Proceedings of GLOGIFT 08
June 14-16, 2008
Stevens Institute of Technology
Hoboken, NJ, pp. 617-626

CO-INTEGRATION OF STOCK MARKETS USING DATA MINING AND WAVELET ANALYSIS

Rajendra Sahu* and Swatantra K. Gupta**

ABSTRACT

Correlation pattern across the various international stock markets provide the international investors edge on cross border investments. However advancement in IT and decrease in cost of transaction have made International Stock Market independent as evident from international spillovers. Effective international investment and diversification demand an in-depth analysis of major market movements and co-integration among markets. This paper exquisitely tries to identify long-term correlation among major stock markets using data mining and wavelet analysis.

Keywords: *Apriori algorithm, Correlation Coefficient, Haar Wavelet, Non-Stationary Time Series, Spearman Rank Correlation, Unit Root Test.*

Introduction

Advancements in Information and Communication and Technology have made the financial markets more integrated and international in character. The investors today can choose to invest in various financial instruments across financial markets to suit individual risk-return profiles.

There are various methods for association mining to analyze correlation among the various stock markets. However, the real time series data being non-stationary in character are not much suited for direct analysis by conventional statistical methods.

This paper identifies long-term associations in selected stock markets using data mining particularly the modified TAPER algorithm (Xiong et al., 2004) and wavelet analysis for association rule mining.

Past Work on Data Mining

The economists have attempted to explain the stock price movements using macroeconomic fundamentals. As real market data being non-stationary in nature, the analysis is made indirectly by transforming the non-stationary series to a stationary series and adopting classical data analysis on the stationary time series.

Study on association rules and correlation using data mining have been presented by various researchers. Srikant et al., (1995) suggested a generalized, multilevel, quantitative association

* ABV-Indian Institute of Information of Technology & Management, NH#92, Morena Link Road, Gwalior – 474 010, India, Telephone: +91 751 2449804 (o), Fax No. 91 751 2449804
E-mail: rsahu@iiitm.ac.in & sahu_rajendra@hotmail.com

** Room No. 155, Boys Hostel No.1, ABV- Indian Institute of Information of Technology & Management, NH#92, Morena Link Road, Gwalior – 474 010, India, Telephone: +91 9425773181
Fax No. 91 751 2449804, E-mail: swatantra.gupta@iiitm.ac.in & swatantra.iiitm@gmail.com

rules mining, Agrawal and Srikant, (1994) suggested mining sequential patterns and episodes, and Srikant et al., (1997) suggested constraint based Association rule mining using data mining. Tung et al., 2003 used association rule mining on stock market data to analyze the share price movements.

Data Collection and Preparation

The data of past 10 years day's i.e., form July, 1997 to June, 2007 closing price on selected seven stock markets were collected (http://finance.yahoo.com). The markets were selected on the basis of *Market Capitalization, Openness,* and *Proximity* among the market.

The selected markets are: Cac40 of France, Xetra Dax of Germany, FTSE 100 of Great Britain, Sensex of India, Nasdaq of USA, HangSeng of Hong Kong, and Kospi of South Korea. The data on the days on which all the above selected stock markets were operational during the period are only considered for the analysis. In total there are 2176 data points.

A visual inspection of the data series indicates the series to be non-stationary. Analysis for non-stationarity was tested statistically in the next section.

Statistical Analysis for Correlation

A two stepped approach was adopted for determination of correlation among the selected stock markets. First, Unit Root Testing was done to check the non stationarity of the financial time series and then Correlation among markets were found using Spearman Rank Correlation.

Unit Root Testing using KPSS Test

Typically, non-stationarity of the series is determined using Unit Root Test. Standard econometric tools are then used to find out correlation among the international stock market pairs.

A strict stationarity is verified when all elements of a distribution elements are periodically moved and the probability distribution does not change Engle et al., 1987). A time series without a deterministic trend have a moving average representation that can be approximated by a finite moving average autoregressive process. One of the most common and useful techniques for Unit Root Testing is Kwiatkowski-Phillips-Schmidt-Shin Test (usually referred to as KPSS test) (Kwiatkowski et al. 1992) is adopted.

