A comparison of clustering algorithms in categorizing economic events based on the behavior of exchange rates

H A Pathberiya
Department of Statistics
University of Sri Jayewardenepura
Nugegoda, Sri Lanka
hasanthi@sjp.ac.lk

L Liyanage
School of Computing and Mathematics
University of Western Sydney
Locked Bag 1797, Penrith South DC 1797, Australia
L.Liyanage@uws.edu.au

C D Tilakaratne
Department of Statistics
University of Colombo
P O Box 1490, Colombo, Sri Lanka
cdt@stat.cmb.ac.lk

R S Lokupitiya
Department of Statistics
University of Sri Jayewardenepura
Nugegoda, Sri Lanka
lokupitiya@yahoo.com

Abstract—**Cluster analysis is used to identify dissimilar subgroups of objects out of a set of objects based on a combination of rules. In the light of cluster analysis, it is possible to treat dissimilar individuals in an appropriate manner by taking their dissimilarity into consideration. This will be resulted in enhancing the accuracy and efficiency of estimation and prediction models. This study aims to evaluate the performance of different partitioning methods namely, k-means, k-medoids (PAM) and fuzzy and hierarchical methods namely, agglomerative nesting and divisive analysis in grouping the economic events affecting the foreign exchange market. Cluster analysis performed on economic indicators data set depicts the structure of clusters resulted from all algorithms are the same except the single linkage of agglomerative nesting. Poor quality of the clustering structure formed by the single linkage method is confirmed by the lower value of average silhouette width. Comparatively high value of agglomerative coefficient associated with the ward's method reveals the better performance of clustering compared to other linkages. Economic indicators under study are found to be clustered in three groups as performing high, moderate and low impact on the movements of exchange rates. High impact of economic indicators on the exchange rates is reflected by the high volatility at release time and shorter prevailing time of the impact after the release.**

Keywords-clustering algorithms; partitioning methods; hierachichal methods; foreign exchange; volatility; economic indicators

I. Introduction

The main goal of cluster analysis is to derive heterogeneous subgroups of objects from a set of objects based on a combination of rules. It is performed by grouping the most similar objects together. Cluster analysis is used in a wide range of disciplines including medicine, marketing, social sciences, etc. In the light of cluster analysis, it is possible to treat dissimilar individuals in different manner by taking their dissimilarity into consideration. This will be resulted in improving the accuracy of estimation and prediction.

Over the past decades, various clustering algorithms [1][2][3][4] have been developed. Apart from that, some researchers [5][6][7] have taken an immense effort to improve the performance of existing clustering algorithms by means of the speed of calculations. The performance of different clustering algorithms depends on the application used and the conditions under consideration [8]. This fact arise the impossibility of acquiring the best clustering algorithm which can universally be used. Comparison of simple KMeans and farthest first clustering algorithm revealed that the time taken to form clusters in a large data set is longer for KMeans algorithm [9]. Reference [10] recommended partitioning algorithms for large data sets and hierarchical algorithms for small data sets by evaluating the performance of k-means, hierarchical clustering, self-organization map, and expectation maximization algorithms. Some past studies, for example [11][12], found that a better structure of clusters can be achieved by density based clustering algorithms compared to centroid based clustering algorithms even though their accuracy is comparatively low.

This study aims at evaluating the performance of different partitioning methods namely, k-means, k-medoids (PAM) and fuzzy and hierarchical methods namely, agglomerative nesting and divisive analysis in identifying the subgroups of economic events affecting the foreign exchange market. To the extent of our knowledge, researches on clustering economic indicators relating to the foreign exchange market has not been carried out in past. Financial markets are expected to be reacting to

Authors are thankful to Dukascopy Bank SA, Switzerland for their assistance in providing data.

economic incidents take place at times. As the largest and the most liquid financial market, the foreign exchange market tends to demonstrate sudden movements with the release of economic indicators. Degree of change in exchange rate may not be the same for different types of indicators such as, employment, housing, inflation, consumer surveys, etc. Hence, the identification of such dissimilarity of economic indicators with respect to factors for instance volatility of the market, impact prevailing time, etc. would be of great interest. Detection of such grouping would be useful in enhancing the accuracy and efficiency of estimation and prediction models.

