Amplitude-Based Time Series Data Clustering Method

Yukari Shirota*, and Basabi Chakraborty**

* Faculty of Economics Gakushuin University Tokyo, Japan
yukari.shirotaATgakushuin.ac.jp

**Faculty of Software and Information Science Iwate Prefectural University Morioka, Japan
basabiATiwate-pu.ac.jp

Abstract

In the paper, we propose an amplitude-based time series data clustering method. When we analyze the trend index movement in economy, shape-based clustering does not work well because the standardization/z-normalization is required in advance on the input data and the standardization removes the amplitude/variance information from the original data. Then, the flat fluctuation may often become a large-variance fluctuation by the standardization, which is a problem. To solve the problem, we proposed a method by Amplitude-based time series data clustering method which uses Euclidean distance of Euclidean distances as the distance measurement. In the paper, we investigate the performance of the method, using the real stock prices data. The data are the indexed growth rate patterns of stock prices. Our proposed method could divide the companies' stocks as we humans did, and the result met our requirements. The proposed amplitude-based time series data clustering method is helpful in economic indexed growth data clustering.

Keywords: amplitude-based time series data clustering, Euclidean distance, k-Shape, DTW, hierarchical clustering

1. Introduction

In the paper, we develop a suitable clustering method for comparison of time series data considering amplitude similarities. There have been many research papers concerning time series data clustering. The methods used are k-Shape method, k-means [1, 2] with DTW (Dynamic Time Warping) [3-5] as the distance, HRP (Hierarchical Risk Parity)[2, 6, 7], and so on. These methods require the standardization (z-normalization) of the input data. For example, k-Shape requires the z-normalization [8, 9].

For example, two stock price series must be standardized per company, in advance, so that we can compare and measure the similarity between them. However, the standardization removes the amplitude/strength information from the original data. Let us suppose that there are two similar fluctuations A and B with different amplitudes. Even if data A presents a rapid growth compared to data B, after the standardization, both shapes become the same similar one.

In Fig. 1, the indexed fluctuations that we often use are illustrated as sample data where the starting figure is set to be 1 as the base line. The upper figure in Fig. 1 is the original data and the lower figure shows its standardized fluctuations. We would like to conduct a clustering which can identify a rapid growth group or a slow growth group and so forth. However, after the standardization, almost every data series shows very similar increase of growth patterns as shown in the lower figure in Fig. 1. In addition, the flat fluctuation shown in purple in the original data became a large variance fluctuation in the standardized version. Our aim is that if that is a flat fluctuation, compared to others, we would like to observe it as a still/flat one.

Our aim is to find a suitable shape-based clustering method which can also keep the amplitude information. In this paper, we have evaluated a hierarchical clustering method by Euclidean distances without the data standardization which is our proposed method. For the performance comparison, we also take the HRP method which uses the correlation coefficients as distances. As a result, we found that the hierarchical clustering with Euclidean distances was superior to the HRP.

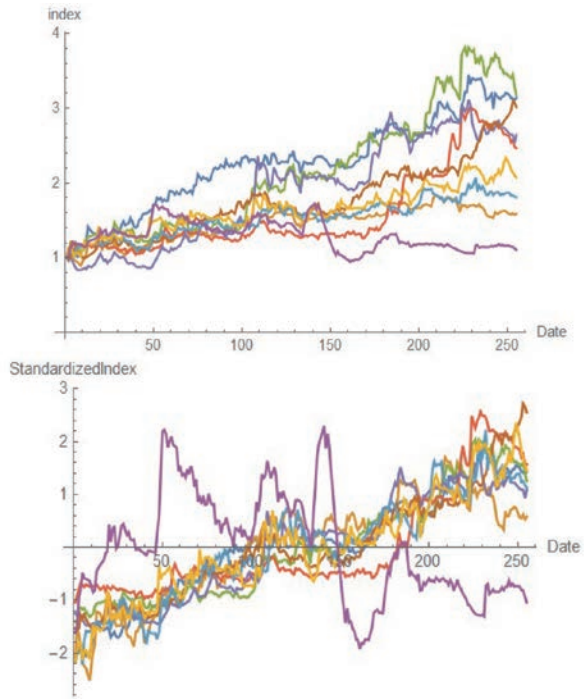

Fig. 1. Stock price movement of automkers in India from 2020/03/24 to 2021/03/22. The lower figure shows the standardized version of the upper figure

In the next section, we explain the method and the data we used in our experiment. In Section3, the clustering results by the two methods with three different data sets are presented. In Section 4, the results are evaluated. Finally, conclusion of the paper is stated in the final section.

2. DATA AND METHOD

The method and the data sets we used in the experiment is described below.