For KPSS testing the series, z(t), is stationary when z(t) = c + u(t), where u(t) is a zero-mean stationary process and "c" is a constant (considered as null hypothesis H_0 for the test). Otherwise, the series z(t) is believed a non-stationary process having a unit roots (considered as alternate hypothesis H_1 for the test). The p-values from KPSS test result for the data series is shown in Table 1.

Table 1: p-values for the Stock Markets

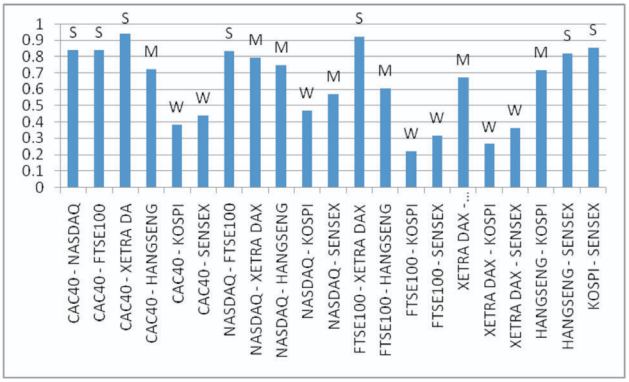
Stock Market	p-value
Cac40	0.6376
Nasdaq	0.6506
FTSE100	1.2928
Xetra Dax	0.7879
Kospi	3.9911
Hangseng	1.7307
Sensex	0.6506

**5% and 10% significance level the critical test value, p-value, is 0.463 and 0.347 respectively.*

As evident from Table 1, all the series have p-value greater than the critical value thus rejecting the null hypothesis. The series are thus non-stationary. But, a first-order differentiation of the series showed the series to be near stationarity. However, presence of Unit Root in the series does not reject relationship among the stock markets. Spearman Rank Correlation Coefficient was determined to identify association among the selected stock markets.

Spearman Rank Correlation Coefficient

Spearman Rank Correlation Method first ranks the data in the series and then determines the correlation coefficient. In financial series the trading strategies are based more on the directional movements, thus the interest lies in determining whether the series goes up or goes down together or not. The stock market pairs are grouped into strong, moderate and weak associations depending on the strength of the associations of movements. The market pairs having correlation coefficient greater than 0.80 are considered as strong associations, and those with correlation coefficient less than 0.50 marked as weak associations. Figure 1 shows the strength of the stock market pairs using the Spearman Correlation Coefficient.



where: *S-Strength, *M-Moderate, *W-Weak

Figure 1: Association among Stock Market Pairs using Spearman Rank Correlation

It is observed from Figure 1 that seven pairs of markets seem to be strongly correlated, seven pairs of markets are moderately correlated and remaining seven pairs of markets are weakly correlated.

Stock Market Correlation using Wavelet Analysis

In stock market data, the most distinguished information is hidden in the frequency content of the signal. The conventional signal transformation methods like the Fourier Transform, FT, etc. provides the frequency-amplitude representation of the signal. However, representation from FT does not reveal whether these frequency components in time domain of the signal exist or not. This information is of paramount importance for non-stationary signals than the stationary signals. As most financial series are non-stationary, the FT is not a suitable approach for data analysis. On the other hand the Wavelet Transform provides the time-frequency representation of the signal simultaneously.

Study on wavelet analysis has been presented by various researchers. Bapna S et al, 2006 suggested a generalized data transformation approach while preserving the original classification, Kantarcioglu et al, 2004 suggested association rule mining on horizontally partitioned data using wavelet analysis and Chen, S.-L et al., 1997 researched upon linear time varying systems by Haar wavelet. Though there are a numerous wavelet analysis algorithms, the Haar wavelet is

the simplest and most suitable technique for the financial time series. (http://www.bearcave.com/misl/misl_tech/wavelets/haar.html).

