DOI: 10.62241/emd.22.0114.2339 Vol. 2, No. 2, year 2024, pp. 01-14

Contents lists available at Scientific Research Discovery

International Journal of Economic and Management Decisions (EMD)

Journal homepage: scird.org/ijemd

Exploring Clustering Techniques for Time Series Analysis: A Case Study on Moroccan Stock Market Data

Mahboub Sabah[a,b], Raby Guerbaz[a]

[a] Department of Modeling Applied to Economics and Management, University Hassan II Faculty of Judicious Economies and Social Sciences, Ain Sebaa, Casablanca, Morocco.

[b] National Institute of Statistics and Applied Economics, Rabat, Morocco.

Article Info

Article history:

Received: 09-04-2024
Revised: 11-05-2024
Accepted: 26-05-2024

Keywords:

Financial analysis
Stock market
Clustering techniques
Time series analysis
Euclidean distance
Dynamic Time Warping (DTW)
Market dynamics
Financial data
Investment strategies

ABSTRACT

This study examines the effectiveness of clustering techniques for analyzing time series data in the context of the Moroccan stock market, focusing on 50 companies. The research utilizes Euclidean and Dynamic Time Warping (DTW) clustering methodologies to endeavor to discern latent trends and patterns in the oscillations of stock values across diverse firms.

The strategy consists of multiple crucial steps. The time series data is first preprocessed to ensure its consistency and dependability. Then, it is determined how many clusters would be most effective in grouping the data. The DTW distance, which identifies nonlinear similarities in time series data, and the Euclidean distance, which emphasizes similarities in pricing patterns, are then used to cluster the data.

The main conclusions of the study point to certain clustering tendencies in the Moroccan stock market dataset. While DTW clustering finds further details in nonlinear patterns that older approaches can miss, Euclidean clustering reveals groups of companies having similar price movements.

These discoveries significantly expand on our understanding of the mechanics and composition of the Moroccan stock market. By using appropriate clustering algorithms, investors and financial analysts can gain valuable insights into market behavior and potentially more precisely pinpoint investment possibilities. The study highlights the need of selecting suitable analytical methods for time series data, especially in financial contexts where the ability to identify subtle patterns and trends is crucial for making well-informed decisions. Ultimately, our study advances academic knowledge and offers practical advice to professionals working in the financial and investing domains.

Corresponding Author:

Mahboub Sabah
Department of Modeling Applied to Economics and Management, University Hassan II Faculty of Judicious Economies and Social Sciences Ain Sebaa, Casablanca, Morocco.
Email: mahboubsabah97@gmail.com

1. INTRODUCTION

Clustering analysis, a basic tool in financial analysis, exposes the dynamics and structures embedded in time series data.Researchers and analysts use clustering to group similar data points and identify latent patterns,

trends, or anomalies missed by examining them separately. Now, with clustering as a powerful data taking tool, it still suffers from choice of distance metric between individual time-series points if we intend to cluster such points based on similarity.

Time series analysis in financial markets is a discipline that has been the subject of extensive study to comprehend movements in price, market behavior and volatility. The initial work of Box and Jenkins [1] laid a baseline for time series model by introducing ARIMA model which is commonly used in forecasting financial records. Later on, Engle [2] built on these ideas when he came up with autoregressive conditional heteroskedasticity models (ARCH) aimed at capturing time varying volatilities within financial arena.

Machine learning's subset, the clustering techniques are now quite useful in the analysis of financial time series. Jain [3] comprehensively reviews clustering methods and demonstrates the relevance of these methods in other fields such as economics. In the financial industry, the earliest application of clustering was done for stock selection purposes using hierarchical clustering as demonstrated by Mantegna [4] who classified stocks according to their historical trading patterns. This technique has shown that clustering can reveal hidden structures within financial data.

Even though they are frequently employed, traditional distance measurements like the Euclidean distance frequently perform poorly when applied to financial time series data. Their incapacity to fully represent the many temporal correlations and inconsistencies present in financial data is the source of this shortcoming [5]. Therefore, there is an urgent need for more advanced distance measures that can take these subtleties into account and improve the precision and consistency of clustering results. For instance, the use of dynamic time warping (DTW) for clustering financial time series has shown promise in aligning and comparing temporal patterns in stock prices [6].