Rest of this paper is organized as follows. Section II describes the data used for the study and the methodology followed outlining the clustering algorithms studied in this paper. Experimental results are comprehensively described in section III. Conclusions are summarized in section IV.

II. Methods and Materials

A. Data

For the purpose of categorization, 28 objects representing the US economic indicators were used (Table I). All the indicators except US Unemployment Claims (UC) are released in monthly basis where UC is released in weekly basis. Euro against US Dollar (EUR/USD) percentage return at five minute frequency during the release days of aforementioned indicators throughout the period of 2007 to 2011 was used as the basis for categorization.

TABLE I. List of US Economic Indicators Under Study

No.	Economic Indicator	Abbreviation
1	ADP Non-Farm Employment Change	ADPNFE
2	Advance GDP Price Index	GDPPI
3	Advance GDP	GDP
4	Average Hourly Earnings	AHE
5	Building Permits	BP
6	CB Consumer Confidence	CC
7	Core CPI	CCPI
8	Core Durable Goods Orders	CDGO
9	Core PPI	CPPI
10	Core Retail Sales	CRS
11	CPI	CPI
12	Durable Goods Orders	DGO
13	Existing Home Sales	EHS
14	Housing Starts	HS
15	ISM Manufacturing PMI	MPMI
16	ISM Non-Manufacturing PMI	NMPMI
17	New Home Sales	NHS
18	Non-Farm Employment Change	NFE
19	Pending Home Sales	PHS
20	Philly Fed Manufacturing Index	MI
21	PPI	PPI
22	Prelim UoM Consumer Sentiment	CS
23	Prelim UoM Inflation Expectations	IE
24	Retail Sales	RS
25	TIC Long-Term Purchases	LTP
26	Trade Balance	TB
27	Unemployment Claims	UC
28	Unemployment Rate	UR

B. Data Analysis

At the initial phase of the analysis, average percentage return of exchange rate for each five minute interval of a day was calculated. This was done using the EUR/USD return at all release days for a given economic indicator. Intraday regimes with different means and variances of percentage return were perceived separately for each economic indicator through change point analysis. Based on the result of change point analysis, the prevailing time of the impact before and after the release of the indicator in minutes was defined as follows:

$$PT_b = (RT – CP_b) * 5 \quad (1)$$

$$PT_a = (CP_a – RT) * 5 \quad (2)$$

where PT_b and PT_a represent the prevailing time before and after release, respectively, CP_b and CP_a represent the change point closest to the release time before and after release, respectively, RT represents the release time of the indicator. Volatility at the release time was defined as the standard deviation of percentage returns relative to the regime which includes the release time.

Clusters of economic indicators were derived depending on three variables namely, volatility at release time, prevailing time of the impact before the release and prevailing time of the impact after the release.

1) Change Point Analysis

Pruned exact linear time (PELT) method with Schwarz information criterion (SIC) as the penalty function was employed to detect multiple change points in mean and variance of five minute percentage returns [13][14].

2) Cluster Analysis

Many clustering algorithms have been developed over the year for studying various applications [1][2][3][4]. Each of these algorithms belongs to one of the two categories; partitioning methods or hierarchical methods. In partitioning methods, the desired number of clusters has to be specified by the user whereas it is not necessary in hierarchical methods. This paper evaluates three partitioning methods namely, K-means, K-medoids (PAM) and fuzzy clustering algorithms and two hierarchical methods namely, agglomerative nesting and divisive clustering algorithms.

a) K-means clustering: Partitioning

K-means clustering algorithm is performed by assigning each object to one of K clusters by minimizing the dissimilarity between object and the cluster center [1]. Dissimilarities are defined either by Euclidean (3) or Manhattan (4) distance measures.

$$\text{Euclidean} = \sqrt{\Sigma_i} (x_i - m_i)^2 \quad (3)$$

$$\text{Manhattan} = \Sigma_i |x_i - m_i| \quad (4)$$

where
x_i = i^{th} attribute of an object, and
m_i = center relevant to i^{th} attribute.

