11th MALAYSIA STATISTICS CONFERENCE 2024
Data and Artificial Intelligence: Empowering the Future
Sasana Kijang, Bank Negara
19th September 2024

COMPARATIVE ANALYSIS OF TIME SERIES CLUSTERING: DYNAMIC TIME WARPING AND EUCLIDEAN DISTANCE MEASURE IN PRICE INDEX OF COMPANIES IN STANDARD AND POOR (S&P) 500 INDEX

Cheong Kah Ken[1]; Dr Norli Anida Binti Abdullah[2]; Dr Nur Anisah Binti Mohamed @ Abdul Rahman[1]; Dr Arief Gusnanto[3]

[1] Institute of Mathematical Sciences, Faculty of Science, University of Malaya, Malaysia
[2] Center for Foundation Studies in Science (PASUM), University of Malaya, Malaysia
[3] School of Mathematics, Faculty of Engineering and Physical Sciences, University of Leeds, United Kingdom (UK)

Abstract:
Time series data clustering is a crucial task in various fields, including pattern recognition and data segmentation. This research project employs k-means clustering in the price index of companies in Standard and Poor (S&P) 500 index with dynamic time warping (DTW) distances for centroid updating, comparing it with traditional Euclidean distance-based clustering. The comparative analysis reveals the differences in cluster formation and structure. Besides, this research project also investigated the distinct properties of specific clusters identified through DTW-based clustering, highlighting its effectiveness in capturing temporal similarities. Results from Rand Index (RI) and Davis-Bouldin Index (DBI) show that DTW-based clustering yields more coherent temporal patterns than that using Euclidean distance-based clustering, achieving more meaningful cluster formation and interpretation. This study is significant in financial analysis, stock market prediction, and investment strategies. By identifying temporal similarities in stock price movements, investors can make more informed decisions, identify market trends, and optimize their portfolios.

Keywords:
Dynamic Time Warping (DTW); k-means Clustering; Euclidean Distance; Standard and Poor (S&P) 500 Index

1. Introduction:
Time series clustering is a powerful technique for grouping similar time series data based on their temporal patterns, helping to uncover inherent patterns in various fields like finance, healthcare, and marketing. A common method used for clustering is the K-means algorithm (Steinhaus, 1949), which aims to group similar data points together while keeping clusters distinct.

In financial markets, analyzing stock prices is challenging due to volatility, temporal dependencies, and external influences. Time series clustering provides valuable insights by grouping similar stock price movements, aiding in informed investment decisions.

This research project aims to employ k-means clustering based on Dynamic Time Warping (DTW) with average centroid updating to leverage its ability to capture temporal similarities between stock price movements. The study focuses on the price index of the share prices in the United States stock exchange, as recorded by the Standard and Poor 500 (S&P 500) group of companies. DTW is chosen because it provides a more robust approach compared to traditional Euclidean distance measures, as it can account for shifts, stretches, and distortions in the time axis, which are common in stock price data.

The objectives of this research project are threefold: (1) to investigate the stock price index data extracted from companies listed in the S&P 500 group of companies using K-means clustering based on DTW distance with average centroid updating, (2) to compare the clustering results obtained with DTW against traditional Euclidean distance to understand the impact on cluster formation, and (3) to interpret the characteristics and temporal patterns within the clusters identified through DTW-based clustering within the companies in the S&P 500 index.

2. Methodology:

The data collected consists of historical stock prices and company sectors for S&P 500 firms. Historical stock prices from 8th February 2013 to 8th February 2018 were sourced from Kaggle, featuring variables such as Date, Open, High, Low, Close, Volume, and Name. Sectors were obtained from Stock Analysis, dividing companies into eleven sectors based on the Global Industry Classification Standard (GICS).

Next, data filtering and transformation was performed. The dataset was filtered to retain only Date, Close, and Name, with 35 companies removed due to unequal time series lengths. Stock prices were transformed into index numbers to standardize the starting point across companies, using the formula

$$I_t = \frac{\text{Stock Price at time } t}{\text{Stock Price at time } 1} \times 100, t = 2,3,4, \ldots \quad (1)$$

where $I_1 = 100$

This formula is modified from (Wong, 2019) to ensure that all the companies have the same starting point.