Apart from the difficulty of choosing a distance metric, choosing the ideal number of clusters continues to be a major problem in cluster analysis. The interpretability and usefulness of the clustering results are significantly impacted by this process, which is called cluster validation or selection. Informed decision-making processes may be hampered by the insights obtained from clustering if the right number of clusters is not well understood [7].

Researchers have created a wide range of methods and strategies to improve the effectiveness of clustering analysis in financial situations in response to these difficulties. These include the use of external validation metrics, like the adjusted Rand index, for comparing clustering solutions against ground truth labels or external criteria, and the application of internal validation metrics, like the Davies-Bouldin index and silhouette coefficient, to evaluate the quality of the clustering [8], [9].

Furthermore, exploratory data analysis approaches, such as visualization methods and cluster stability analysis, offer valuable tools for acquiring deeper insights into the underlying structure and stability of clusters [3]. Through the application of these methodologies, scholars can progressively enhance their clustering strategies, guaranteeing congruence with the distinct attributes of the dataset and the analysis's goals.

2. METHODOLOGIE

2.1. Data Acquisition and Preprocessing:

The first step in the research process is obtaining financial time series data from a CSV file that includes the daily stock values for several companies over a given time frame. The dataset is preprocessed after acquisition to improve analysis readiness and alleviate data quality concerns. Among these preparation procedures are data normalization and management of missing values. Data integrity is ensured by substituting the mean of the corresponding column for any missing values. The dataset's numerical properties are then standardized to a shared scale to enable efficient clustering analysis.

```r
# Read data from the CSV file
stock_data <- read.csv(file = "stock_data.csv")

# Convert Date column to Date type
stock_data$Date <- as.Date(stock_data$Date, format = "%m/%d/%Y")

# Replace NA values with the mean of the respective column
stock_data[is.na(stock_data)] <- sapply(stock_data, function(x) mean(x, na.rm = TRUE))

# Normalize the data
stock_data[, -1] <- scale(stock_data[, -1])
```

Figure 1. Data Preprocessing

2.2. Algorithm Selection and Clustering Analysis:

The algorithm selection process plays a major role in the clustering analysis, affecting the way financial time series data are organized into relevant clusters. Two crucial factors taken into account in this procedure are the choice of a distance metric and a clustering technique. The chosen distance metric which can be either Euclidean or Dynamic Time Warping (DTW) distance determines the technique for calculating the degree of similarity between time series data points.

❖ **Explanation of Algorithms and Techniques**

➢ **Partitioning Around Medoids (PAM)**

PAM is a clustering technique that minimizes the sum of dissimilarities between points and their nearest medoids. It follows these steps:

1. **Initialization**: Select k initial medoids from the dataset.
2. **Assignment**: Assign each data point to the nearest medoid based on a chosen distance metric, such as Euclidean distance. The assignment step can be expressed as:

$$C(i) = \underset{j}{argmin}\{d(x_i, m_j)\}$$

where $d(x_i, m_j)$ is the distance between data point and medoid m_j, and $C(i)$ is the cluster assignment of x_i [10].

3. **Optimization**: Iteratively swap medoids with non-medoids and reassign points to the nearest medoid if the swap reduces the total cost:

$$Cost = \sum_{i=1}^{n} d(x_i, m_{C(i)})$$

4. where n is the number of data points.
5. **Convergence**: Repeat the optimization step until no further reduction in cost is possible.

➢ **Dynamic Time Warping (DTW)**

DTW is a distance metric that calculates similarity between two time series by allowing for non-linear alignments. It follows these steps:

1. **Cost Matrix Construction**: Create a cost matrix *D* where each cell $D(i,j)$ represents the distance between points x_i and y_j:

$$D(i,j) = (x_i - y_j)^2$$

2. **Path Finding**: Find the optimal path through the matrix that minimizes the cumulative distance. The DTW distance is calculated as:

$$DTW(X,Y) = \underset{\pi}{min}\left\{\sum_{(i,j)\in\pi} D(i,j)\right\}$$

where π is a path through the matrix [11].

