Volume: 3, Issue: 2
Page: 146-166
2019

ISSN 2520-4750 (Online) & ISSN 2521-3040 (Print)

International Journal of Science and Business

Comparing Clustering Algorithms using Financial Time-series data

Duangrux Tangsirisakul

Abstract:

Data clustering is one of the most popular unsupervised machine learning approaches. Clustering data can help identify the pattern of what seems to be similar data and leads to the best solution for all commercial problems. For example, taxi booking application, customer's data can be clustered to match supply with demand, to detect fraud pattern of an e-commerce transaction or clustering customers in dating application, etc. In order to carry out the best calculation of clustering certain requirement is needed in each method and approach such as the basic assumption of data. When analyzing data with a wrong assumption, it results in low-quality outcomes. So we would like to study and compare this type of data in an in-depth manner. Time-series analysis is used in many future prediction tasks based on previously observed values, mixing cluster analysis and time-series data to serve the initial purpose that researcher would like to share to the public for better understanding of the clustering, researcher would also like following researchers to refer to this work and develop this theory and apply in wider issues in future. In this paper, the focus is on comparing time-series clustering algorithm with financial time-series data, which is common data such as cryptocurrency, exchange rate currency, the Shanghai Stock Exchange (SSE50), and the stock exchange of Thailand 50 (SET50). The paper introduces the importance of data mining, machine learning, and time-series clustering and some related methods, which lays a theoretical foundation for the formal research of this paper. By analyzing the structure of time-series clustering, that consists of several parts, including distance measurement, time-series prototype, a clustering algorithm, and clustering evaluation. From research result, the hierarchical algorithm is the most efficient algorithm for unequal length of cryptocurrency series and SSE 50. In another hand, the partitional algorithm is the most efficient for an equal length of exchange rate currency and SET 50.

IJSB
Accepted 27 March 2019
Published 31 March 2019
DOI: 10.5281/zenodo.2617341

Keywords: time-series clustering, machine learning, dynamic time warping, crypto-currency

About Author

Duangrux Tangsirisakul, Department of Mathematics, China University of Mining and Technology, Jiangsu, China.

1. Introduction

Time-series data is a kind of data that was collected from the data points. It is a continuous sequence of time such as daily stock data, daily currency exchanged rate, daily temperature data or cancer growth rate etc. Current time-series data plays an important in research of various disciplines, such as bioinformatics, robotics, medicine, chemistry, gesture recognition, speech recognition, tracking, finance, biometrics, astronomy, manufacturing, etc. Data mining of time-series data is an interesting topic. To analyze and drilling the time-series data of the relationships, models or insights, which are hidden, useful information based on the principles of mathematics, statistics, database, recognition and learning of the machine (Machine Learning such as association rule, classification, prediction, clustering, anomaly detection, and visualization). This paper raises the clustering analysis to experiment by comparing 3 scenarios of clustering algorithm with various time-series data (crypto-currency, exchange rate currency, the Shanghai Stock Exchange and the Stock Exchange of Thailand 50) using the principle of distance measurement which is Dynamic Time Warping techniques (DTW), however there are still problems caused by DTW techniques calculation that is very dynamic, so it would take time to calculate and could not speed up easily. After finished data clustering, clustering evaluation is used to deciding clustering algorithm scenarios quality and indicated the dataset that is fit and suitable to which clustering algorithm scenario.

2. Related work

Recently, the big data, machine learning and AI topic are interested in many industries. This increasing amount of digital data consequently effects to the role of data analysis as well. In which reflected in the continuously rising amount to researches related to the clustering algorithm in part years, (Liao T. W. (2005), Rokach, L. (2009), Ling H E et al. (2007), Nayak J, Naik B and Behera H S. (2015), Sasirekha, Sasirekha, K. and Baby, P. (2013) and Rui, X., and D. Wunsch. (2005)). Especially for time-series clustering which is reviewed in-depth on and processing detail of the theory by Aghabozorgi, S. et al. (2015). All process of time-series clustering which are distance measurement, time-series prototype, a clustering algorithm, and cluster evaluation. For distance measurement in time-series clustering, Berndt DJ and Clifford J proposed dynamic time warping (DTW) to find patterns in time-series data which time-series clustering by approximate prototypes are propose by Ville Hautamäki et al (2008) to decide cluster presenter. The other important point of this research is cluster validity indices (CVIs) that will be explained by Arbelaitz, O. et al. (2013). And in order to complete this research R software was applied using by Sardá-Espinosa (2018)'s manual. Due to much interesting in time-series clustering, therefore there are studies on this topic; Tsay, R. S. (2010) did cluster analysis with time-series data of American unemployment rate Niennattrakul, V. , & Ratanamahatana, C. A. . (2006) applied time series clustering to compare the efficiency of multimedia data using representation method and multimedia data using traditional processing, so this research has demonstrated the profit of time-series representation method. Saikhamwong N and Rimcharoen S. (2002)did cluster analysis applied with stock data.

3. Time-series clustering

Time-series clustering consists of several parts, including distance measurement, time-series prototype, clustering algorithm, and clustering evaluation. Table 1 shows the overall of each step of time-series clustering. This section briefly describes basic time-series clustering, which used in this work. This paper, the time-series data has 2 types, equal length and non-equal length. Each type, we will compare the clustering algorithm in 3 scenarios. The cryptocurrency data and the Shanghai Stock Exchange (SSE 50) are non-equal lengths, so the algorithm which clustering time-series is hierarchical clustering, partitional clustering with k-medoid and partitional clustering with k-shape. In another data type which is exchange rate currency and the stock exchange of Thailand (SET 50) that use hierarchical clustering, partitional clustering with k-medoid and partitional clustering with TADPole. To compare these scenarios of cluster algorithm whether they are suitable for this dataset or not. We use clustering evaluation approaches such as Silhouette index, COP index, DB index, DB* index and CH index. The flow chart of this research framework shown in Figure 1.

Table 1. The overall each step of time-series clustering

Time-series representations	**Distance measurement**	**Time-series prototypes**	**Clustering algorithm**	**Clustering evaluation**
Data adaptive	Dynamic Time Warping (DTW) - Global DTW constraints - Lower bounds for DTW	Mean and median	Hierarchical clustering	**Internal evaluation** Crisp partitions - Silhouette index - Dunn index - COP index - Davies-Bouldin index - Modified Davies-Bouldin index - Calinski-Harabasz index - Score Function Fuzzy partitions - MPC index - K index - T index - SC index - PBMF index
Non-data adaptive	Global alignment kernel distance	Partition around medoids (PAM)	Partitional cluster - k-medoids - TADPole - k-Shape	
Model-based	Soft-DTW	DTW barycenter averaging (DBA)	Fuzzy clustering	
Data dictated	Shape-based distance (SBD)	Soft-DTW centroid		
		Shape extraction		
		Fuzzy based prototype		**External evaluation** Crisp partitions - Rand Index - Adjusted Rand Index - Jaccard Index - Fowlkes-Mallows - Variation of Information Fuzzy partitions - Soft Rand Index. - Soft Adjusted Rand Index - Soft Variation of Information - Soft Normalized Mutual Information