A. Data

We used real stock price data as the sample data. The data are the stock price data of manufacturing companies in India which were retrieved from ORBIS company database by Bureau van Dijk. Its last data update date is 2021/03/22. The missing data were linearly interpolated after removing every Saturday and Sunday. We used data of three manufacturing fields as follows

- Automakers (9 companies)
- Computer, electronic and optical products (8 companies)
- Electrical equipment (7 companies)

In Fig. 2, the top 9 automakers' data that we retrieved are illustrated. First, we took the top 250 automakers' data after removing the companies if they have over 50 day missing data during 2 years from 2019 Apr. to 2021 Mar. Finally the number of companies became just 9. Around the world, the manufactures had been severely damaged owing to the COVID-19 from Feb. in 2020. The worst stock prices were found around 24th March, 2020 from which many companies started the recovery. Therefore we selected 2020/03/24 as the recovery starting date. The vertical line in Fig. 2 shows the date. As we would like to compare the recovery speed and strength, we took the period from **2020/03/24 to 2021/03/22** and set 2020/03/24 to be 1. The resultant movement was shown in the upper figure in Fig. 1.

Fig.2. Stock price movement of the 9 automakers in India.

Regarding the field of computer and the electrical equipment, we took the same process to make the movements in the graphs as the upper figure in Fig. 1. When we analyze stock price data, in general, the input data is the matrix of natural logarithmic return values (hereafter return) which is defined as follows:

$$G_{i,j} = ln\left(\frac{S_{i,j}}{S_{i,j-1}}\right)$$

$S_{i,j}$ represents the i-th company's stock price on j-th day. In the individual company's return data, the standardization is conducted. In a case of stock prices, a standardization is required. In the stock data clustering for portfolio construction, in general, the input data are the **standardized return values.**

However, in this study, we did not use the return values; we used the raw stock values. The reasons are:

(1) We would like to observe shapes of the raw data fluctuation, not shape of the return values. It is difficult to read the return value sequence. Just from the sequence of the return values, it is too difficult for us to make the original stock movement.

(2) Our target data are various time series data. For example, a GDP growth rate and many key trend indicators are included.

In this study, we use the index fluctuation data as shown in the upper figure in Fig. 1.

B. Method

Here we compare the two clustering methods. They are (1) the HRP's clustering method and (2) Amplitude-based clustering with Euclidean distances. Both the methods use hierarchical clustering algorithms which are explained in [10, 11].

(1) HRP's clustering method with correlation coefficients

In stock price analysis, Prado's proposed HRP (Hierarchical Risk Parity) method has been widely used for portfolio construction, because HRP's outperformance over Markowitz's Critical Line Algorithm (CLA) and traditional risk parity's Inverse-Variance Portfolio (IVP) [6]. After HRP, many researches of a hierarchical clustering for portfolio had been conducted [12-15]. Therefore, we use the hierarchical clustering with the same distance as the distance in HRP. The distance is based on the Pearson's correlation coefficients [16, 17].

Given the matrix $\{G_{i,k}\}$ where is k is the parameter of date, we calculate the correlation coefficient matrix $\rho=\{\rho_{i,j}\}_{i,j=1,\ldots,N}$ where $\rho_{i,j}$ is the correlation coefficient between i-th and j-th companies.

In HRP, the distance d is defined as

$$d_{i,j} = \sqrt{\frac{1}{2}(1-\rho_{i,j})}$$

We adopted this distance definition using the correlation coefficients. Then the distance matrix $D=\{d_{i,j}\}_{i,j=1,\ldots,N}$ is obtained. Next, selecting any 2 distance columns, we calculate the Euclidean distance as follows:

$$\tilde{d}_{i,j} = \sqrt{\sum_{n=1}^{N}(d_{n,i}-d_{n,j})^2}$$

The distance $d_{i,j}$ represents the relationship between the two companies' data. On the other hand, $\tilde{d}_{i,j}$ represents a relationship between the two companies' data in the N dimensional world. We call $\{\tilde{d}_{i,j}\}$

distance-distance matrix.

We conduct the hierarchical clustering with the input as the distance-distance matrix $\{ \tilde{d}_{i,j} \}$. As HRP uses the single linkage method, we used the same metric, following the HRP. Then, using the distance between the nodes, we sort the distance matrix, which is called a matrix serialization; we sort the companies, so that small distance figures may be laid diagonally as much as possible. Prado calls the matrix serialization process “quasi-diagonalization” [2].