Initial cluster centers are selected randomly from a set of objects. At the end of each iteration, cluster centers are recalculated using the mean of all objects in the cluster. This procedure is repeated until the centers do not change.

b) Partitioning Around Medoids (PAM): Partitioning

PAM is an upgraded version of K-means algorithm. PAM differs from K-means due to the method it uses to calculate cluster centers. Instead of using mean it uses median to represent cluster centers. Use of median avoids the effect from outliers. PAM is more robust and it is found that the sub groups resulted from PAM is more natural compared to that of K-means algorithm [15]. However, it becomes computationally complex for large sample sizes and large number of clusters [16].

c) Fuzzy clustering: Partitioning

In K-means or PAM clustering methods, each object belongs exactly to only one cluster. In these methods, even though the distances between an object to each cluster center is very much close, the object is assigned to the cluster which has the lower distance. Fuzzy clustering [2] provides additional information in terms of a membership value which measures the closeness of the object to the cluster center. Hence, fuzzy clustering enables to identify objects locate in margins of different clusters.

In fuzzy clustering algorithm, membership (u_{ij}) for each data point and center for each cluster (C_j) are defined by (5) and (6), respectively at the end of each iteration where the initial cluster centers are selected randomly.

$$u_{ij} = 1/ \Sigma_k (d_{ij} / d_{ik})^{(2/m-1)} \quad (5)$$

$$C_j = \Sigma_i (u_{ij})^m x_i / \Sigma_i (u_{ij})^m \quad (6)$$

where

d_{ij} = Euclidean or Manhattan distance between i^th data point and j^th cluster center,

m = fuzziness exponent, and

x_i = i^th attribute of an object.

This procedure is repeated until the cluster centers remain unchanged.

d) Agglomerative Nesting: Hierachichal

Agglomerative nesting is referred as bottom up algorithm. It initially considers all the objects as separate clusters. Afterwards at each step a pair of clusters with the smallest dissimilarity is merged to form a new cluster. This procedure is repeated until all the objects belong to one large cluster. Dissimilarity between any two objects are calculated by Euclidean or Manhattan distance measures.

$$\text{Euclidean} = \sqrt{\Sigma_i (x_{ij} - x_{ik})^2} \quad (7)$$

$$\text{Manhattan} = \Sigma_i |x_{ij} - x_{ik}| \quad (8)$$

where

x_{ij} = i^th attribute of j^th object.

To calculate the dissimilarity (d) between clusters A and B where at least one of A and B consist of more than one object, one of the following methods are used.

Single linkage:

$$d = \min d_{ij}; \text{ where } i \in A \text{ and } j \in B \quad (9)$$

Complete linkage:

$$d = \max d_{ij}; \text{ where } i \in A \text{ and } j \in B \quad (10)$$

Group average method:

$$d = \text{mean } (d_{ij}); \text{ where } i \in A \text{ and } j \in B \quad (11)$$

Ward's method can also be used to form clusters in agglomerative nesting algorithm. Instead of using two objects to calculate the dissimilarity, ward's method uses one object and the mean of the cluster to be formed. Hence, Euclidean and Manhattan distance measures are defined as,

$$\text{Euclidean} = \sqrt{\Sigma_i \Sigma_j (y_{ij} - \bar{y}_{i.})^2} \quad (12)$$

$$\text{Manhattan} = \Sigma_i \Sigma_j |y_{ij} - \bar{y}_{i.}| \quad (13)$$

where

y_{ij} = i^th attribute of j^th object, and

$\bar{y}_{i.}$ = mean of i^th attribute,

$j \in A \cup B$.

e) Divisive Analysis: Hierachichal

Compared to the agglomerative nesting, divisive analysis builds the hierarchy of clusters in reverse order. Hence, it is referred as top down algorithm. This algorithm initially considers all objects as a single cluster and then split it into two clusters. This procedure is repeated until clusters consist of a single object are found. Assignment of objects into clusters is done according to the dissimilarities between objects measured by the Euclidean or Manhattan distance measures. Algorithm consists of the following steps,

Step 1: consider all objects as a one cluster

Step 2: select the object with the highest average dissimilarity to all the other objects within the cluster. Assign this object to a new cluster "splinter group".