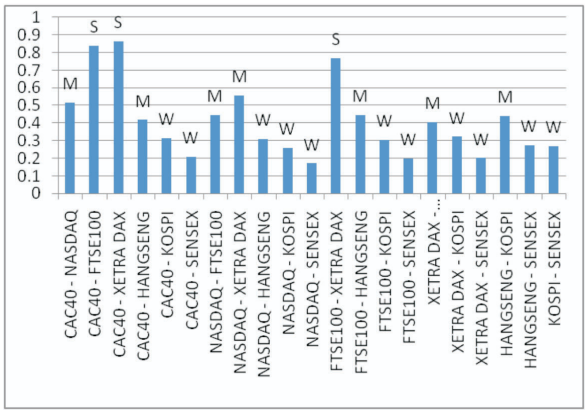
Methodology adopted for Wavelet Analysis

The following are the steps undertaken for the Wavelet Analysis:

Step I: *Determination of Approximation Coefficient and the Detail Coefficient Using Haar Wavelet*: Using Haar Wavelet Transform obtains the two types of coefficients i.e., the *Approximation Coefficient* and the *Detail Coefficient*. The Detail Coefficients provides information regarding the short bursts in the financial data and the Approximation Coefficients denote the average behavior of the markets in the long run.

Step II: *Determination of the Strength of Market Pair Relationship using Pearson Correlation Coefficient:* The Approximation Coefficients obtained in Step-I is used to find the strength of relationship among the pair wise stock markets using Pearson Correlation Coefficient.

Step III: *Categorization of Market Strength:* Pearson Correlation Coefficient obtained in Step-II, is used to categorize the associations into strong, moderate and weak associations. The market pairs having correlation coefficient greater than 0.75 are considered as strong associations, and those with correlation coefficient less than 0.40 marked as weak associations. The strength of the relationship is shown in Figure 2.



where: *S-Strength, *M-Moderate, *W-Weak

Figure 2: Stock Market Pairs Strength using Pearson Correlation Coefficient

It is observed that three pairs of markets seem to be strongly correlated, seven pairs of markets are moderately correlated, and rest eleven pairs of markets are weakly correlated.

Association Rule Mining using modified TAPER Algorithm

Pearson Correlation Coefficient that represents the strength of the relation can also be calculated by the TAPER algorithm (2D Filtering method)(Xiong et al.,2004).The TAPER algorithm computes all-strong-pairs correlation query that returns pairs of high positively correlated items. The all-strong-pairs correlation query problem can be formalized as follows: Given a user-specified minimum correlation threshold theta and a market-basket database with N items and T transactions, an all-strong-pairs correlation query finds all item pairs with correlations above the minimum correlation threshold theta. The 2D Filtering TAPER Algorithm is as follows:

Inputs: S: an item list sorted by item supports in non increasing order.

Theta: a user-specified minimum correlation threshold.

Output

P: the result of all-strong-pairs correlation query.
Variables:
n: the size of itemset S .
A: the item with larger support.
B: the item with smaller support.

2D Filtering (S",theta)
1. n=size (S"), P=Ø (null)
2. startposi = 2
3. for i from 1 to n-1
4. A=S [i]
5. for j from i+1 to n
6. flag=0
7. B=S"[j]
8. if(j e ≥ startposi)
9. upper(ø) =√[(supp(B)) (1-supp(A))/ (supp(A)) (1-supp(B))]
10. if (upper(ø) < theta then
    //reducing computation of upper bound
11. if (j > i+1 || startposi= = n) then
12. startposi = j
13. else
14. startposi = j+1 // break from inner loop
15. P=P U Refine (A,B,theta)
16. If (startposi = = (i+1) && flag= =0) startposi++

// The Refinement Step
Refine (A,B,*theta*)
1. Get the supp (A,B) of item set {A,B}
2. ø =[supp(A,B) - supp(A)supp(B)]/[(supp(A)supp(B)(1-supp(A))(1-supp(B))]
3. if ø< theta then
4. Return Ø (null)
5. else
6. Return { {A,B}, ø}

Methodology Adopted for TAPER Algo

The algorithm is suited for database with binary variables. For computation Pearson's correlation coefficient the following methodology is adopted.

Step I: To calculate the directional movement of the stock markets we consider the rise and fall of market as 1 and 0 respectively and prepare the binary database of all the stock markets.

Step II: Calculate the support for each stock market and stock market pair from the binary database as per the following:

If A and B be two different stock markets, then Support of A is equal to the total number of data points in which item A appears divided by total number of data points in the database. Similarly Support of B can also be calculated. Then, Support of pair A & B is equal to the total number of data points in which both A and B appear together divided by the total number of data points in the database.

Step III: Using the 2D Filtering TAPER Algorithm calculate the strength of associations among stock market pairs. The market association strengths are shown in Table 2.