➢ **DTW Barycenter Averaging (DBA)**

DBA averages a set of time series under DTW distance. The process involves:

1. **Initialization**: Choose an initial template sequence T.
2. **Alignment**: Align each time series *X* to the template *T* using DTW.
3. **Template Update**: Update each point T_i in the template by averaging the aligned points from all series:

$$T_i = \frac{1}{k}\sum_{j=1}^{k} X_{a_j(i)}$$

where k is the number of time series and $a_j(i)$is the aligned index of X_j to T_i.

4. Iteration: Repeat alignment and updating until the template stabilizes [12].

❖ **Euclidean Distance**

Euclidean distance measures the straight-line distance between two points in Euclidean space. For time series, it is calculated as:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where *x* and *y* are the time series, and *n* is the length of the series [13].

❖ **Clustering Analysis**

Rather than depending solely on the PAM algorithm, we use a variety of clustering strategies. We use PAM for Euclidean distance and the DBA algorithm for DTW distance. This comprehensive method allows us to consider the unique qualities of financial time series data and investigate clustering patterns from multiple angles.

❖ **Determining the Number of Clusters**

Another crucial stage in the clustering analysis process is determining the number of clusters. To find the ideal number of clusters, we apply the gap statistic method instead of relying solely on within-cluster sum of squares (WSS). This approach provides a more thorough evaluation of clustering quality by considering the within-cluster dispersion distribution compared to a reference distribution. The steps are:

1. **Calculate Dispersion**: For each potential number of clusters k, calculate the within-cluster dispersion W_k:

$$W_k = \sum_{i=1}^{k}\sum_{x \in C_i} d(x, m_i)$$

   where C_i is the *i-th* cluster and m i is the medoid of C_i
2. **Generate Reference Data**: Create B reference datasets $\{X^b\}_{b=1}^{B}$ with similar distributions to the original data.
3. **Gap Statistic Calculation**: Compute the gap statistic for each k:

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log\left(W_k^b\right) - log(W_k)$$

   where W_k^b is the within-cluster dispersion for the *b-th* reference dataset.
4. **Optimal Cluster Number**: Identify the number of clusters *k* where the gap statistic is maximized, indicating the best balance between cluster complexity and meaningfulness [14].

   In our analysis, the gap statistic method suggests that 5 clusters provide an optimal balance. Subsequently, we perform clustering using both Euclidean distance and DTW distance with the PAM and DBA algorithms, respectively. This approach allows us to explore and compare the clustering structures and patterns within the financial time series data more comprehensively, ensuring robust and insightful clustering results.

```
# Determine the optimal number of clusters using the gap statistic method
fviz_nbclust(close_prices, pam, method = "gap_stat") +
  labs(title = "")

# According to the graph we got 5 is the optimal number of clusters
k <- 5
```

Figure 2. Code to plot Gap Statistic Method and Determining Optimal Number of Clusters