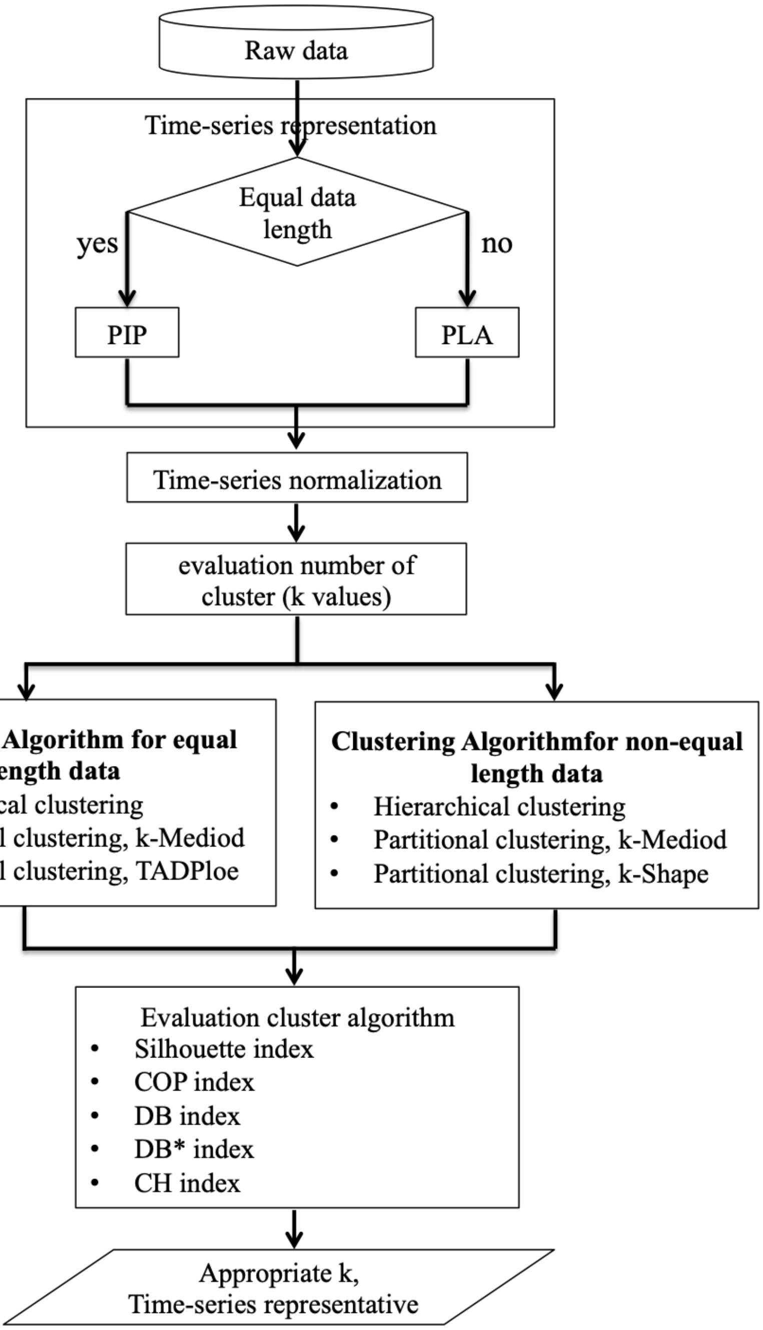

Raw data
Time-series representation
Equal data length
yes
no
PIP
PLA
Time-series normalization
evaluation number of cluster (k values)
Clustering Algorithm for equal length data
• Hierarchical clustering
• Partitional clustering, k-Mediod
• Partitional clustering, TADPloe
Clustering Algorithmfor non-equal length data
• Hierarchical clustering
• Partitional clustering, k-Mediod
• Partitional clustering, k-Shape
Evaluation cluster algorithm
• Silhouette index
• COP index
• DB index
• DB* index
• CH index
Appropriate k,
Time-series representative

Figure 1. Flow chart of this research framework

3.1 Time-series representation

Time-series representation method is dimension reduction, which was the raw represented the raw time-series in another space by transforming data to the lower space dimension or by extracting features. Its advantages are decreasing of the time series dimensionality, emphasizing on essential shape characteristics, noise management, reduce mandatory memory and calculation complexity of machine learning reactions.

Definition: Time-series representation, given a time-series data $S_i = \{s_1, \dots, s_t, \dots, s_T\}$, representation transformed the time-series to another dimensionality reduced vector $S_i = \{s_1, \dots, s_y\}$ where $y < T$ and if two series are similar in the initial space, their representations will be similar in the transformed spaces. Time-series representation methods are broken down to four groups (Saeed Aghabozorgi (2015)), such as data adaptive, non-data adaptive, model-based, and data dictated (clipped data). Data adaptive representation method is used to minimize the reconstruction error of time-series dataset using non-equal length segment. Non-data adaptive representation method is suited for time-series dataset that has equal length segment, while the comparison of representations of several time-series is very direct. Model-based representation is suitable with time-series with stochastic results, Statistical models, and Time-series models. Finally, the representation method is data dictated, the proposal that the condensation-ratio is described based on original time-series data like in the related work Ratanamahatana, C. (2005)

In this paper, we used the non-data adaptive representation method. Perceptually Important Point (PIP) (Aghabozorgi, S. et al. (2015) and Tak Chung Fu et al. (2001)) was used for SET50 and exchange rate currency dataset because time-series dataset is equal length segment. This data adaptive representation method, Piecewise Linear Approximation (PLA) was used for cryptocurrency and SSE 50 dataset, which each cryptocurrency born time is different so time-series length is also different.

3.2 Distance measurement

Dynamic Time Warping (DTW)

Dynamic Time Warping is the dynamic programming method, which used for measuring the similarity between the 2 time-series data. By the results, the distance and the alignment are the best value between two data, which can stretch and shrink for accommodating variations in the axis of time that is shown in Figure2. One point can calculate distance with many points. For example, we have two time-series, a series Y of length a, and series Z of length b. The gab of two time-series is defined as:

$$DTW(Y,Z) = \sum_{i=1}^{a,b} |y_i - z_i| \quad (1)$$

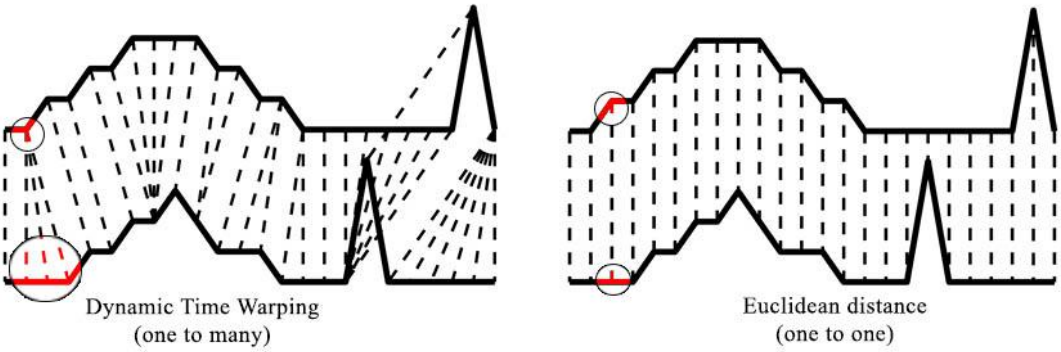


Figure 2. Comparing Dynamic Time Warping distance measurement and Euclidean distance measurement

The features of the time-series data that is different from generic data, causes many problems, which are very high dimensional indices (difficulty to identify the locations of data in high dimension) and other problems that cause by Dynamic Time Warping techniques that is dynamic leads to long–time of calculation and could not speed up.

Shape-based distance (SBD)

The Shape-Based Distance (SBD) was taken a component of K-Shape clustering algorithm. It depends on the Cross-Correlation with Coefficient Normalization (NCC_c) sequence between two series and is sensitive in scale, so normalized data is the suggestions; the distance formula is defined as:

$$SBD(Y,Z) = 1 - \frac{\max(NCCc(y,z))}{\|x\|_2 \|y\|_2} \qquad (2)$$

3.3 Time-series prototype

Partition Around Medoid (PAM)

The prototyping function could be called time-series averaging also. It is used for solidifying the concert or to perform time-series categorization. This step is very important for time-series clustering. Partition Around Medoid approach is the common type of time-series prototype. This approach is suitable with time-series data which structure is not altered. In the implementing process, k series data is randomly chosen as initial centroids. Then the initial centroids and other series calculate distance and series will be appointed to the cluster of its nearest centroid. And the total minimize distance of series is assigned to the new centroid. This iteratively calculates for new centroid until no series could change clusters.

Shape extraction

Another part of k-Shape clustering algorithm is shaping extraction. This centroid function should apply with z-score series data, furthermore, this centroid function can be done between series with the different length, the shape extraction also applied to multivariate series or each variable of all series.

3.4 Clustering algorithm

Hierarchical clustering

Hierarchical clustering(Han, J. and Micheline Kamber (2012)) is an algorithm of cluster analysis. The approaches for hierarchical clustering fallen into 2 types, such as agglomerative and divisive. The agglomerative approach is a bottom to top direction, all observations start as one pair of clusters will be combined as together and moving to top hierarchy. The divisive approach is top to bottom direction, all observations start as single cluster and split. Then executed in circular moving up the hierarchy. This method can display the result as the "Dendrogram" graph. Hierarchical algorithm implementation shows as Table 2

Table 2. Hierarchical algorithm

Hierarchical algorithm
Input: D is a dataset containing n objects. Output: a set of k clusters Method: 1. Compute the distance of the object. Input data in the distance matrix. 2. Repeat 3. Examine for two most related clusters or objects. 4. Combine the two clusters or objects to create a cluster which has 2 objects at least. 5. Update the matrix recalculate the distances between this new cluster and all other clusters again to until no change.

Partitional clustering, k-Medoid

Partitional clustering is different from hierarchical that strategically used to create partitions. The k-shape and k-medoids are the most popular partitional algorithms. For this algorithm, we must assign the number of the cluster as "k" value for calculation. The optimal k value can be indicated by using the cluster evaluation procedure. The partitional algorithm with k-medoid implementation shows as Table 3.

Table 3. The partitional algorithm with k-medoid

The partitional algorithm with k-medoid

```
Input:          k is the amount of clusters, D is a dataset containing n objects.
Output:  a set of k clusters
Method:
1. Promptly select k objects from D as the starting representative object or seed;
2. Repeat
3.      Designate remaining objects onto the cluster with the closest representative object;
4.      Casually choose one non-representative  object, $O_{random}$;
5.      Calculate total cost, S, of switching representative object, $O_j$, with $O_{random}$;
6.      If $S < 0$ then switch $O_j$ with $O_{random}$ to the new set of $k$ the representative object until no change
```

Partitional clustering, k-Shape

Paparrizos, J. , and Gravano, L. . (2016) researched the k-Shape clustering algorithm was by in 2016. This partitional algorithm needs the custom distance measure as SBD and the custom centroid function as shape extraction. It also requires z-normalization to default data. The partitional algorithm with k-shape implementation shows as .

Table 4.

Table 4. The partitional algorithm with k-shape

Partitioning algorithm with k-shape

```
Input:          k is the number of clusters, D is a dataset containing n objects.
Output:  a set of k clusters
Method:
1. Promptly select choose k objects from D as the starting representative object or seed;
2. repeat
3.      Base on the mean value of the objects in the cluster, designate remaining objects onto the cluster with a similar object
4.      Recomputed the mean value of the objects in each cluster to update the cluster until no change
```

Partitional clustering, TADPole

The TADPole clustering was proposed by Begum, N. et al. (2015), this time-series clustering use DTW as the distance measurement and use PAM as the centroid. TADPole depends on the DTW bounds defining for time-series with equal length only.