Simple Moving Average (SMA) was used for data smoothing to prevent overfitting. The optimal window size for the SMA was determined through cross validation. For each stock, we divide the data into initial training period of 200 time points and a testing period comprising about 15% of the data and shifted forward by the set window size. The root mean square error is used to measure predictive accuracy. This process is iterated across all stocks with window sizes ranging from 2 to 40. The optimal window size for each company is determined by the lowest RMSE.

Three k-means clustering algorithms were implemented based on DTW-Average, DTW-Median, and Euclidean distance measures. For each algorithm, initial k centroids were randomly selected at first. Next, the distances between each time series and centroids were calculated. Centroids were updated by averaging or taking the median of each time

step across assigned time series and the centroid updating process was repeated until there's only minimal changes in the centroids.

Furthermore, the elbow plot and silhouette plot were used to identify the optimal number of clusters. The elbow method identified the optimal number of clusters by plotting within-cluster sum of squares (WCSS) against the number of clusters, seeking the point where the rate of decrease sharply changes whereas silhouette plot used the silhouette scores to evaluate the clustering quality, with values ranging from -1 to 1, where higher scores indicate better-defined clusters (Rousseeuw, 1987).

For model performance evaluation, the Rand Index (RI) (Hubert & Arabie, 1985) and Davis-Bouldin Index (DBI) (Davies et al., 1979) were used. RI compared the similarity between clustering results, with higher RI values indicating greater similarity. It calculated the ratio of true positive (TP) and true negative (TN) decisions to all decisions. DBI assessed clustering quality by comparing the average similarities between the most similar clusters. Lower DBI values indicated well-separated and compact clusters. The most acceptable model will be further analysed.

The methodology is summarized in Figure 2.1 below:

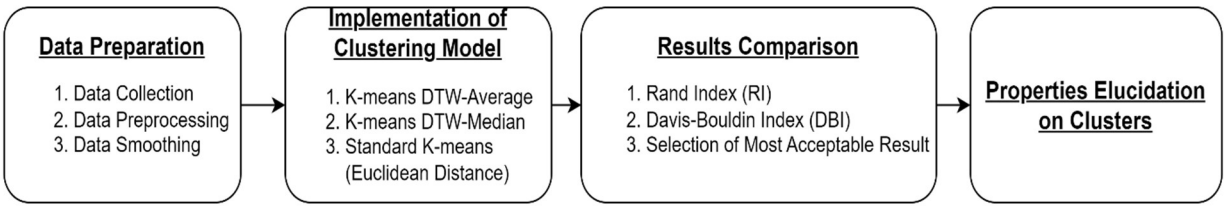

Figure 2.1: Summary of Methodology

3. Result:

The median of the optimal window sizes will be taken as the optimal window size and applied to the SMA because it is a more robust choice for data with potential outliers, as it's less likely to be pulled towards extreme values. In this case, the median is around ten. So, SMA-10 will be applied to all the time series data for smoothing. to outliers as compared to the mean. In this case, the median is around ten. So, SMA-10 will be applied to all the time series data for smoothing.

The optimal number of clusters from the three clustering methods are shown in Table 3.1.

Table 3.1: Summary of Optimal Number of Clusters

Method	Optimal Number of Clusters		
	DTW-Average	DTW-Median	Euclidean
Elbow Plot	3	4	4
Silhouette Plot	2	2	2

The optimal number of clusters suggested by the silhouette plot is not meaningful. This is because it is overly simplistic. Two clusters will essentially divide the stocks data into "up" and "down" trends. This provides limited insight into the complex dynamics of stock prices.

Besides, with only two clusters, significant patterns and relationships between stocks might be hidden, leading to decision-making based on incomplete information. In addition,

two clusters might be unable to capture the patterns present in stock price movements accurately. Most stocks will have unique characteristics that are not well-represented in just two groups. So, three or four number of clusters are more acceptable in this case.