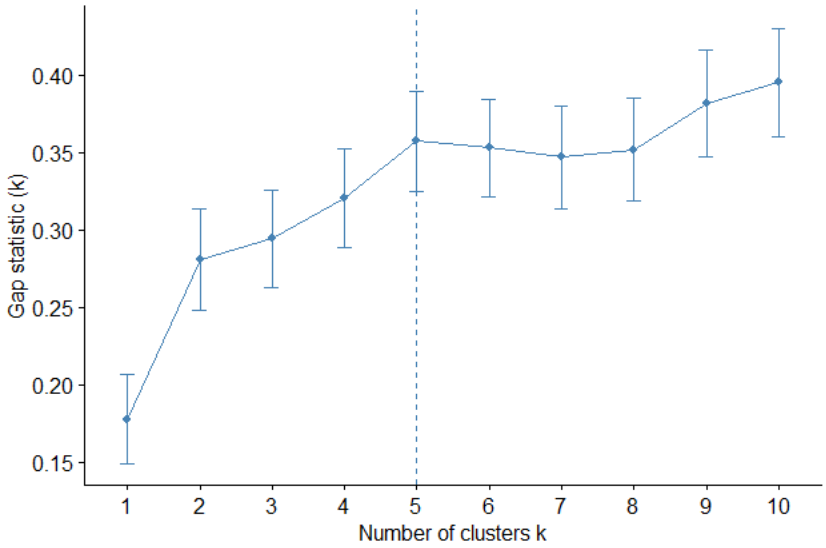

Figure 3. Gap Statistic Method for Determining Optimal Number of Clusters

```
# Perform clustering using Euclidean distance
eu_cluster <- tsclust(close_prices, type = "partitional", k,
                      distance = "Euclidean", centroid = "pam", seed = 1234,
                      trace = TRUE)

# Perform clustering using DTW distance
dtw_cluster <- tsclust(close_prices, type = "partitional", k,
                       distance = "dtw_basic", centroid = "dba", seed = 1234,
                       trace = TRUE)
```

Figure 4. Code to Perform both EU and DTW clustering

2.3. Evaluation Metrics for Clustering Algorithms

In this review, we will study how different computers work. What are some of the consumer interfaces that LDA distributions involve? Specifically, what are the (parameters to the) user interfaces involved? Which human sensors are important for planning? (What kind of human sensors are important for planning?).

❖ **Silhouette Score**

The Silhouette Score measures how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, with higher values indicating better-defined clusters.

➢ **Formula:**

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

The average distance between the sample and all other points in the same cluster is represented by *a(i)* while the average distance between the sample and all points in the next nearest cluster is represented by *b(i).*

➢ **Implementation:**

```
# Function to calculate Silhouette Score
silhouette_score <- function(data, cluster_labels) {
  silhouette_values <- silhouette(cluster_labels, dist(data))
  mean(silhouette_values[, 3])
}
```

Figure 5. Code for Silhouette Score function

❖ **Rand Index**

The Rand Index measures the similarity between the predicted and true clusters. Higher values indicate better agreement.

➢ **Formula:**

$$RI = \frac{a + b}{\binom{n}{2}}$$

Where *a* and *b* are the number of pairs of elements that are correctly clustered together or apart, respectively.

➢ **Implementation:**

```
# Function to calculate Rand Index
rand_index <- function(cluster_labels1, cluster_labels2) {
  rand.index(cluster_labels1, cluster_labels2)
}
```