3.5 Cluster evaluation

The cluster evaluation can be split into two types such as external and internal evaluations. If we know the group truth, we can use external evaluation to recheck the clustering with the group truth and re-measure without the group truth, we can use the internal evaluation to measure separating of cluster. In this paper, we will use Silhouette index, COP index, Davies-Bouldin index, Davies-Bouldin star index, and Calinshi-Harabasz index of internal evaluation to compare time-series clustering.

Silhouette index

Silhouette index is the internal validation of consistency within cluster data. Some algorithm needs to know k-value before, therefore silhouette is used to measure k-value which control the amount of clusters in a dataset and relationship between objects in the dataset. The silhouette coefficient value should always be the maximum number after measurement.

Definition: The dataset D, of n objects, suppose D is segregated to k clusters, $C_1, \dots, C_k$. For each object $u \in D$, we compute $a(u)$ as the average distance of u and other objects in the cluster where u belongs. Likewise, $b(u)$ is the minimum average distance from u to call clusters to where u does not belongs. Formally, suppose $u \in C_i (1 \le i \le k)$; then

$$a(u) = \frac{\sum u' \in C_i, u \neq u' \ \mathrm{dist}(u,u')}{|C_i| - 1} \qquad (2)$$

and

$$b(u) = \min_{C_j : 1 \le j \le k, j \neq i} \frac{\sum_{u' \in C_j} \mathrm{dist}(u,u')}{|C_j|} \qquad (3)$$

The Silhouette coefficient (Han, J. (2005)) of u is then defined as

$$s(u) = \frac{b(u) - a(u)}{\max \{a(u), b(u)\}} \qquad (4)$$

COP index

It was first presented to adopted in the partnership of a cluster hierarchy post-processing algorithm, Traditional cluster validity indices could also adopt this. It is a ratio-type index, where the coherence is indicated by the distance from the position in a cluster to its centroid. The partition establish based on the furthest border width is defined as:

$$\mathrm{COP}(C) = \frac{1}{N} \sum_{c_k \in C} |c_k| \frac{\frac{1}{|c_k|} \sum_{x_i \in c_k} d_e(x_i, \bar{c_k})}{\min_{x_i \notin c_k} \max_{x_j \in c_k} d_e(x_i, x_j)} \qquad (5)$$

Davies-Bouldin index (DB)

This index is commonly used in cluster validity comparison studies. It assesses the coherence formed on the distance from the positions in a cluster to centroid and partition based on the distance of centroids. It is defined as

$$\mathrm{DB}(C) = \frac{1}{K} \sum_{c_k \in C} \max_{c_l \in C \setminus c_k} \left\{ \frac{S(c_k) + S(c_l)}{d_e(\bar{c_k}, \bar{c_l})} \right\} \qquad (6)$$

where

$$S(c_k) = \frac{1}{c_k} \sum_{x_i \in c_k} d_e(x_i, \bar{c_k}) \qquad (7)$$

Davies-Bouldin star index (DB star)

This index is the variation of the Davies-Bouldin index. It is defined as

$$\mathrm{DB}^*(C) = \frac{1}{K} \sum_{c_k \in C} \frac{\max_{c_l \in C \setminus c_k} \{S(c_k) + S(c_l)\}}{\min_{c_l \in C \setminus c_k} \{d_e(\bar{c_k}, \bar{c_l})\}} \qquad (8)$$

Calinshi-Harabasz index (CH)

This ratio index cohesion is predicted from the distance from the point in a cluster to its centroids. Its partition is a measurement of the distance from the centroids to the global centroid. This can be defined as:

$$CH(C) = \frac{N-K}{K-1}\,\frac{\sum_{c_k \in C} |c_k|\, d_e(\overline{c_k}, \overline{X})}{\sum_{c_k \in C} \sum_{x_i \in c_k} d_e(x_i, \overline{c_k})} \tag{9}$$

4. Experiment result

4.1 Dataset

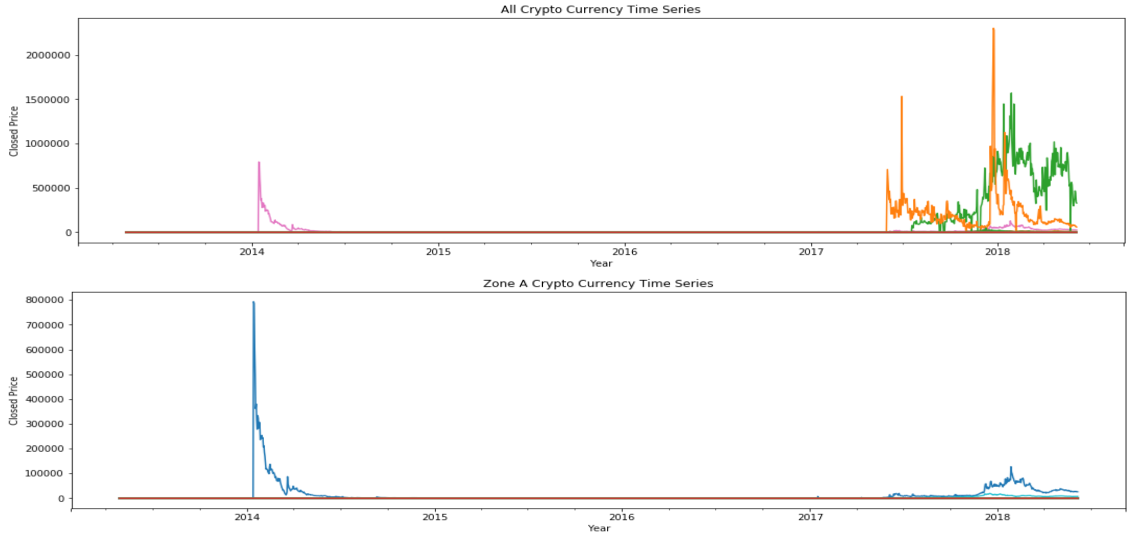
This paper has 4 dataset experiments, which are cryptocurrency dataset, SSE 50 dataset, exchange rate currency dataset, and SET 50 dataset. Dataset structure detail shows as Table 5.

Table 5. Dataset structure detail

Dataset	**Length**	**No. Currency / Stock**	**No. Data Points**
Cryptocurrency, Zone A	1,866 (Unequal length)	154	234,949
Cryptocurrency, Zone B	1,096 (Unequal length)	297	130,663
Cryptocurrency, Zone C	522 (Unequal length)	1,192	265,547
SSE50	2,435 (Unequal length)	43	99,353
Exchange rate currency	92 (Equal length)	146	13,432
SET50	244 (Equal length)	50	12,200

Cryptocurrency dataset

Cryptocurrency dataset is all historical closing price of all cryptocurrencies from the Kaggle website (www.kaggle.com/jessevent/all-crypto-currencies/home). All datasets have 631,159 observations, the used variables are currency, data, closing price and period of dataset between 28, April 2013 to 21, May 2018. But each cryptocurrency has a different length, therefore we split cryptocurrency to 3 zones; zone A has 154 cryptocurrencies that have 1,866 lengths of 234,949 data points, zone B has 297 cryptocurrencies that have 1,096 lengths of 130,663 data points, and zone C has1,192 cryptocurrencies that have 522 lengths of 631,159 data points. Time-series dataset of each zone shows as Figure 3.



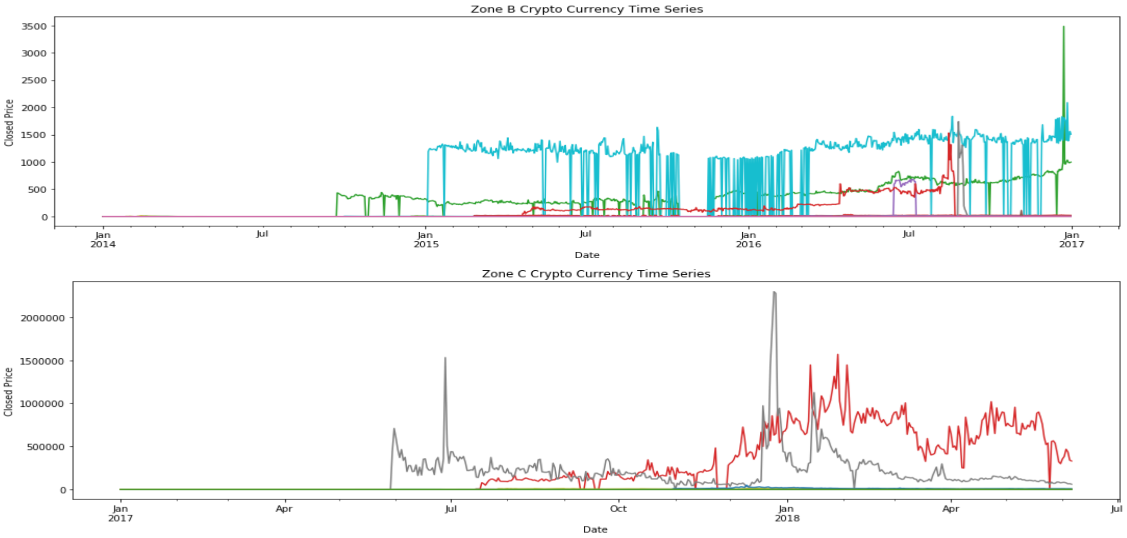


Figure 3. Time-series of all cryptocurrency and breaking down as zone A, B, and C

Shanghai Stock Exchange 50 Index (SSE 50)

Shanghai Stock Exchange, SSE 50 Index, is a capitalization-weighted index based on the top 50 stocks listed in China SSE index. It has high market capitalization and high liquidity. SSE 50 dataset is the historical closing price for 50 stock indices from finance yahoo.com website (https://finance.yahoo.com). But some stock indices have too much of null value, so we cut 7 stock indices off. All datasets have the unequal length of 99,353 data points of 43 stock indices; the used variables are currency data, closing price and period of dataset between 15, December 2008 to 15, December 2018, shows as Figure 4.