Step 3: for each object *i* in the "main group" calculate,

$$ds_i = \text{average } (d_{ij}); \text{ where } j \in \text{splinter group} \quad (14)$$

$$dm_i = \text{average } (d_{ij}); \text{ where } j \in \text{main group} \quad (15)$$

$$D_i = dm_i - ds_i \quad (16)$$

Step 4: assign the object with the maximum D_i to the splinter group.

Step 5: repeat step 3 and 4 until all D_i s are negative.

Step 6: select the cluster with the largest dissimilarity between any pair of objects within the cluster and repeat steps 2 to 5 for this cluster.

Step 7: repeat step 6 until clusters with single object are formed.

3) Comparison criteria

Following criteria were used to determine the optimal number of clusters in partitioning methods.

a) Sum of squared error (SSE)

Selection of the optimal number of clusters in K-means algorithm was done by comparing the SSE for different number of clusters. "SSE is defined as the sum of the squared distance between each member of a cluster and its cluster centroid" [17]. The point in the scree plot of SSE has a bend (elbow) was taken as the optimal number of clusters for the algorithm.

b) Average silhouette width(ASW)

The number of clusters with the highest ASW was selected as the optimal number of clusters in PAM and Fuzzy clustering algorithms. Silhouette width for any object i is defined as,

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\} \tag{17}$$

where

$a(i)$ = average dissimilarity of i to all other objects within the same cluster, and

$b(i)$ = lowest average dissimilarity of i to any other cluster

$-1 \leq s(i) \leq 1$.

Values of $s(i)$ close to 1 implies object i is well classified whereas the values close to -1 implies object i is badly classified. Moreover, this criterion is used to compare the results of different clustering algorithms as well.

c) Agglomerative coefficient(AC)

AC is considered when comparing the performance of different linkages used in agglomerative nesting. Higher values of AC reflect better structure of clustering.

$$d(i) = d_1(i) / d_p(i) \tag{18}$$

$$AC = \Sigma_i [1- d(i)] / n \tag{19}$$

where

$d_1(i)$ = dissimilarity of the object i to the first cluster it is merged with, and

$d_p(i)$ = dissimilarity of the merger in the last step of the algorithm.

III. Results

Change point analysis in mean revealed that the mean of percentage returns remains the same throughout the day even though there are releases of economic indicators. However, multiple change points in intraday variance were detected for each economic indicator. Fig. 1 illustrates that the volatility at the release time of GDP/ GDPPI and NFE/ AHE/ UR is the highest. It is almost the same for all the other economic indicators except NHS which exhibits combative high value.

Prevailing time before the release of GDP/ GDPPI is the shortest and it is about five minutes. Moreover, shorter prevailing times before the release are exhibited at the release of NFE/ AHE/ UR, PPI/ CPPI, UC, NHS and TB. Prevailing times after the release of GDP/ GDPPI and NHS are the shortest and they are about five minutes. In addition to that, after the release of NFE/ AHE/ UR, it shows a comparatively shorter prevailing time. High volatility at the release and shorter prevailing times related to GDP/ GDPPI and NFE/ AHE/ UR imply that the exchange rates move faster for very short period of time around the release of these economic indicators.

Fig. 2 shows the SSE calculated for different number of clusters resulted from McQueen algorithm of K-means clustering. It suggested that having three clusters as the optimal number of clusters since the scree plot illustrates a bend at 3.

K-means, PAM and Fuzzy partitioning methods concluded that the maximum ASW can be observed by categorizing the objects into three clusters. Clusters resulted from all three partitioning methods are the same (Fig. 3).