Table 2: Mkt. Pairs Strength using TAPER Algo

Stock Market Pairs	Association Strength
CAC40 - NASDAQ	Weak
CAC40 - FTSE100	Weak
CAC40 - XETRA DAX	Weak
CAC40 - HANGSENG	Weak
CAC40 - KOSPI	Moderate
CAC40 - SENSEX	Weak
NASDAQ - FTSE100	Moderate
NASDAQ– XETRA DAX	Moderate
NASDAQ – HANGSENG	Moderate
NASDAQ - KOSPI	Weak
NASDAQ - SENSEX	Weak
FTSE100 - XETRA DAX	Strong
FTSE100 – HANGSENG	Strong
FTSE100 - KOSPI	Weak
FTSE100 - SENSEX	Weak
XETRA DAX–HANGSENG	Moderate
XETRA DAX - KOSPI	Weak
XETRA DAX - SENSEX	Weak
HANGSENG - KOSPI	Weak
HANGSENG - SENSEX	Weak
KOSPI - SENSEX	Weak

We obtain two pairs of stock markets to be strongly correlated, five pairs of stock markets are moderately correlated and rest fourteen pairs are weakly correlated.

Comparison of Results from Wavelet and Modified TAPER Algorithm

The comparison of association types from above three methods wavelet analysis modified TAPER Algorithm and Spearman Rank Correlation method is shown in Table 3.

Table 3: Stock Market Pairs Strength using Wavelet Analysis and TAPER Algorithm

Stock Market Pairs	Association Strength by Spearman Correlation	Association Strength by Wavelet Analysis	Association Strength by Modified TAPER Algorithm
CAC40 - NASDAQ	Strong	Moderate	Weak
CAC40 - FTSE100	Strong	Strong	Weak
CAC40 - XETRA DAX	Strong	Strong	Weak
CAC40 - HANGSENG	Moderate	Moderate	Weak
CAC40 - KOSPI	Weak	Weak	Moderate
CAC40 - SENSEX	Weak	Weak	Weak

NASDAQ - FTSE100	Strong	Moderate	Moderate
NASDAQ - XETRA DAX	Moderate	Moderate	Moderate
NASDAQ- HANGSENG	Moderate	Weak	Moderate
NASDAQ - KOSPI	Weak	Weak	Weak
NASDAQ - SENSEX	Moderate	Weak	Weak
FTSE100 - XETRA DAX	Strong	Strong	Strong
FTSE100- HANGSENG	Moderate	Moderate	Strong
FTSE100 - KOSPI	Weak	Weak	Weak
FTSE100 - SENSEX	Weak	Weak	Weak
XETRADAX- HANGSENG	Moderate	Moderate	Moderate
XETRA DAX - KOSPI	Weak	Weak	Weak
XETRA DAX - SENSEX	Weak	Weak	Weak
HANGSENG - KOSPI	Moderate	Moderate	Weak
HANGSENG- SENSEX	Strong	Weak	Weak
KOSPI - SENSEX	Strong	Weak	Weak

It is observed from the table 3 that the wavelet method and the Spearman Correlation techniques provide near similar results. The stock market pairs CAC40 - XETRA DAX, CAC40 - FTSE100 and FTSE100 - XETRA DAX exhibit strong correlation while CAC40 – KOSPI, CAC40 – SENSEX, NASDAQ – KOSPI, FTSE100 – KOSPI, FTSE100 – SENSEX, XETRA DAX – KOSPI and XETRA DAX - SENSEX exhibit weak correlation with both the techniques.

As evident from Table 3 approximately 43% (9 out of 21) market pairs were exhibit similar relationship by all the three above methods. Again, 33% of market pairs (7 out of 21), wavelet analysis and TAPER algorithm give near-similar results with a single level difference of associations i.e., Strong - Moderate , Moderate – Weak. However, five market pairs show the anonymity of relationships by the three methods.

In financial markets, the directional movements are considered more important than the magnitude of the movements. Again, a multiple comparison of markets might provide better results than the market pair wise comparisons. To test for significant market associations among markets Apriori algorithm is more suitable. To data series are divided into two sets: (i) upward movement of data and (ii) downward movement of the data series. Association rule mining using Apriori is done on both the data sets separately.