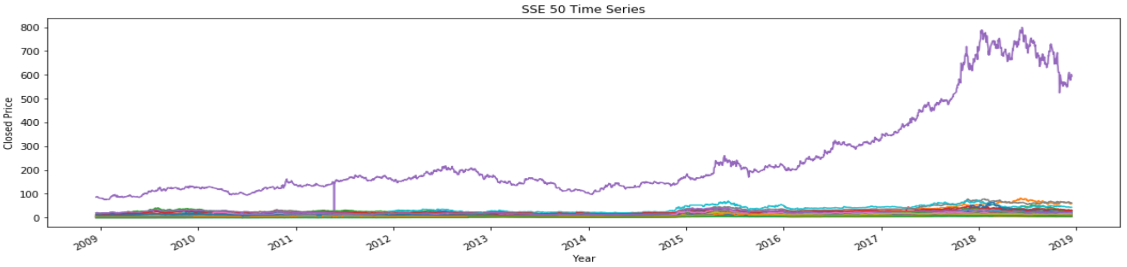


Figure 4. Time-series of SSE 50 stock index

Exchange rate currency dataset

Exchange rate currency dataset is the historical closing price for all currencies in the world that has 92 currencies from finance yahoo.com website (https://finance.yahoo.com). All datasets have the same length of 13,432 data points; the used variables are currency, data, closing price, and period of dataset between 1, July 2018 to 30, September 2018, shows as Figure 5.

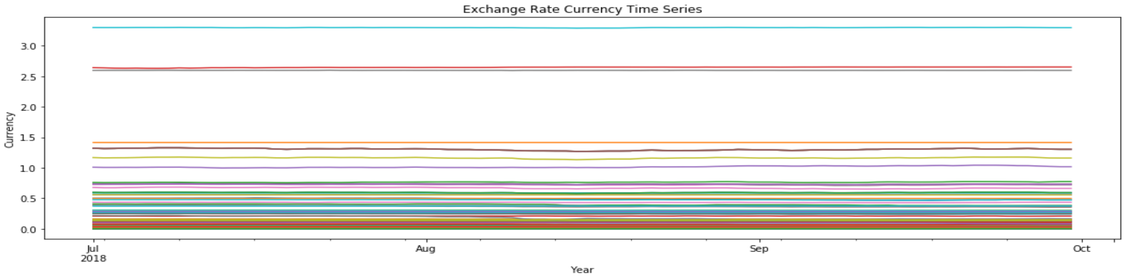


Figure 5. Time-series of all exchange rate currency

The Stock Exchange of Thailand 50 (SET 50) dataset

The stock exchange of Thailand, SET 50 Index, is a capitalization-weighted index based on the top 50 stocks listed in Bangkok. Thailand SET index has high market capitalization and high liquidity. SET 50 dataset is the historical closing price of 50 stock indices from investing website (www.investing.com). All dataset has the same length of 12,200 data points; the used variables are currency data, closing price and period of dataset between 1, October 2017 to 30, September 2018, shows as Figure 6.

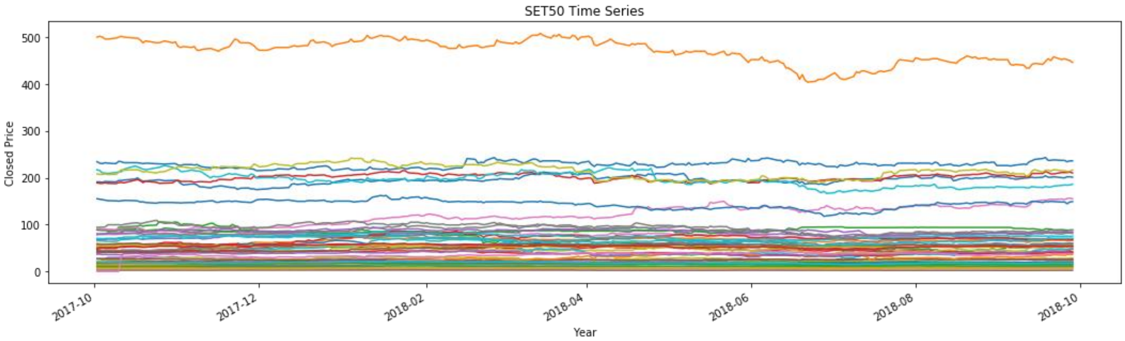


Figure 6. Time-series of SET 50 index stock

4.2 Cluster algorithm

This paper will break down 3 scenarios for each data type of experiment. The scenario of the clustering algorithm detail describes in Table 6. The hierarchical and partitional algorithm was needed to normalize data and cluster evaluating to find the best number of cluster (k-value).

Table 6. The scenario of the clustering algorithm by a column of distance measurement and centroid

Scenario	**Distance measurement**	**Centroid**	**Clustering algorithm**	**Data type**
hc	DTW	PAM	Hierarchical	Equal or non equal length
pc_dtw	DTW	PAM	Partitional with k-medoid	Equal or non equal length
pc_sbd	SBD	Shape extraction	Partitional with k-shape	Non-equal length only
pc_tp	DTW	PAM	Partitional with TADPole	Equal length only

To evaluate the clustering algorithm, we compare 3 scenarios of the clustering algorithm, time series representation, a common image clustering algorithm, and clustering by the color of currency values.

4.3 The time-series clustering experiment result

The cryptocurrency clustering result

Before running the clustering algorithm, we have to prepare time-series data by doing representation methods to reduce noise and normalize time-series. To evaluate the number of clusters (k-value), we calculate Silhouette coefficient value to identify k-value of each clustering algorithm of zone A, B, and C that show as Figure 7, Figure 8, and Figure 9 respectively. In zone A, the k-value of hierarchical and partitional with k-medoid scenario resulted in for 2 clusters as best, while the partitional with k-shape is 4 clusters. The series and centroid result of the hierarchical scenario, the partitional with the k-medoid scenario, and the partitional with k-shape scenario are shown as Figure 10, Figure 11, and Figure 12 respectively.

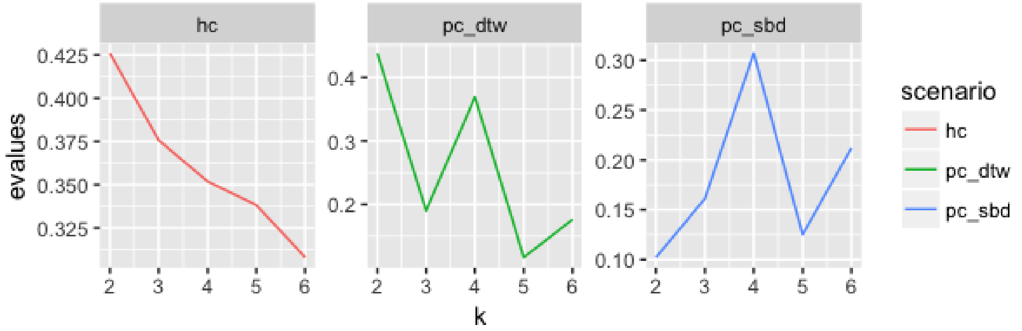


Figure 7. Cryptocurrency, zone A of Silhouette index of each scenario algorithm and each k clusters

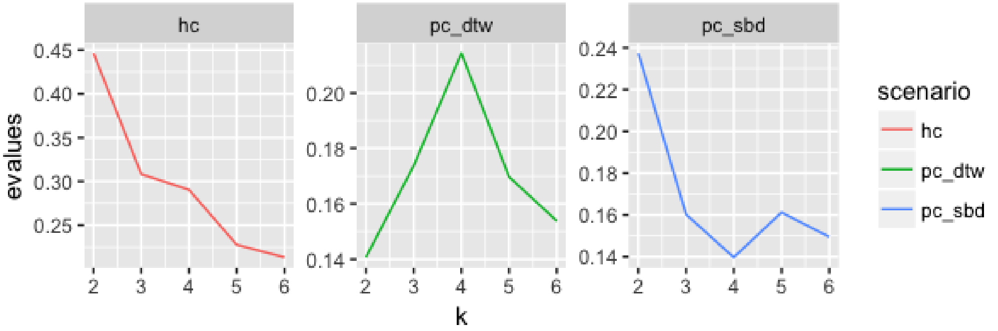


Figure 8. Cryptocurrency, zone B of Silhouette index of each scenario algorithm and each k clusters

