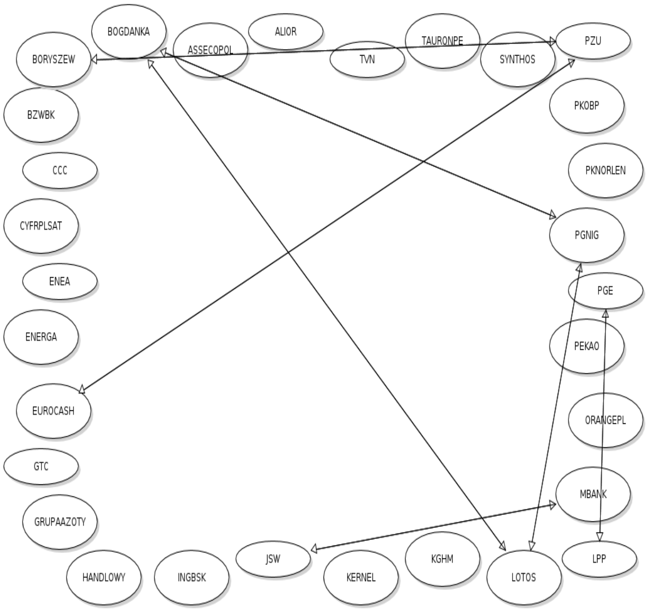

Fig. 10. Example of high correlations among the world indices: the arrows indicate that the Pearson correlation coefficient between the retarded ($m = 3$) and unretarded indices are larger than 0.8 with p-value smaller than 0.01. The arrows start at the stock taken with retarded prices and end unretarded

References

[1] Murphy, J.: Technical Analysis of Financial Markets, New York Institute of Finance, New York (1999).

[2] Kaufman, P.: Trading Systems and Methods, John Wiley and Sons, New York (2013).

[3] Malkiel, B.: A Random Walk Down the Wall Street, Norton, New York (1981).

[4] Fama, E., Blume, M.: E.f. fama and m. blume filter rules and stock-market trading, Journal of Business **39** (1966) 226–241.

[5] Brock, W., Lakonishok, J., LeBaron, B.: Simple technical trading rules and the stochastic properties of stock returns, Journal of Finance **47**(5) (1992) 1731–1764.

[6] Lo, A., MacKinley, A.: Stock market prices do not follow random walks: Evidence from a simple specification test, Review of Financial Studies **1** (1988) 41–66.

[7] Lo, A., MacKinley, A.: A Non-Random Walk down Wall Street, Princeton University Press, Princeton (1999).

[8] Lo, A., Mamaysky, H., Wang, J.: Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation, Journal of Finance **55**(4) (2000) 1705–1765.

[9] MacQueen, J.: Some methods for classification and analysis of multivariate observations, In Cam, M.L., Neyman, J., eds.: Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability. Volume 1., Berkeley, University of California Press (1967) 281–297.

[10] Steinhaus, H.: Sur la division des corps matériels en parties, Bull. Acad. Polon. Sci. **4**(12) (1957) 801–804.

[11] Lloyd, S.: Least square quantization in PCM (1957) Bell Telephone Laboratories Paper,

[12] Forgy, E.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications, Biometrics **21**(3) (1965) 768–769.

[13] Hartigan, J.: Clustering Algorithms, John Wiley and Sons, New York (1975).

[14] Granger, C. W. J.: Investigating Causal Relations by Econometric Models and Cross-spectral Methods, Econometrica **37** (3) (1969) 424–438.

[15] Pearson, K.: Notes on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London **58** (1895) 240–242.

[16] Scikit-learn Community: Scikit-learn – machine learning in python (2015) http://scikit-learn.org.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python, Journal of Machine Learning Research **12** (2011) 2825–2830.