(2) Amplitude-based clustering with Euclidean distances

We explain the second hierarchical clustering method which is our proposed one. As the distance definition, we adopted the Euclidean distance

$$ED_{i,j} = \sqrt{\sum_{k=1}^{T} (G_{i,k} - G_{j,k})^2}$$

where $G_{j,k}$ is the index data of j-th company on k-th day and T is the number of days. Then the distance-distance is defined as the Euclidean distance as follows:

$$\widetilde{\mathrm{ED}}_{i,j} = \sqrt{\sum_{n=1}^{N} (ED_{n,i} - ED_{n,j})^2}$$

where N is the number of companies. As well as the HRP, we input the matrix $\{ \widetilde{\mathrm{ED}}_{i,j} \}$, conduct the hierarchical clustering and the matrix serialization.

In the next section, we compare the clustering results.

3. Comparison of two clustering results

In the section, two clustering methods are compared. The data we used are the India manufacturers’ stock prices. There are three fields (A)Automakers, (B) Computer, electronic and optical products, and (C) Electrical equipment manufactures.

A. Automakers

First, we illustrate results by the HRP’s clustering method with correlation coefficients (see Fig. 3). In Fig. 3, the distance-distance matrix heat maps are shown where the left is one before the serialization and the right is one after the serialization. The serialization makes white cells placed as near to the diagonal line as possible. The color white shows that the distance equals to 0. The red color shows a long distance. The serialized dendrogram is illustrated in the lower figure in Fig. 3. If the number of clusters k = 3, the three clusters are Cluster#0: {1,7,8,5,6,3,2}, Cluster#1: {4}, and Cluster#2: {9}.

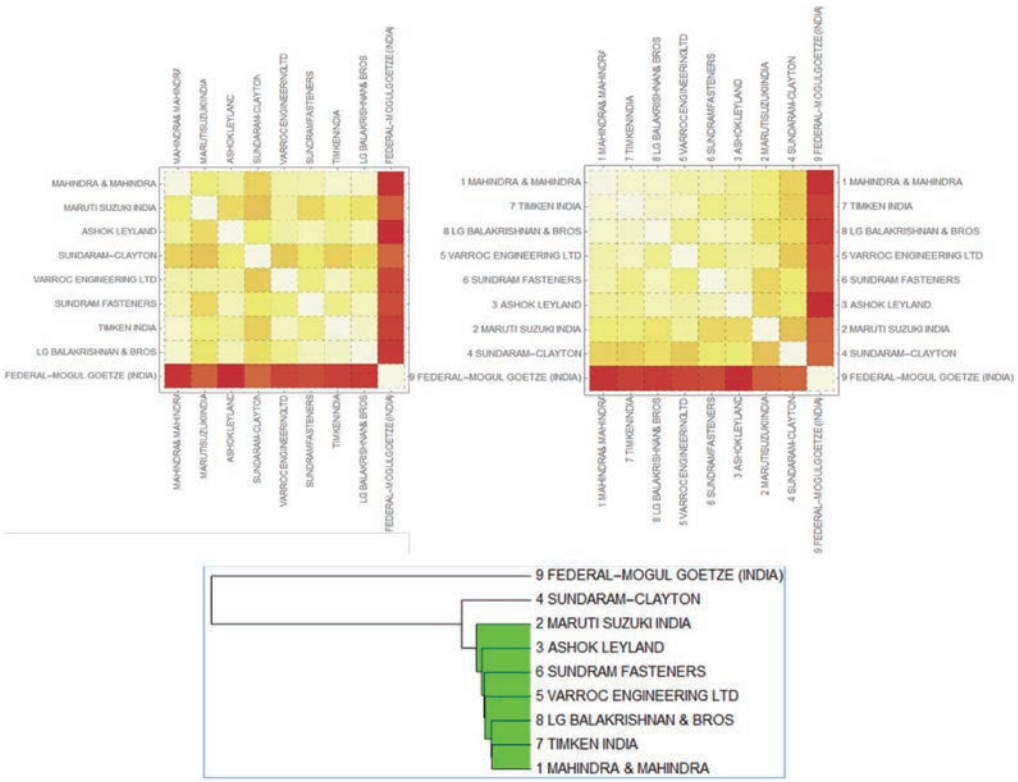

Fig. 3. Results by the HRP's clustering method with correlation coefficients

Next results by the Amplitude-based clustering with Euclidean distances are shown in Figure 4. If the number of clusters k = 3, the three clusters are Cluster#0: {FEDERAL-MOGUL}, Cluster#1: {MAHINDRA & MAHINDRA, ASHOK LEYLAND}, and Cluster#2: {others}.