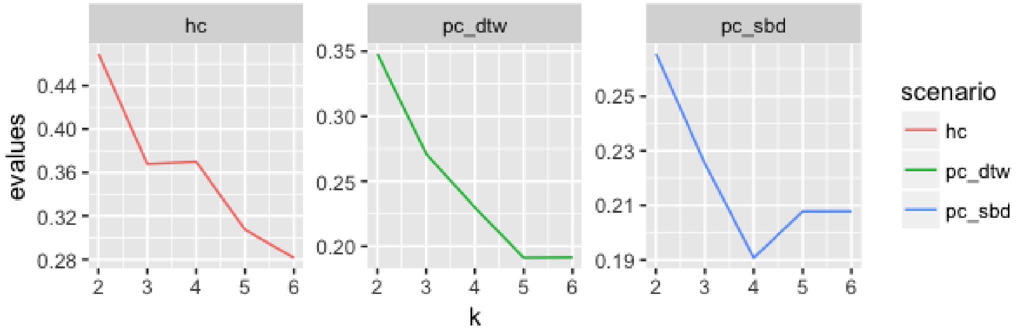


Figure 9. Cryptocurrency, zone C of Silhouette index of each scenario algorithm and each k clusters

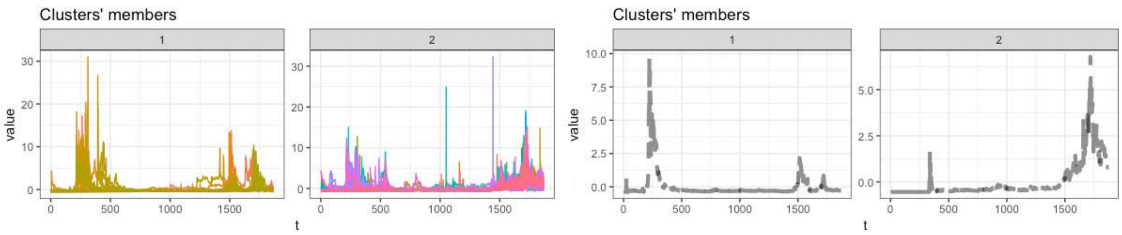


Figure 10. The series (left) and centroid (right) result of the hierarchical scenario of cryptocurrency time-series, zone A

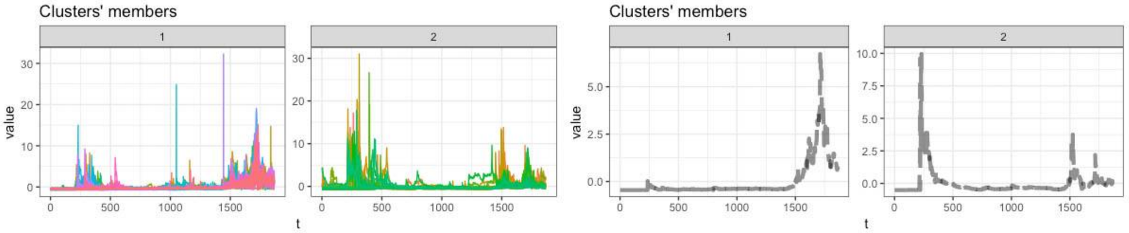


Figure 11. The series (left) and centroid (right) result of partitional with the k-medoid scenario of cryptocurrency time-series, zone A

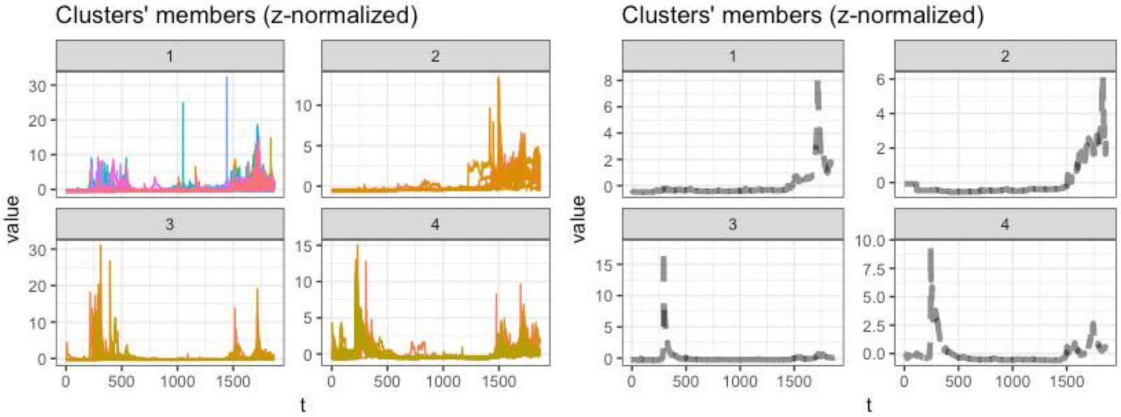


Figure 12. The series (left) and centroid (right) result of partition with the k-shape scenario of cryptocurrency time-series, zone A

In zone B, the k-value of hierarchical and partitional with k-shape scenario resulted in 2 clusters as best, while the partitional with k-medoid is 4 clusters. The series and centroid result of the hierarchical scenario, the partitional with the k-medoid scenario, and the partitional with k-shape scenario are shown as Figure 13, Figure 14, and Figure 15 respectively.

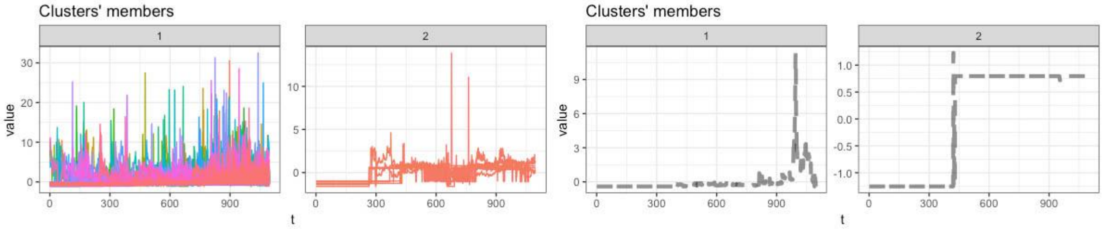


Figure 13. The series (left) and centroid (right) result of the hierarchical scenario of cryptocurrency time-series, zone B

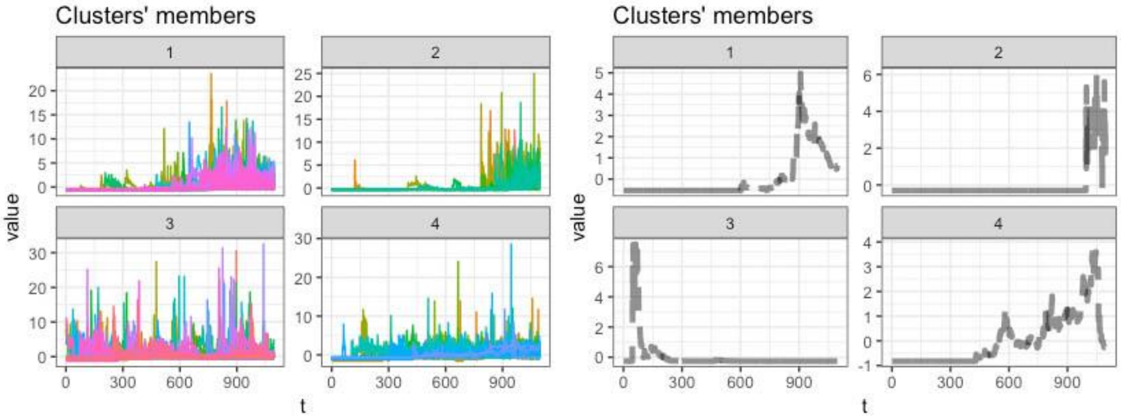


Figure 14. The series (left) and centroid (right) result of partitional with the k-medoid scenario of cryptocurrency time-series, zone B

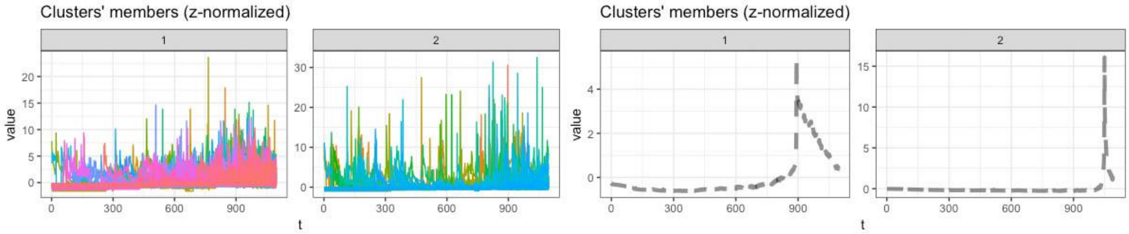


Figure 15. The series (left) and centroid (right) result of partitional with the k-shape scenario of cryptocurrency time-series, zone B.

In zone C, The k-value of hierarchical, partitional with k-shape and partitional with k-medoid scenario resulted in 2 clusters as best. The series and centroid result of the hierarchical scenario, the partitional with the k-medoid scenario and the partitional with k-shape scenario are shown as Figure 16, Figure 17, and Figure 18 respectively.