Association Rule Mining using Apriori Algorithm

Association rule mining using Apriori algorithm tries to discover interesting association relationships among data series. Apriori Algorithm provides the frequent itemsets from those data sets having high confidence and support value than the prescribed minimum confidence and minimum support respectively are considered as strong association. The apriori algorithm for finding frequent itemsets is as follows:

Input: Database, D, of transactions; minimum support threshold, min sup.

Output: L, frequent itemsets in D.

Method

1) L_1 = find frequent 1-itemsets(D);
2) for (k = 2; L_{k-1} !=Φ; k++) {
3) C_k = apriori gen (L_{k-1}, min sup);
4) for each transaction t ε D {
   // scan D for counts
5) Ct= subset (C_k , t);
   // get the subsets of t that are candidates
6) for each candidate c ε Ct
7) c.count++;
8) }
9) Lk = { c ε C_k | c.count ≥ min_sup }
10) }
11) return L = U_k L_k ;

Procedure Apriori_Gen (L_{k-1}:frequent (k-1)-itemsets; min sup: minimum support)

1) for each itemset l_1 ε L_{k-1}
2) for each itemset l_2 ε L_{k-1}
3) if (l_1 [1] = l_2 [1]) ^ (l_1 [2] = l_2 [2]) ^ ….^ (l_1 [k-2] = l2 [k-2]) ^ (l_1 [k-1]< l2 [k-1]) then {
4) c = l1 join l2 ; // join step: generate candidates
5) if has infrequent subset(c, L_{k-1}) then
6) delete c; // prune step: remove unfruitful candidate
7) else add c to C_k ;
8) }
9) return C_k ;

procedure has infrequent subset(c: candidate k-itemset; L_{k-1}: frequent (k-1)-itemsets);
// use prior knowledge

1) for each (k-1)-subset s of c
2) if s ε L_{k-1} then
3) return TRUE;
4) return FALSE;

Confidence is referred as the conditional probability of the itemset evaluated as:

Confidence (A=>B) = support (AUB)/support (A),

where support(AUB) is the number of transactions containing the itemsets A and B, support(A) is the number of transactions containing the itemset A.

Methodology adopted for Association Rule Mining using Apriori Algorithm

The following methodology has been adopted for determining Correlation coefficient:

Step -1: Find the frequent itemsets using Apriori Algorithm.

Step-2: Generate association rules from frequent itemsets as follow:

- for each frequent itemset l, generate all the nonempty subsets of l.

"s $\Longrightarrow$ (l-s)" if $\frac{support(l)}{support(s)}$ e" min_conf

- for every nonempty subset s of l ,output the rule where min_conf is the minimum confidence threshold.

Step-3: Evaluate the correlation coefficients and strength of association rules of stock markets.

Two events A and B are independent if P(A)^P(B) = 1, otherwise A and B are dependent and correlated. This definition can easily be extended to more than two variables. The correlation between A and B can be measured by computing Corr (A, B) $=\frac{P(A \wedge B)}{P(A).P(B)}$

where, Corr (A,B) is the correlation value among A and B; P(A) and P(B) show the probability of A and B respectively(A^B) shows the probability of A and B simultaneous occurrence. If the resulting value of Corr (A, B) is less than 1, then A and B are negatively correlated and each event discourages the occurrence of the other. If the resulting value is greater than 1, then A and B are positively correlated and each event encourages the occurrence of other. If the resulting value is equal to 1, then A and B are independent and there is no correlation between them. Association rules for rising and falling pattern of market obtained by above methodology are tabulated as follow:

Table 4: Association Rules for Rising Markets

If Set	**Then Set**	**Type of Correlation**
Cac40,Xetra Dax	FTSE 100	-ve
Xetra Dax,FTSE 100	Cac40	+ve
Cac40,Nasdaq	Xetra Dax	+ve
Xetra Dax,Nasdaq	Cac40	+ve
Cac40,Nasdaq	FTSE 100	+ve
FTSE 100,Nasdaq	Cac40	+ve