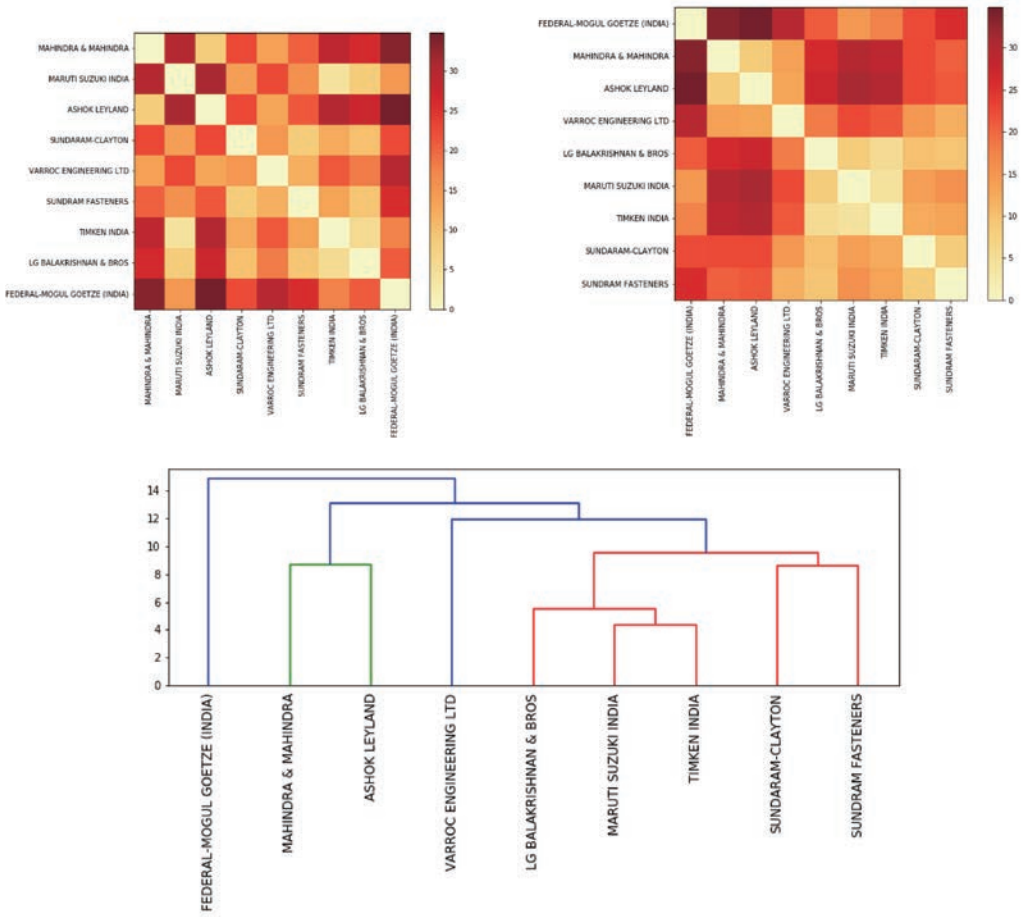

Fig. 4. Results by the Amplitude-based clustering method with correlation coefficients

Using the input indexed stock movements, we evaluated the two clustering results (see Fig. 5). The upper figure in Fig. 5 illustrates the result by HRP and the lower does the result by Amplitude-based clustering. The term “centroid” means an average movement of each cluster. In the HRP result, there are three clusters which are (1) a large cluster in black, (2) just one company in pink, and (3) just one company in blue. Cluster#1 does not meet our expectation, because we think that it should be involved in the large Cluster#0. On the other hands, the result by Amplitude-based clustering meets our expectations. The rapid growth two companies are separated as the cluster which are MAHINDRA & MAHINDRA and ASHOK LEYLAND. The results in Fig. 5 convinced us that the Amplitude-based clustering method with Euclidean distances could correctly evaluate both shape and amplitude.