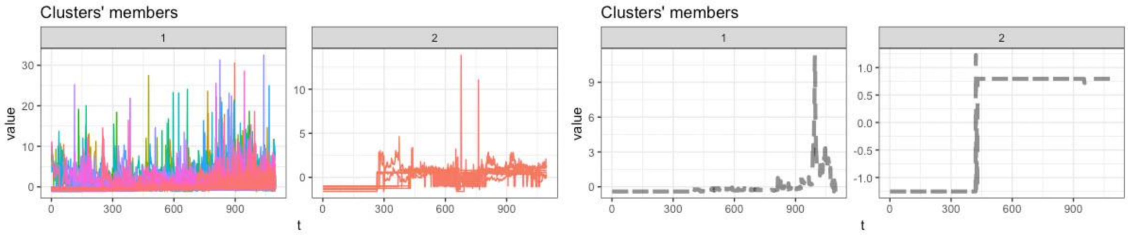


Figure 16. The series (left) and centroid (right) result of the hierarchical scenario of cryptocurrency time-series, zone C

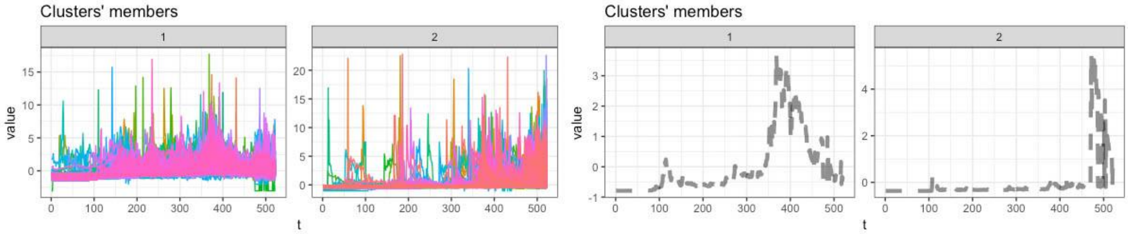


Figure 17. The series (left) and centroid (right) result of partitional with the k-medoid scenario of cryptocurrency time-series, zone C

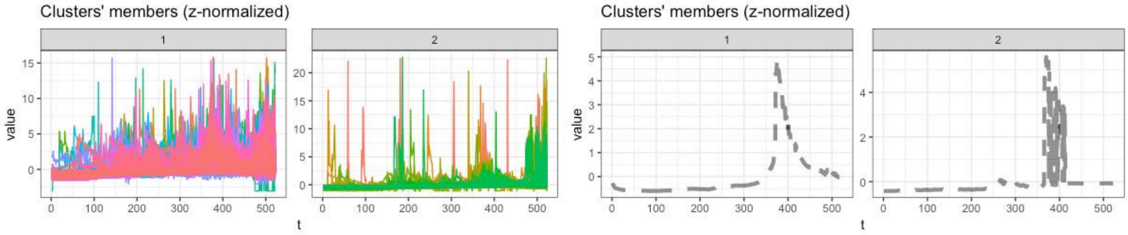


Figure 18. the series (left) and centroid (right) result of partitional with the k-shape scenario of cryptocurrency time-series, zone C

The Shanghai Stock Exchange 50 Index (SSE 50) clustering result

For Shanghai stock exchange 50 indices, the k-value of hierarchical, partitional with k-shape, and partitional with k-medoid scenario resulted in 3, 2, and 4 clusters respectively as best, the result showed as Figure 19. The series and centroid result of the hierarchical scenario, the partitional with the k-medoid scenario, and the partitional with k-shape scenario are shown as Figure 20, Figure 21, and Figure 22 respectively.

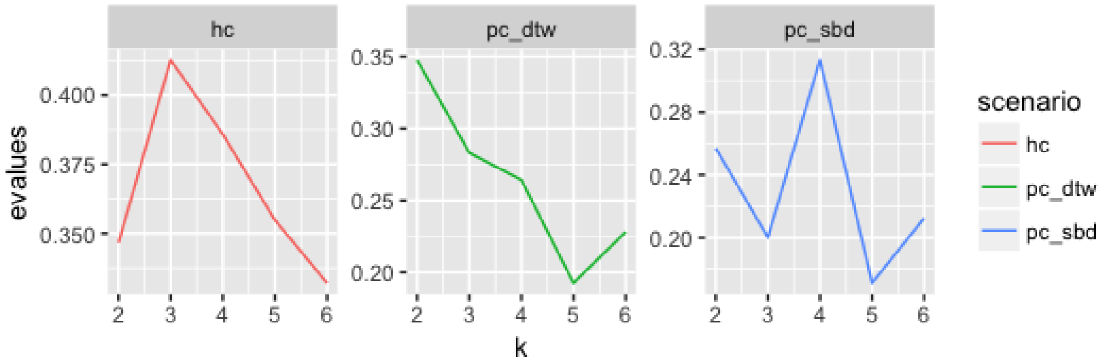


Figure 19. The Shanghai Stock Exchange 50 Index (SSE 50) of Silhouette index of each scenario algorithm and each k clusters

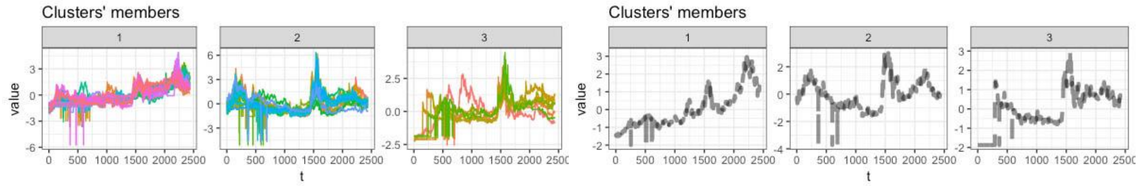


Figure 20. The series (left) and centroid (right) result of hierarchical scenario of SSE 50

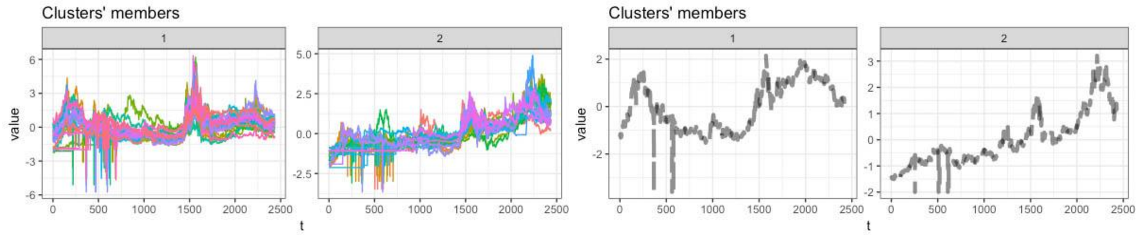


Figure 21. The series (left) and centroid (right) result of partitional with k-medoid scenario of SSE 50

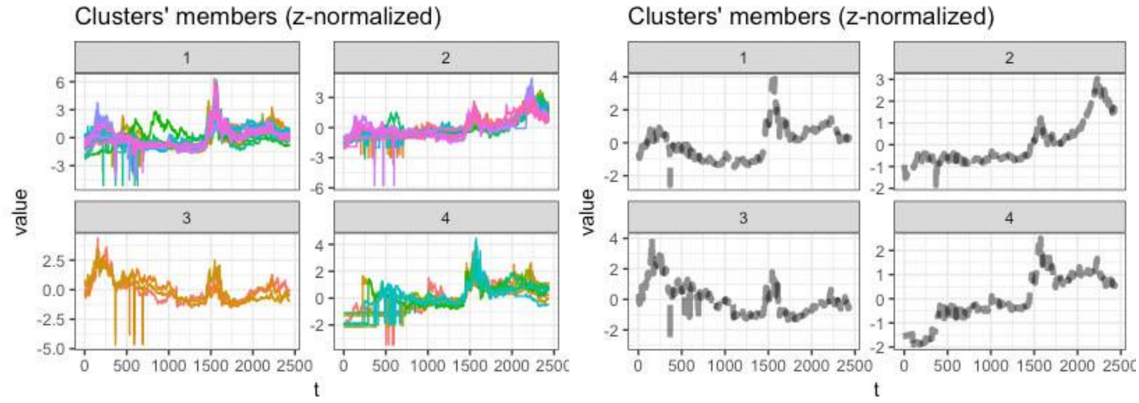


Figure 22. The series (left) and centroid (right) result of partition with the k-shape scenario of SSE 50

The exchange rate currency clustering result

For the exchange rate currency, the k-value of all scenarios resulted in 2 clusters as best, the result shown as Figure 23. The series and centroid result of the hierarchical scenario, the partitional with the k-medoid scenario, and the partitional with k-shape scenario are shown as Figure 24, Figure 25, and Figure 26 respectively.