Table 5: Association Rules for Falling Markets

If Set	Then Set	Type of Correlation
Cac40,Xetra Dax	FTSE 100	-ve
Cac40,FTSE 100	Xetra Dax	-ve
Cac40,Nasdaq	Xetra Dax	+ve
Cac40,Xetra Dax, Nasdaq	FTSE 100	+ve
FTSE 100,Nasdaq	Cac40, Xetra Dax	+ve

Following are the observations as evident from Table 4 and Table 5.Two association relations Cac40, Xetra Dax with FTSE100 and Cac40; FTSE100 with Xetra Dax are negatively correlated while Cac40, Nasdaq with Xetra Dax; Xetra Dax, Nasdaq with Cac40; Cac40, Nasdaq with FTSE 100; FTSE 100,Nasdaq with Cac40; Cac40,Xetra Dax, Nasdaq with FTSE 100; FTSE 100, Nasdaq with Cac40, Xetra Dax are positively correlated.

This technique analyzes how the market pairs change their pattern with simultaneous rise or fall of the other markets. We observe the following results from Table 4 and Table 5.

- The simultaneous rise and fall of market Cac40 and Xetra Dax discourage the moving pattern of FTSE 100 market.
- Cac40 and Nasdaq markets simultaneously support the changing pattern of the markets Xetra Dax and FTSE 100 separately.
- The simultaneous directional movement pattern of the markets FTSE100 and Nasdaq encourages the same movement with the Cac40 and Xetra Dax market pattern.

These results obtained above support the findings observed by the table3 and ends the anonymity in the results.

Conclusions

This paper provides wavelet analysis and data mining approaches to establish long-run relationship among major stock markets. It is observed that the wavelet analysis provide near similar results with traditionally adopted statistical techniques like the Spearman Rank Correlation. The modified TAPER algorithm though has an inbuilt approach for correlation coefficients is found inferior to the wavelet analysis.

An analysis of directional movements of markets using Apriori Algorithm indicates that there exists strong correlation among the European and US markets whereas the Asian stock markets does not indicate such integration among themselves.

The present work can be further extended to consider multiple stock market correlations that exhibit a real life global market scenario.

References

- Agrawal R., Imielinski T., and Swami A., 1993, "Mining Association Rules Between Sets of Items in Large Databases," Proc. 1993 ACMSIGMOD International Conference on Management of Data, 207-216
- Agrawal R. and Srikant R., 1994, "Fast Algorithms for Mining Association Rules," Proc. 1994 International Conference on Very Large Data Bases, 487-499
- Bapna S, Gangopadhyay A, 2006, "A Wavelet-Based Approach to Preserve Privacy for Classification Mining", 37(4), 623-642.
- Chen S L, Lai H C, 2006, "Identification of linear time varying systems by Haar wavelet", *International Journal of Systems Science*, 37(9) 619-628.
- Engle, R.F. and Granger, C.W.J., 1987, "Co-Integration and Error Correction Representation, Estimating and Testing", Econometrica, Nº 55, 251-276.
- Kantarcioglu M, Clifton C. 2004. "Privacy-preserving distributed mining of association rules on horizontally partitioned data". IEEE Transactions on Knowledge and Data Engineering, 16, 1026–1037.
- Klemettinen M., Mannila H., Ronkainen P., Toivonen H., and Verkamo A.I., 1994, "Finding Interesting Rules from Large Sets of Discovered Association Rules," Proc. Third International Conference on *Information and Knowledge Management*, 401-408
- Kwiatkowski, D., Phillips, P., Schmidt, P. and Shin, Y., 1992, "Testing the Null of Stationarity against the Alternative of a Unit Root ", *Journal of Econometrics*, 159-178.
- Meo R., Psaila G., and Ceri S., 1996, "A New SQL-Like Operator for Mining Association Rules," Proc. 1996 International Conference on Very Large Data Bases, 122-133
- Srikant R. and Agrawal R., 1995, "Mining Generalized Association Rules," Proc. 1995 International Conference on Very Large Data Bases, 407-419
- Srikant R., Vu Q., and Agrawal R., 1997 "Mining Association Rules with Item Constraints," Proc. Third International Conference on *Knowledge Discovery and Data Mining* (KDD 97), 67-73
- Tung A.K.H., Lu H., Han J., and Feng L., 2003, "Efficient Mining of Intertransaction Association Rules" *IEEE Transactions on Knowledge and Data Engineering*, 15(1), 43 - 56