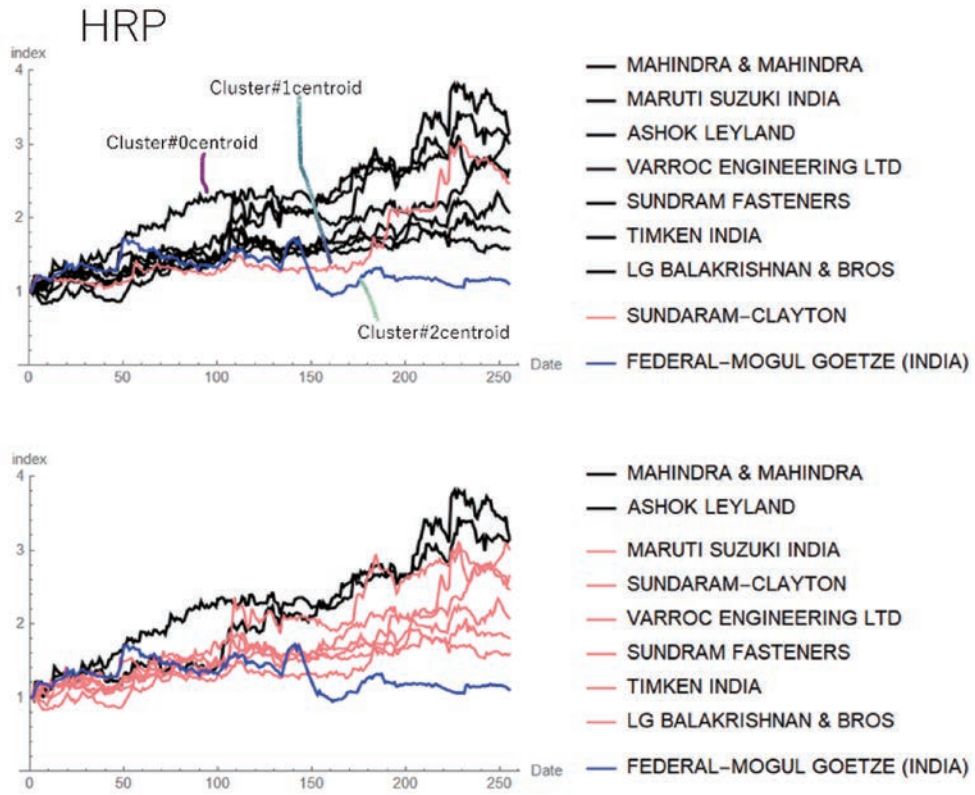

Fig. 5. Comparison of the two methods HRP and Amplitude-based clustering with the automakers' data.

B. Computer manufacturers

Next, we shall see the results on the computer manufactures' index data. In Fig. 6, the results by Amplitude-based clustering method are illustrated.

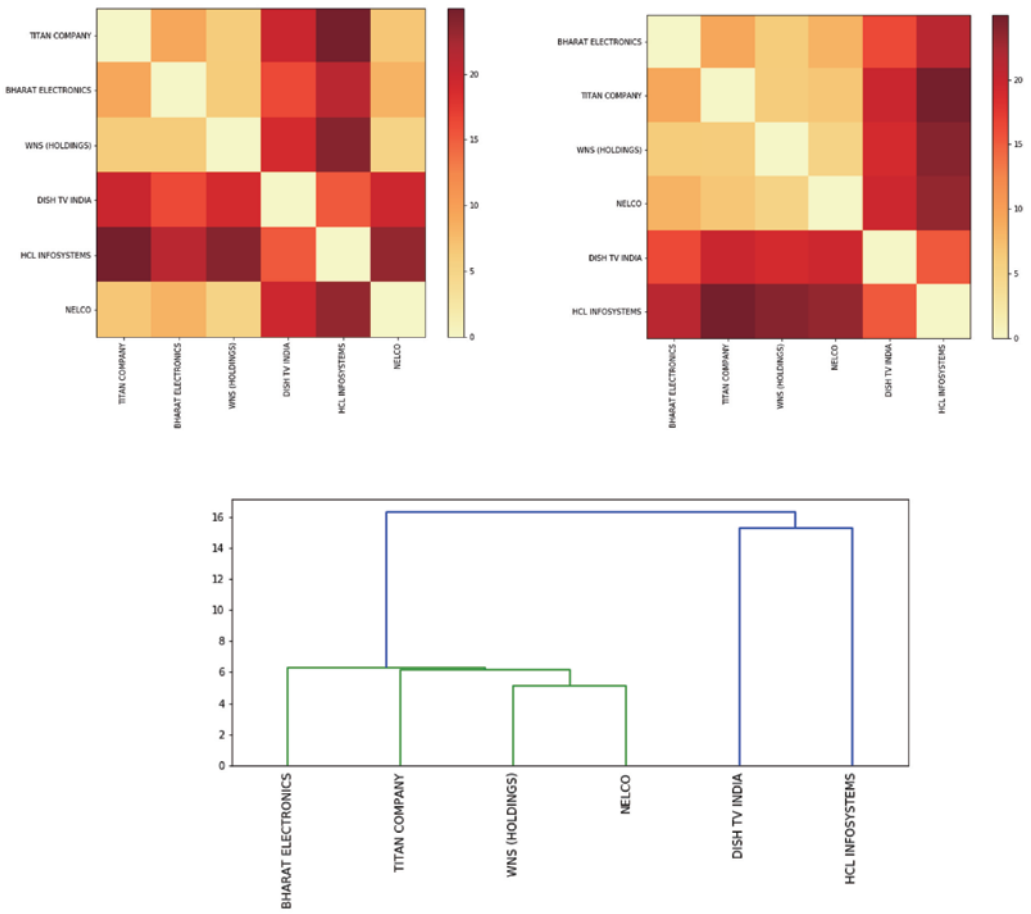

Fig. 6. Resultant heatmaps and dedtrogram by Amplitude-based clustering method with computer manufacturers' data.