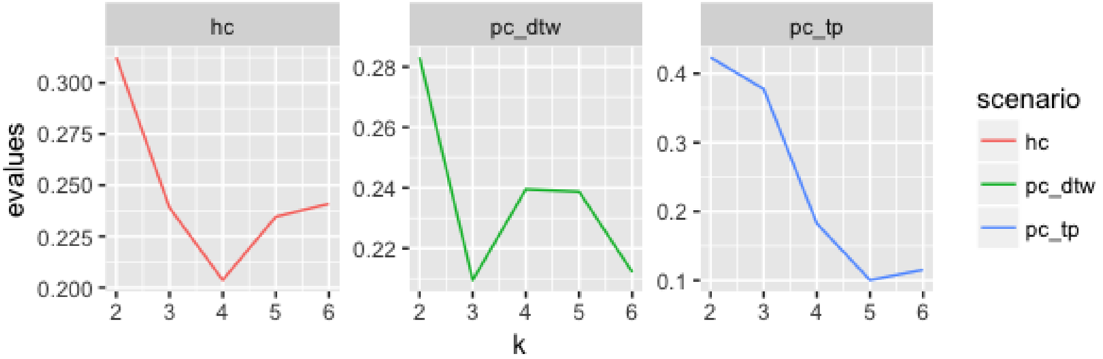


Figure 23. The exchange rate currency of Silhouette index of each scenario algorithm and each k clusters

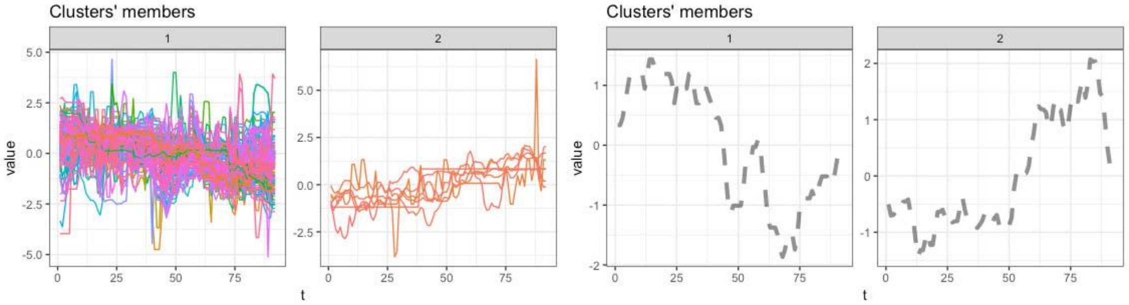


Figure 24. The series (left) and centroid (right) result of hierarchical scenario of the exchange rate currency

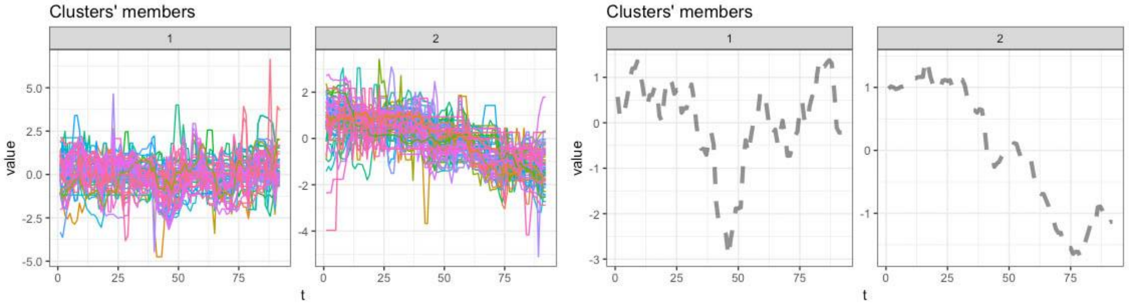


Figure 25. The series (left) and centroid (right) result of partitional with the k-medoid scenario of the exchange rate currency

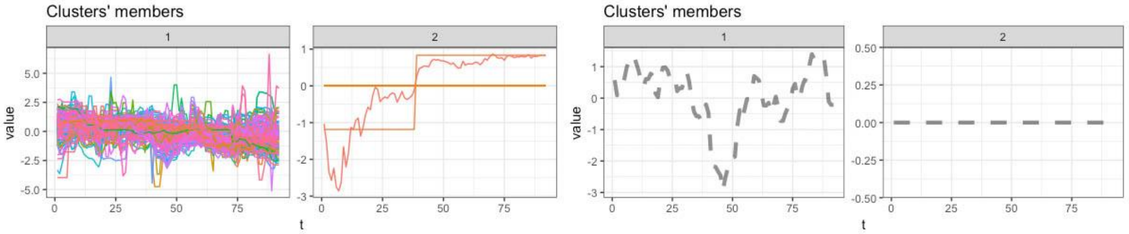


Figure 26. The series (left) and centroid (right) result of partitional with TADPole scenario of the exchange rate currency

The Stock Exchange of Thailand 50 (SET 50) clustering result

For SET 50, The k-value of hierarchical and partitional with k-medoid scenario resulted in 2 clusters as best, the partitional with k-shape is 4 clusters, the result showed as Figure 27. The series and centroid result of the hierarchical scenario, the partitional with the k-medoid scenario and the partitional with TADPole scenario are shown as Figure 28 Figure 29 and Figure 30 respectively.

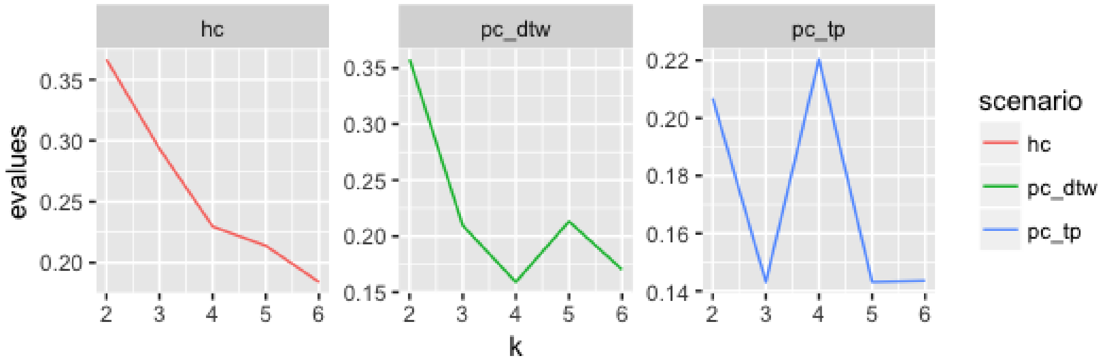


Figure 27. SET 50 of Silhouette index of each scenario algorithm and each k clusters

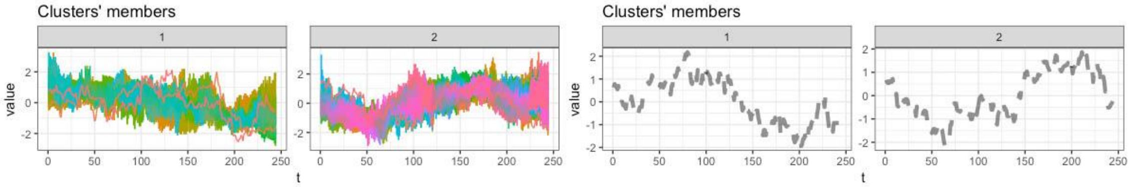


Figure 28. The series (left) and centroid (right) result of the hierarchical scenario of SET 50

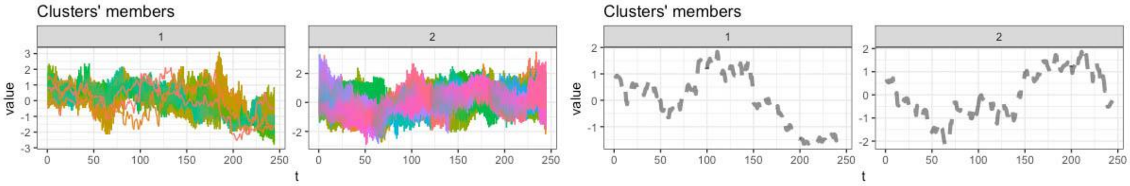


Figure 29. The series (left) and centroid (right) result of partitional with the k-medoid scenario of SET 50

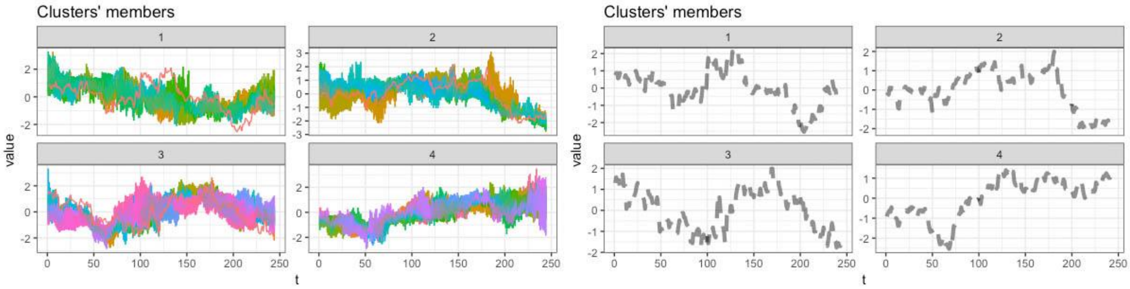


Figure 30. The series (left) and centroid (right) result of partitional with TADPole scenario of SET 50

4.3 Comparing time-series clustering result

To evaluate our proposed method, we run 3 zones of cryptocurrency series using 3 clustering algorithms such as the hierarchical scenario, the partitional with the k-medoid scenario, and the partitional with k-shape which are evaluated by Silhouette index, COP index, DB index, DB star index, and CH index.

Table 7,

	Sil (max)	COP (min)	DB (min)	DBStar (min)	CH (max)
HC	**0.45**	0.41	**0.70**	**0.70**	74.62
DTW	0.44	0.39	0.91	0.91	**148.81**
PC_k-Shape	0.23	**0.24**	1.30	1.39	39.93

Table 8 8, 9 and 10 report the clustering evaluation value that represent in the same direction, the hierarchical scenario is the most effective for the unequal dataset; cryptocurrency and SSE 50 time-series. For exchange rate currency and SET 50 which are the equal length time-series, their evaluation values are shown as

Table 11 and

Table 12 respectively. The most effective algorithm with exchange rate currency and SET 50 is partitional with TADPole and partitional with k-medoid respectively.