In addition to the grouping done by partitioning methods, fuzzy algorithm provides fuzzy memberships indicating how strong each object belongs to the corresponding cluster. According to Table II even though PHS belongs to cluster 3 the strength of its attachment to this cluster is only 49%. It can also be considered as belonging to cluster 1 with strength 31%. This implies that PHS is in the margins of cluster 1 and cluster 3.

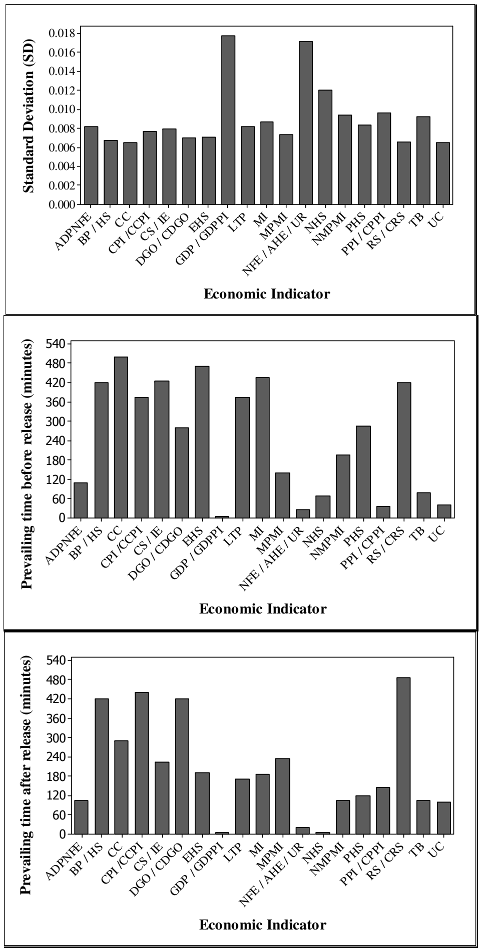


Figure 1. volatility around the release time and prevailing time of the impact before (PT_b) and after(PT_a) the release

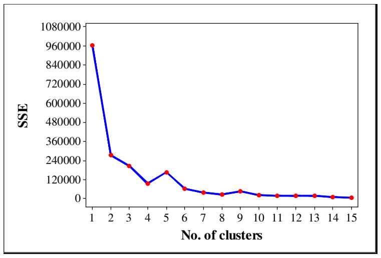
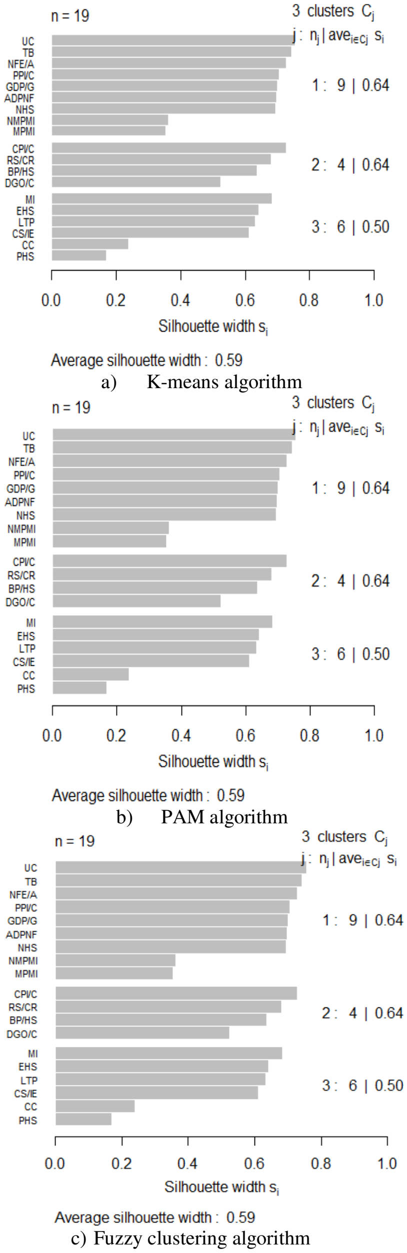