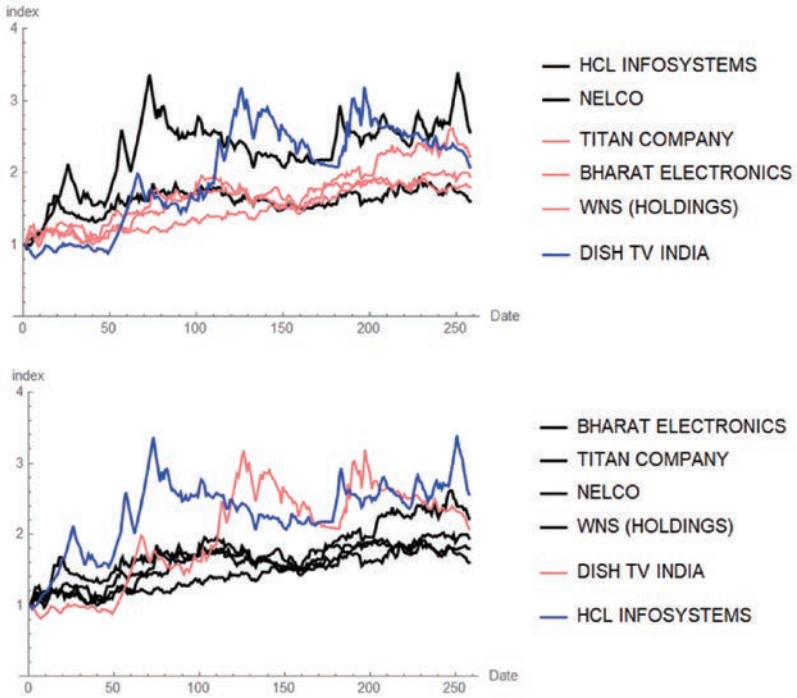

Fig. 7. Comparison of the two clustering methods with computer manufactures' data.

In Fig. 7, the indexed stock movement of the computer manufacturers' data is illustrated by the two clustering methods. The upper figure shows the HRP's result and the lower one does Amplitude-based clustering method. In the HRP result, the flat movement of NELCO and the higher growth HCL are members of the same cluster, which cannot meet our expectation. On the other hand, in the Amplitude-based clustering, HCL is separately classified as a cluster and NELCO becomes a member in the flat movement cluster with 3 other companies. These four companies have a similar movement which is clear in the dendrogram of Fig. 6. The second largest variance company DISH TV is separately selected as a cluster there. Therefore, Amplitude-based clustering is superior to the former.

C. Electrical equipment manufactures

Next, we explain the results on the electrical equipment manufactures' index data. The results by Amplitude-based clustering method are shown in Fig. 8. The resultant hierarchy is clearly ranked from AMBER ENTERPRISES INDIA to ORIENT ELCTRIC.