Table 7. Cryptocurrency, zone A clustering evaluation of each scenario

	Sil (max)	COP (min)	DB (min)	DBStar (min)	CH (max)
HC	**0.45**	0.41	**0.70**	**0.70**	74.62
DTW	0.44	0.39	0.91	0.91	**148.81**
PC_k-Shape	0.23	**0.24**	1.30	1.39	39.93

Table 8. Cryptocurrency, zone B clustering evaluation of each scenario

	Sil (max)	COP (min)	DB (min)	DBStar (min)	CH (max)
HC	**0.45**	0.53	**0.89**	**0.89**	16.28
PC_DTW	0.18	**0.28**	1.96	2.20	103.47
PC_k-Shape	0.24	0.37	1.28	1.28	**144.10**

Table 9. Cryptocurrency, zone C clustering evaluation of each scenario

	Sil (max)	COP (min)	DB (min)	DBStar (min)	CH (max)
HC	**0.51**	1.13	**0.31**	**0.31**	3.25
PC_DTW	0.35	0.35	1.51	1.51	**1,075.66**
PC_k-Shape	0.27	**0.30**	1.85	1.85	550.91

Table 10. SSE 50 clustering evaluation of each scenario

	Sil (max)	COP (min)	DB (min)	DBStar (min)	CH (max)
HC	**0.41**	**0.30**	**0.99**	**1.00**	23.52
DTW	0.35	0.49	1.54	1.54	**28.76**
PC_k-Shape	0.10	0.59	2.42	2.88	11.56

Table 11. Exchange rate currency clustering evaluation of each scenario

	Sil (max)	COP (min)	DB (min)	DBStar (min)	CH (max)
HC	0.31	0.70	1.02	1.02	16.78
PC_DTW	0.28	**0.46**	1.15	1.15	**111.73**
PC_TADPole	**0.42**	0.98	**0.72**	**0.72**	13.38

Table 12. SET 50 clustering evaluation of each scenario

	Sil (max)	COP (min)	DB (min)	DBStar (min)	CH (max)
HC	0.36	**0.43**	1.33	1.33	**26.59**
PC_DTW	**0.37**	**0.43**	**1.19**	**1.19**	25.41
PC_TADPole	0.22	0.55	2.82	3.18	14.72

5. Conclusion

According to Table 13, which we have demonstrated the clustering time-series of some kind of financial time-series. We use 4 time-series datasets, which can split into 2 data type (equal and unequal length). This experiment, comparing time-series clustering using 3 scenarios of cluster algorithm for each time-series data set and evaluating clustering algorithm using 5 indices to identify the validity of each clustering algorithm. From research result, the hierarchical algorithm is the most efficient algorithm for unequal length of cryptocurrency series and SSE 50. In another hand, the partitional algorithm is the most efficient for an equal length of exchange rate currency and SET 50.

Table 13. Conclusion of research

Dataset	Data type	The Best Algorithm	No. of Cluster
Cryptocurrency, Zone A	Unequal length	Hierarchical	2
Cryptocurrency, Zone B	Unequal length	Hierarchical	2
Cryptocurrency, Zone C	Unequal length	Hierarchical	2
SSE50	Unequal length	Hierarchical	3
Exchange rate currency	Equal length	Partitional with TADPole	2
SET50	Equal length	Partitional with k-medoid	2

Exploring more on the source of the result, some element of the unequal length data might contribute to the result such as the fact that original data is very large in both the data period and type of currency dimensions. when we execute and cluster data, it is hard to do it in equal length manner, where SSE50, includes 50 stock indices and data period is 10 years while cryptocurrency, includes 643 currencies in 5 years. Therefore, the hierarchical algorithm is suitable with the larger and high dynamic variant datasets. On another hand, SET50 includes 50 stock indices with fixed data only period of 1 year while exchange rate currency in the study includes 92 currencies around the world within 3-month period. With these low amount and much less variant dataset, a partitional clustering algorithm is more suitable. We hope our research would be good motivation for researchers to study further on time-series clustering and its application in much wider area including biometric clustering, exchange rate currency and stock clustering to manage the trading portfolio.

Acknowledge

Firstly, I would first like to thank my thesis advisor Professor Li Jin Yu of the School of Mathematics at China University of Mining and Technology. The door to Prof. Li office was always open whenever I ran into trouble or had a question about my research. His opened mind allows me to put my curiosity at its best and push this workpiece beyond its stand by steering me on in the right direction and provide constructive comment to it. I would also like to acknowledge Mr. Pakorn Leewasuthorn, my forever friend and Ms. Zhu Lin my senior apprentice sister at faculty educational companion as the second reader of this thesis. I am gratefully indebted to his/her for very valuable comments on this thesis. I would also like to acknowledge Jiangsu Province and China University of Mining and Technology who supported scholarship and *this valuable opportunity to study in China. I would also like to acknowledge my senior apprentice Zhu Lin, Zhao Xing Wei, Xu Yuan Yuan, Chinese classmates, Wang Pei, Dong Min Jie, Yang Jia Hui, Cheng Li, Aommy, Jo Jo, Maylo, all my Chinese and International friends that gave me long last friendship，and emotional support, take care and made my life in China meaningful. Finally, I must express my very profound gratitude to my parents and older brothers, my grandma, and my aunt for providing me with unfailing support, give me freedom and continuous encouragement throughout my years of study and through the processes of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.*

Reference

Ville Hautamäki, Pekka Nykänen, & Pasi Fränti. (2008). Time-series Clustering by Approximate Prototypes. International Conference on Pattern Recognition. IEEE.

Niennattrakul, V. , & Ratanamahatana, C. A. . (2006). Clustering Multimedia Data Using Time Series. International Conference on Hybrid Information Technology. IEEE.

Gullo, F. , Ponti, G. , Tagarelli, A. , Tradigo, G. , & Veltri, P. . (2012). A time series approach for clustering mass spectrometry data. Journal of Computational Science, 3(5), 344-355.

Izakian H, Pedrycz W, Jamal I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. Engineering Applications of Artificial Intelligence, 39, 235-244.

Liao, T. W. . (2005). Clustering of time series data—a survey. Pattern Recognition, 38(11), 1857-1874.

PA Wang, W. , & Zhang, Y. . (2007). On fuzzy cluster validity indices. Fuzzy Sets and Systems, 158(19), 2095-2117.

Berndt DJ, Clifford J. (1994). Using dynamic time warping to find patterns in time series. In KKD workshop, 10(1), 359-370.

Rokach, Lior, and Oded Maimon.(2005). "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 321-352.

Begum, N. , Ulanova, L. , Wang, J. , & Keogh, A. E. . (2015). Accelerating dynamic time warping clustering with a novel admissible pruning strategy.

Ratanamahatana, C., Keogh, E., Bagnall, A. J., & Lonardi, S. (2005). A Novel Bit Level Time Series Representation with Implication of Similarity Search and Clustering. Pacific-asia Conference on Advances in Knowledge Discovery & Data Mining.

HHan, J. (2005). Data Mining: Concepts and Techniques.

Paparrizos, J. , & Gravano, L. . (2016). K-shape: efficient and accurate clustering of time series. ACM SIGMOD Record, 45(1), 69-76.

Aghabozorgi, S. , Shirkhorshidi, A. S. , & Wah, T. Y. . (2015). Time-series clustering - A decade review. Elsevier Science Ltd.

Tak Chung Fu - F.L. Chung - Robert Wing Pong Luk - Vincent T. Y. NgVincent. (2001). Flexible time series pattern matching based on perceptually important points. JT Conference on Artificial Intelligence Workshop, 1-7

Arbelaitz, O. , Gurrutxaga, I. , Muguerza, J. , Jesús M. Pérez, & Iñigo Perona. (2013). An extensive comparative study of cluster validity indices. pattern recognit. Pattern Recognition, 46(1), 243-256.

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193-218.

Sardá-Espinosa. (2018). Comparing Time-Series Clustering Algorithms in R using the dtwclust Package - Retrieved from https://cran.r-project.org/web/packages/dtwclust/vignettes/dtwclust.pdf

Tsay, R. S. . (2010). Multivariate Time Series Analysis and Its Applications. Analysis of Financial Time Series, Second Edition. John Wiley & Sons, Inc.

Rokach, L. . (2009). A survey of clustering algorithms. Data Mining & Knowledge Discovery Handbook, 16(3), 269-298.

Ling H E , Ling-Da W U , Yi-Chao C . (2007). Survey of Clustering Algorithms in Data Mining. Application Research of Computers, 24(1), 10-13.

Nayak J, Naik B, Behera H S. (2015). Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014. Computational Intelligence in Data Mining, 2(1).

Sasirekha, K. , & Baby, P. . (2013). Agglomerative hierarchical clustering. Electronic Design, 17-17.

Rui, X., & D. Wunsch. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645-678.

Saikhamwong N - Rimcharoen S.(2002).K-Mean Clustering of the Stock Exchange of Thailand 50 (SET50) for Portfolio Diversification. 11th International Conference on e-Business (iNCEB 2013).

Cite this article:

Tangsirisakul, D. (2019). Comparing Clustering Algorithms using Financial Time-series data. ***International Journal of Science and Business,*** *3(2), 146-166*. doi: https://doi.org/10.5281/zenodo.2617341

Retrieved from http://ijsab.com/wp-content/uploads/334.pdf

Published by

IJSB International