Figure 2. Scree plot of SSE at different number of clusters



a) K-means algorithm

b) PAM algorithm

c) Fuzzy clustering algorithm

Figure 3. Silhouette plots of K-means, PAM and Fuzzy algorithms

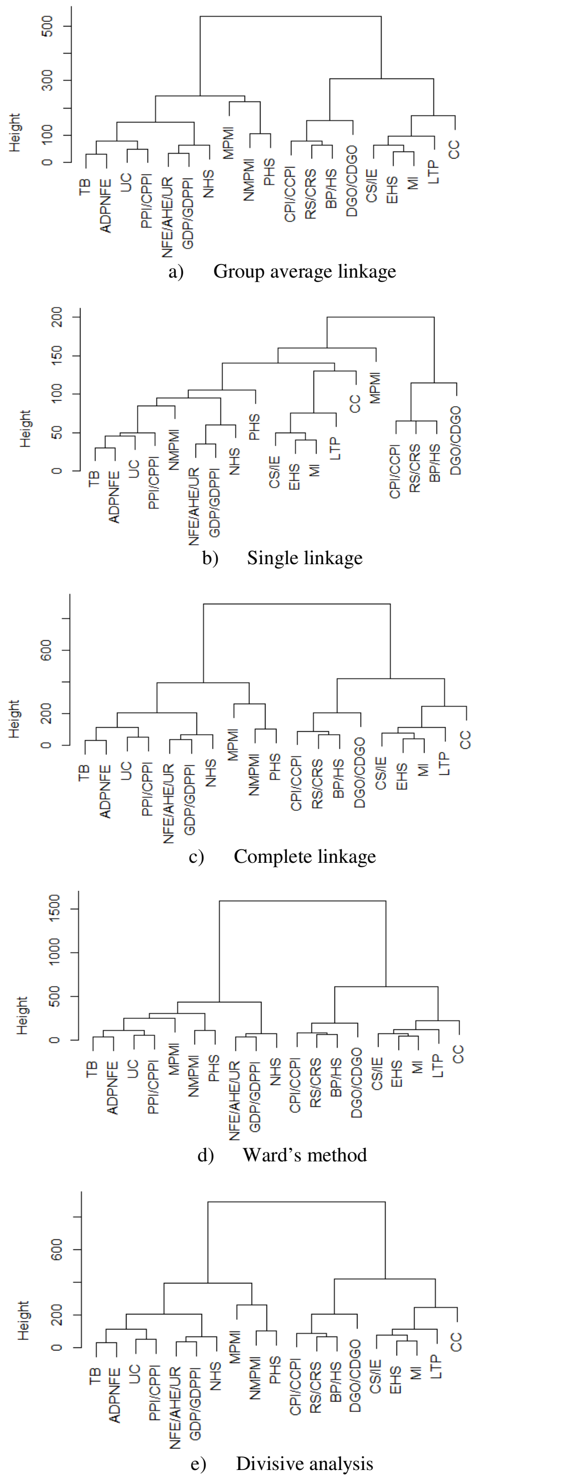
The structure of clusters resulted from agglomerative nesting and divisive analysis is almost the same except the single linkage method of agglomerative nesting (Fig. 4) Single linkage method of agglomerative nesting yields maximum ASW of 0.4 with four clusters. Compared to ASW of 0.6 which is related to the optimal number of clusters decided by other linkages and divisive analysis, the ASW of single linkage method is comparatively low. Moreover, the optimal number of clusters which maximize the ASW by other linkages and divisive analysis is three. These results imply that the single linkage of agglomerative nesting is performing quite different manner compared to other linkages. Comparison of AC for different linkages revealed that the ward's method with the highest AC of 0.95 outperform the other linkages.

There is no difference in the clusters resulted from three partitioning methods and two hierarchical methods except the single linkage of agglomerative nesting algorithm (Table III). Single linkage exhibited lower ASW of 0.4 where all the other algorithms show ASW of 0.6.