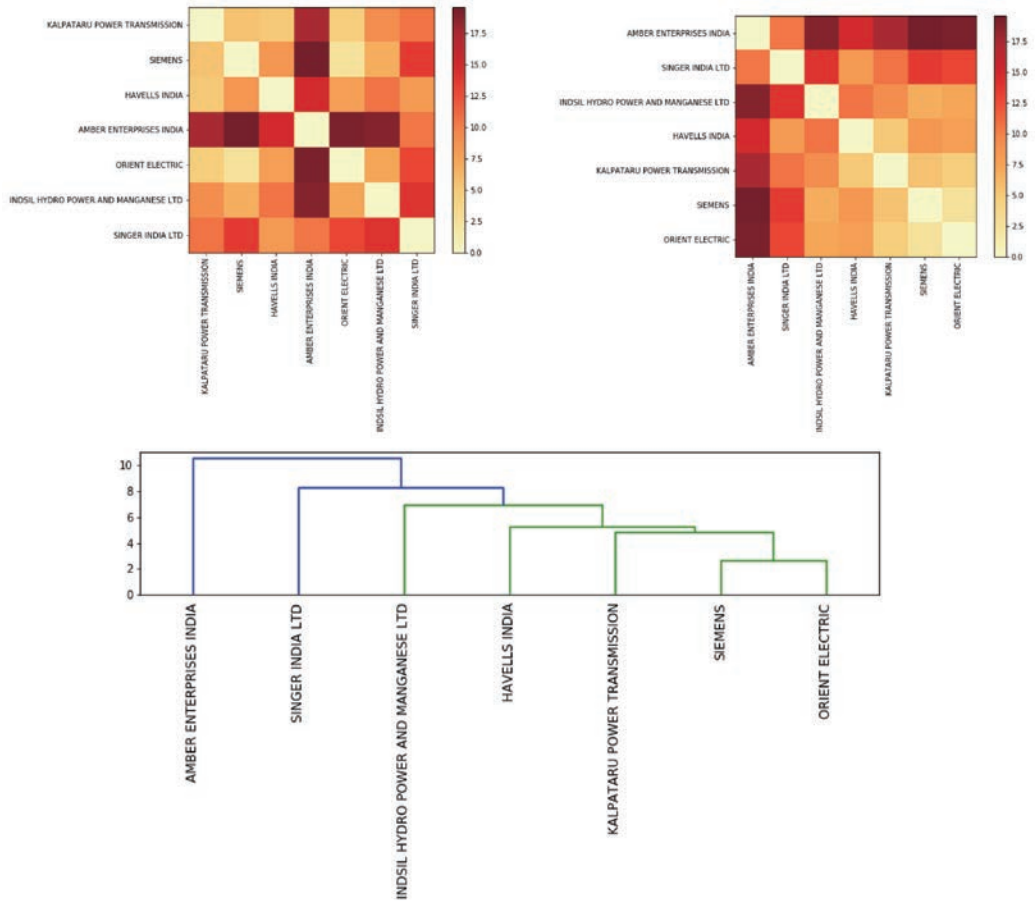

Fig. 8. Resultant heatmaps and dedtrogram by Amplitude-based clustering method with the electrical equipment manufacturers' data.

In Fig. 9, the comparison of the two methods in the indexed stock movement is shown. In the upper portion, result by HRP, the rapid growth of AMBER is not separated. On the other hand, in the result by Amplitude-based clustering, the largest AMBER and the second SINGAR INDIA are taken as the top two clusters which meets our expectations.

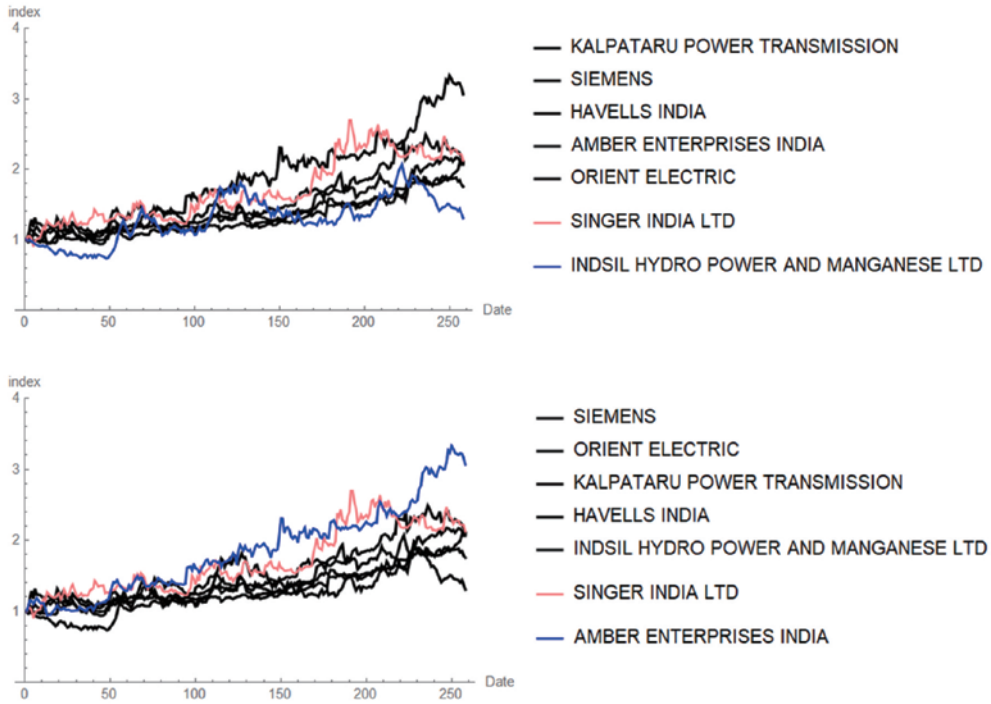

Fig. 9. Comparison of the two hierarchical clustering methods on electrical equipment manufactures.