Now, the RI and DBI based on the results from elbow plot are shown in Table 3.2.

Table 3.2: Summary of Model Performance

Method	DTW-Average	DTW-Median	Euclidean
Optimal Number of Clusters	3	4	4
RI	1.00	0.83	0.96
DBI	0.46	1.23	0.87

The result shows that three clusters using k-means clustering based on DTW-Average provide the most acceptable result as compared to the other two. This is because it has the highest value of RI and the lowest value of DBI. Moreover, it is also the most appropriate to choose the result based on DTW-Average as DTW distance is able to consider the varying speed that occurs in the time series data whereas Euclidean distance only treats all time points equally.

Next, the results from DTW-Average were further elucidated. The centroids of each cluster are shown in Figure 3.1.

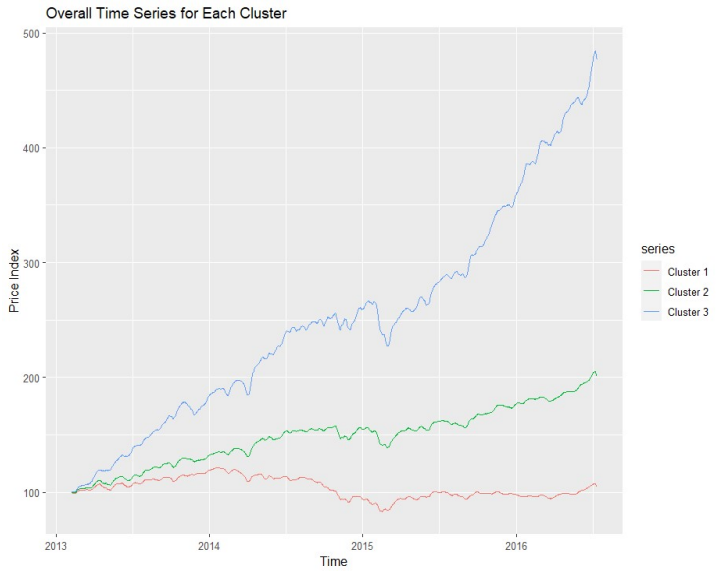

Figure 3.1: Centroids of Each Cluster

From the result, the companies could be categorized into high, moderate and low investment return clusters. Next, all these clusters are in an upward trend. Then, there is a sharp decline in price index around the first quarter of 2015. This is most probably due to three main factors which are the global economic slowdown, the drop of oil price in 2014, and the ending of quantitative easing in the last quarter of 2014.

The S&P 500 companies are categorized into eleven distinct sectors according to the Global Industry Classification Standard (GICS). The proportion of the company sectors for each cluster are shown as pie chart in Figure 3.2.

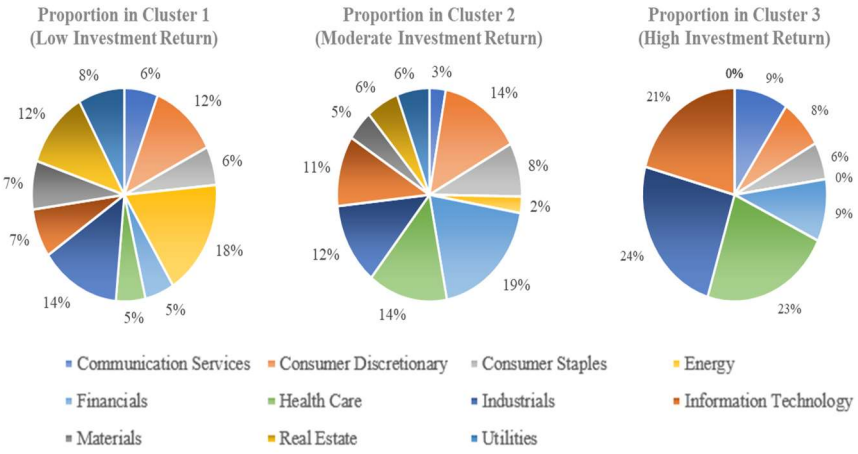

Figure 3.2: Sectorial Breakdown for Each Cluster

From the observation in Figure 3.2, certain sectors are categorized into certain clusters. To understand whether there is a difference in the proportion of industrial sectors in the clusters, we performed Chi-squared test for association (Pearson, 1900).