Clusters recognized by the clustering algorithms are illustrated in Fig. 5. Economic indicators which belong to cluster 1, exhibit higher volatility at the release time and comparatively lower impact prevailing time both before and after the release. This fact indicates that this cluster of economic indicators cause high impact on the movements of exchange rates.

TABLE II. FUZZY MEMBERSHIPS OF OBJECTS

Economic Indicator	Cluster 1	Cluster 2	Cluster 3
TB	87	7	5
ADPNFE	82	11	8
MPMI	48	27	24
NMPMI	55	28	17
UC	88	7	6
NFE / AHE / UR	83	10	7
PPI / CPPI	79	11	10
CPI /CCPI	4	8	89
RS / CRS	6	12	81
CS / IE	6	83	11
LTP	10	77	13
BP / HS	5	11	84
DGO / CDGO	13	21	65
GDP / GDPPI	79	12	9
EHS	7	81	12
NHS	80	11	8
MI	5	88	8
CC	11	55	34
PHS	31	49	20



a) Group average linkage

b) Single linkage

c) Complete linkage

d) Ward's method

e) Divisive analysis

Figure 4. Dendrograms resulted from agglomerative nesting and divisive analysis

TABLE III. Summary of Comparison of Clustering Algorithms

Algorithm	Linkage	SC	AC	No. of Culsters
KMEANS		0.6		3
PAM		0.6		3
FAZZY		0.6		3
AG. NES.	average	0.6	0.85	3
AG. NES.	single	0.4	0.66	4
AG. NES	complete	0.6	0.9	3
AG. NES.	Ward's	0.6	0.95	3
DI. ANA.		0.6		3

Moreover, the volatility of cluster 2 and cluster 3 are approximately the same and lower than that of cluster 1. However, the prevailing time of this lower volatility is longer for cluster 3 compared to cluster 2. Lower impact of this cluster on the movements of exchange rates is pointed out by this fact.