4. Evaluation

In this section, we evaluate the results of the previous section. We compared the HRP clustering method by correlation coefficient-based distance and our proposed Amplitude-based clustering method. As a result, the latter, our proposed one is found to be suitable for the indexed movement comparison.

The reason is whether the standardization is conducted on the data or not. In the HRP's method, the correlation coefficients are used. The correlation coefficient uses in the definition the standardization where the covariance is divided by the two standard deviations. Therefore, the resultant clustering is the same as one when the input data is standardized, even if the input data is not standardized. On the other hand, Amplitude-based clustering method offers the clustering result that meets our expectation, because in the algorithm **no standardization** is conducted. The reason why we selected Euclidean distances is that Euclidean distance is simple and robust.

k-Shape method cannot be used for this indexed movement analysis. Shape-based distances (SBD) in k-Shape method cannot be used without standardization of the input data.

The important point of our proposed method is its distance-distance matrix. Using the distance-distance matrix, a better clustering result can be obtained, compared to that by simply using a distance matrix. The advantage of that is that they can consider the relationship of the whole data.

5. Conclusion

In this paper, we proposed the amplitude-based time series data clustering method named Amplitude-based clustering. Our target data are the growth rate fluctuations such as stock prices or GDP values. Shape-base time series data clustering methods such as k-Shape method requires standardization/ z-normalization of the input data. However, the standardization process removes the amplitude information from the original data. Hence, the flat fluctuation becomes a large variance movement after the standardization. That does not meet our requirement for the growth rate clustering. To keep the amplitude information, we adopted Euclidean distances as the distance. Because Euclidean distances are simple and robust, we can calculate that without standardization. As the clustering algorithm, we adopted the hierarchical clustering.

In addition, as the input to the hierarchical clustering, we used the Euclidean distance of the Euclidean distances, instead of the Euclidean distances. This is because not only the two patterns relationship but also all patterns information can be included in the result.

The experiments were executed on the India manufactures' stock price data. Three different data clustering were conducted and as a result the amplitude-based time series data clustering method seems to produce high performance; The rapid growth company groups and the slow growth company groups could be extracted from each data set. The results meet our expectations.

A still better algorithm might be found for the indexed movement clustering. However, the proposed clustering method meets at least our requirement. We hope that the Amplitude-based clustering method could be widely used in many kinds of key trend index analysis in economy.

Acknowledgment

The research was partly supported by JSPS KAKENHI Grant Number 20H01537.

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[2] M. L. De Prado, *Advances in financial machine learning.* John Wiley & Sons, 2018.

[3] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD workshop*, 1994, vol. 10, no. 16: Seattle, WA, USA:, pp. 359-370.

[4] R. B. Querino, R. C. D. MORAES, and R. A. Zucchi, "Relative warp analysis to study morphological variations in the genital capsule of Trichogramma pretiosum Riley (Hymenoptera: Trichogrammatidae)," *Neotropical Entomology*, vol. 31, no. 2, pp. 217-224, 2002.

[5] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," *Pattern recognition*, vol. 44, no. 9, pp. 2231-2240, 2011.

[6] M. L. de Prado, "Building diversified portfolios that outperform out of sample," *The Journal of Portfolio Management*, vol. 42, no. 4, pp. 59-69, 2016.

[7] M. L. de Prado, *Machine Learning for Asset Managers*. Cambridge University Press, 2020.

[8] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in

Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015, pp. 1855-1870.

[9] J. Paparrizos and L. Gravano, "Fast and accurate time-series clustering," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 2, pp. 1-49, 2017.

[10] K. Koutroumbas and S. Theodoridis, "Pattern Recognition (4th edition)," ed: Elsevier, 2009.

[11] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists* O'Reilly Media, 2016.

[12] N. Hautsch and S. Voigt, "Large-scale portfolio allocation under transaction costs and model uncertainty," *Journal of Econometrics*, vol. 212, no. 1, pp. 221-240, 2019.

[13] T. Raffinot, "Hierarchical clustering-based asset allocation," *The Journal of Portfolio anagement,* vol. 44, no. 2, pp. 89-99, 2017.

[14] J. Pfitzinger and N. Katzke, "A constrained hierarchical risk parity algorithm with cluster-based capital allocation," Stellenbosch University, Department of Economics, 2019.

[15] G. Konstantinov, A. Chorus, and J. Rebmann, "A Network and Machine Learning Approach to Factor, Asset, and Blended Allocation," *The Journal of Portfolio Management*, no. Multi-Asset Special Issue pp. 1-18, 2020.

[16] P. S. Mann, *Introductory statistics*. John Wiley & Sons, 2007.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second ed)*. Springer, 2009.