The frequency of the companies based on their sectors and assigned clusters respectively are tabulated as a contingency table. The result shows that the Chi-squared test statistic, χ^2 equals to 96.85 and the critical value $\chi^2_{0.05,20}$ equals to 31.41. Since the Chi-squared test statistic is greater than the critical value, the null hypothesis is rejected at five percent level of significance. Thus, there is significant association between the company sectors and their assigned clusters. This also means that certain clusters will tend to focus on certain sectors, which verifies the observation. The clustering result is meaningful.

4. Discussion and Conclusion:

This research focuses on analyzing stock clusters with high and low investment returns to provide significant insights into investment performance and trends, specifically excluding the moderate return cluster.

The high return cluster includes sectors such as Health Care, Information Technology, and Industrials. For Health Care Sector, its high returns are attributed to policy changes and technological advancements. The Affordable Care Act (ACA) significantly increased access to health care services, benefitting companies like Aetna Inc. and UnitedHealth Group Inc. These companies saw growth due to increased enrolments and strategic adaptation to new regulations. Technological breakthroughs, such as Boston Scientific's WATCHMAN device and Vertex Pharmaceutical's cystic fibrosis treatments, also drove substantial revenue growth and market expansion.

Next, Information Technology Sector's growth was driven by advancements in cloud computing and a gaming boom. Companies like Amazon.com Inc. and Microsoft Corporation experienced significant revenue increases due to rising demand for secure cloud solutions. NVIDIA Corporation benefited from the growing popularity of gaming and the application of its GPUs in AI and data centres, leading to a 55% revenue increase in 2017.

Industrials Sector, particularly the aerospace and defense subsectors, saw high returns due to consistent global demand. Boeing Company and Lockheed Martin Corporation exemplified this growth, with Boeing achieving record airplane deliveries and Lockheed Martin benefiting from increased defense contracts and the success of the F-35 program.

The low investment return cluster is dominated by the Energy and Industrials sectors. The oil price crash in 2014 had a profound impact on oil and gas companies like Exxon Mobil Corporation and Chevron Corporation in the Energy sector. The significant decline in oil prices led to reduced revenues, capital expenditures, and long-term growth prospects. Consequently, both companies reported substantial revenue decreases in 2015, highlighting the sector's vulnerability to oil price fluctuations.

Moreover, Industrials sector's performance was adversely affected by global economic conditions, particularly in emerging markets. Caterpillar Inc., for instance, faced slow growth in markets like China and Brazil, coupled with reduced commodity prices and significant restructuring costs. These challenges led to a notable decline in sales and revenues, exacerbated by the oil price drop that severely impacted their Energy & Transportation segment.

By identifying clusters representing high, moderate, and low investment returns, the study offers valuable sector-specific insights, guiding investors to optimize their portfolios for better returns and manage risks more effectively to hedge against market volatility. Additionally, the findings provide policymakers with data-driven insights to support struggling sectors through targeted economic policies, such as subsidies and research funding, thereby enhancing overall economic stability. The research also aids corporate strategies, enabling companies to benchmark performance and reassess strategic directions, particularly in underperforming sectors.

References:

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224-227.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software*, *31*, 1-24.

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

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157-175.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Steinhaus, H. (1949). Sur la division pragmatique. *Econometrica: Journal of the Econometric Society*, 315-319.

Wong, M. K., Musa, Z., Ahmad, S., Ahmad Zaki, N., & Burham@Borhan, Z. H. (2019). *KSSM Form 4 Additional Mathematics Textbook* (I. Ibrahim, N. M. Zakaria, W. F. Wan Ismail, & N. S. Osman, Eds.; C. H. Yew & C. M. Lee, Trans.). Penerbitan Pelangi Sdn. Bhd.