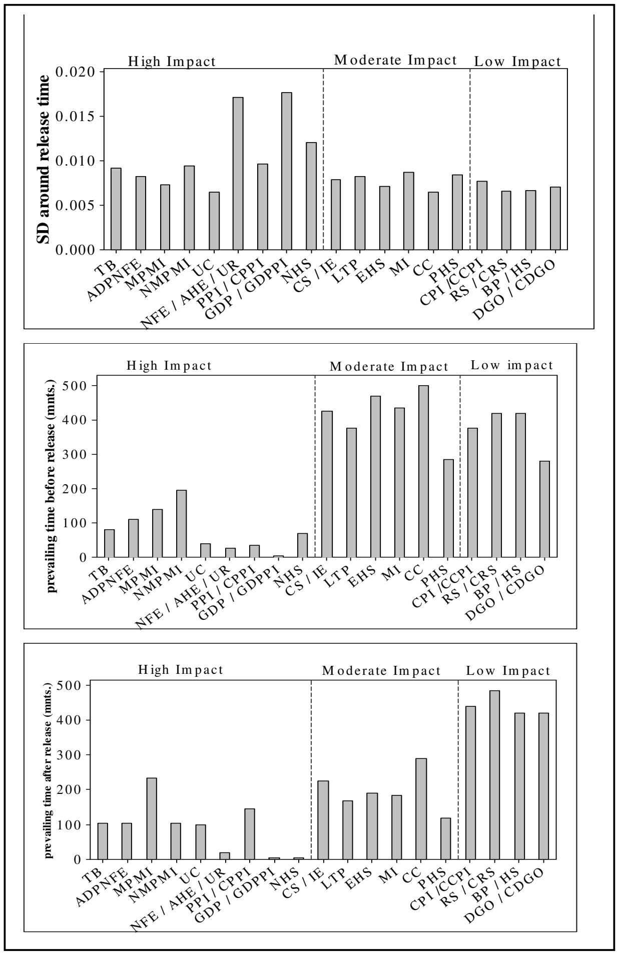


Figure 5. Clusters identified by clustering algorithms

IV. Conclusions

Cluster analysis performed on economic indicators data set depicts the structure of clusters resulted from all algorithms are the same except the single linkage of agglomerative nesting. Clusters are validated using ASW. Value of ASW is 0.4 for the single linkage of agglomerative nesting and it is 0.6 for all the other algorithms. This confirmed that the comparatively poor performance of single linkage of agglomerative nesting. Fuzzy memberships provide additional information to identify the objects located at margins of several clusters where this kind of identification is impossible with K-means or PAM algorithms. Ward's method of agglomerative nesting shows the highest AC reflecting a better structure in clustering compared to other linkages. Internal hierarchy of objects which is invisible in partitioning methods is also visible in hierarchical methods. Hence, it can be considered in further researches to improve the accuracy of results.

Study revealed that the indicators can be categorized into three clusters based on the volatility at the release time, prevailing time before and prevailing time after the release. Cluster 1 comprise of economic indicators with highest volatility at release time and shortest prevailing time both before and after the release. Cluster 2 comprise of economic indicators with lower volatility at release time and moderate prevailing time after the release. Cluster 3 comprise of economic indicators with lower volatility at release time and longest prevailing time after the release. This indicates that during the study period the economic indicators belonging to cluster 1 and cluster 3 caused high and low impact on the movements of exchange rates, respectively.

References

[1] J. MacQueen, Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability vol. 1(14), pp. 281-297, June 1967.

[2] J. C. Bezdek, R. Ehrlich and W. Full, FCM: The fuzzy c-means clustering algorithm, Computers & Geosciences, vol. 10(2), pp. 191-203, 1984.

[3] Jr. Ward and H. Joe, Hierarchical grouping to optimize an objective function, Journal of the American statistical association, 58(301), pp. 236-244, 1963.

[4] L. Kaufman and P. J. Rousseeuw, Finding groups in data: An introduction to cluster analysis, New York: Wiley, 1990.

[5] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, An efficient enhanced k-means clustering algorithm, Journal of Zhejiang University SCIENCE A, 7(10), pp. 1626-1633, 2006.

[6] C. Böhm, K. Railing, H. P. Kriegel and P. Kröger, Density connected clustering with local subspace preferences, Data Mining, 2004. ICDM'04, Fourth IEEE International Conference, pp. 27-34, November 2004.

[7] D. Jiang, J. Pei and A. Zhang, DHC: a density-based hierarchical clustering method for time series gene expression data, Bioinformatics and Bioengineering, 2003, Proceedings of Third IEEE Symposium, pp. 393-400, March 2003.

[8] R. Xu and D. Wunsch, Survey of clustering algorithms, Neural Networks, IEEE Transactions, vol. 16(3), pp. 645-678, 2005.

[9] S. Revathi and T. Nalini, Performance comparison of various clustering algorithm, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3(2), pp. 67-72, 2013.

[10] O. A. Abbas, Comparisons between data clustering algorithms, Int. Arab J. Inf. Technol., vol. 5(3), pp. 320-325, 2008.

[11] J. Erman, M. Arlitt and A. Mahanti, Traffic classification using clustering algorithms, Proceedings of the 2006 SIGCOMM workshop on Mining network data, pp. 281-286, September 2006.

[12] M. Ester, H. P. Kriegel, J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, In Kdd, vol. 96(34), pp. 226-231, August 1996.

[13] R. Killick, P. Fearnhead and I. A. Eckley, Optimal detection of changepoints with a linear computational cost, Journal of the American Statistical Association, vol. 107(500), pp. 1590-1598, 2012.

[14] G. Schwarz, Estimating the dimension of a model, The annals of statistics, vol. 6(2), pp. 461-464, 1978.

[15] A. Struyf, M. Hubert and P. Rousseeuw, Clustering in an object-oriented environment, Journal of Statistical Software, vol. 1(4), pp. 1-30, 1997.

[16] P. Andritsos, Data clustering techniques, Rapport technique, University of Toronto, Department of Computer Science, 2002.

[17] Peeples and A. Matthew, (2011), R Script for K-Means Cluster Analysis. [online]. available at: http://www.mattpeeples.net/kmeans.html, accessed on: August 7th 2015.