Figure 6. Code for Rand Index function

3. RESULTS AND DISCUSSION

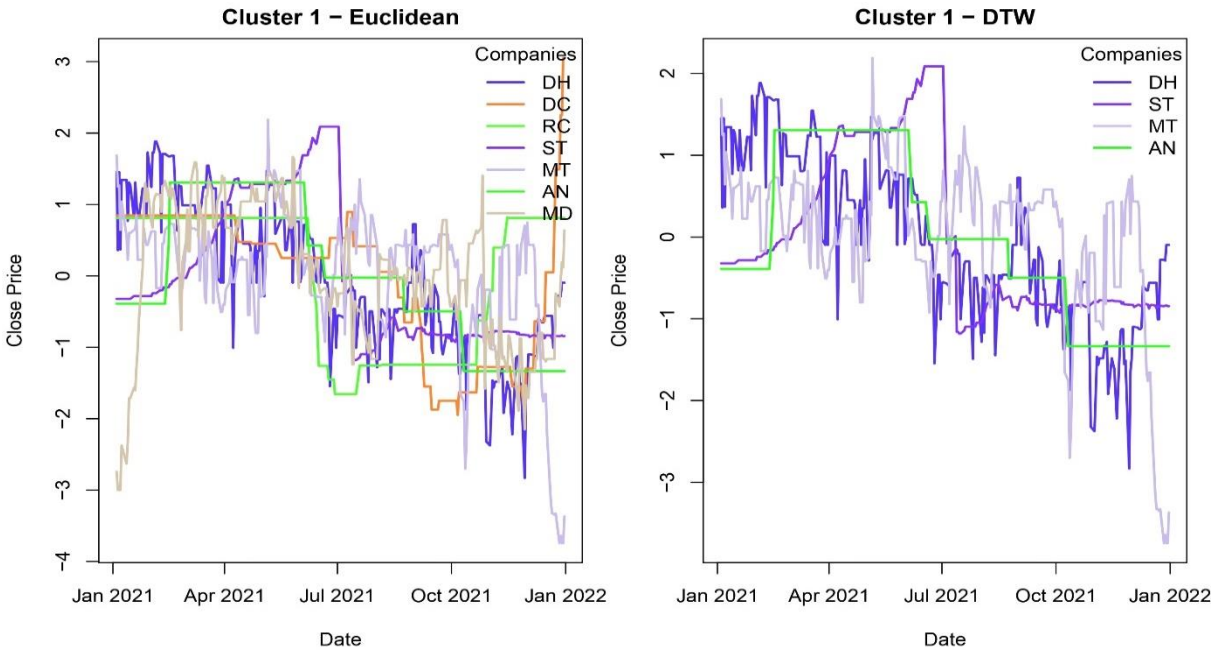

Figure 7. Cluster 1 for both eu clustering and dtw clustering

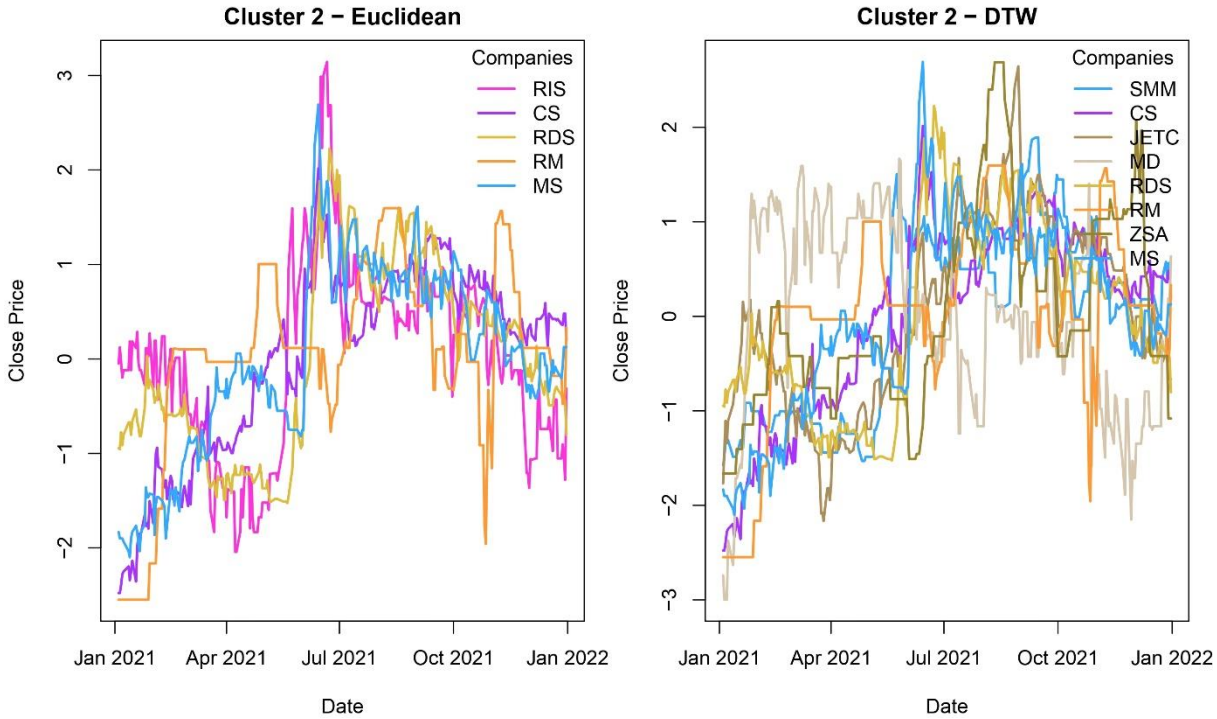

Figure 8. Cluster 2 for both eu clustering and dtw clustering

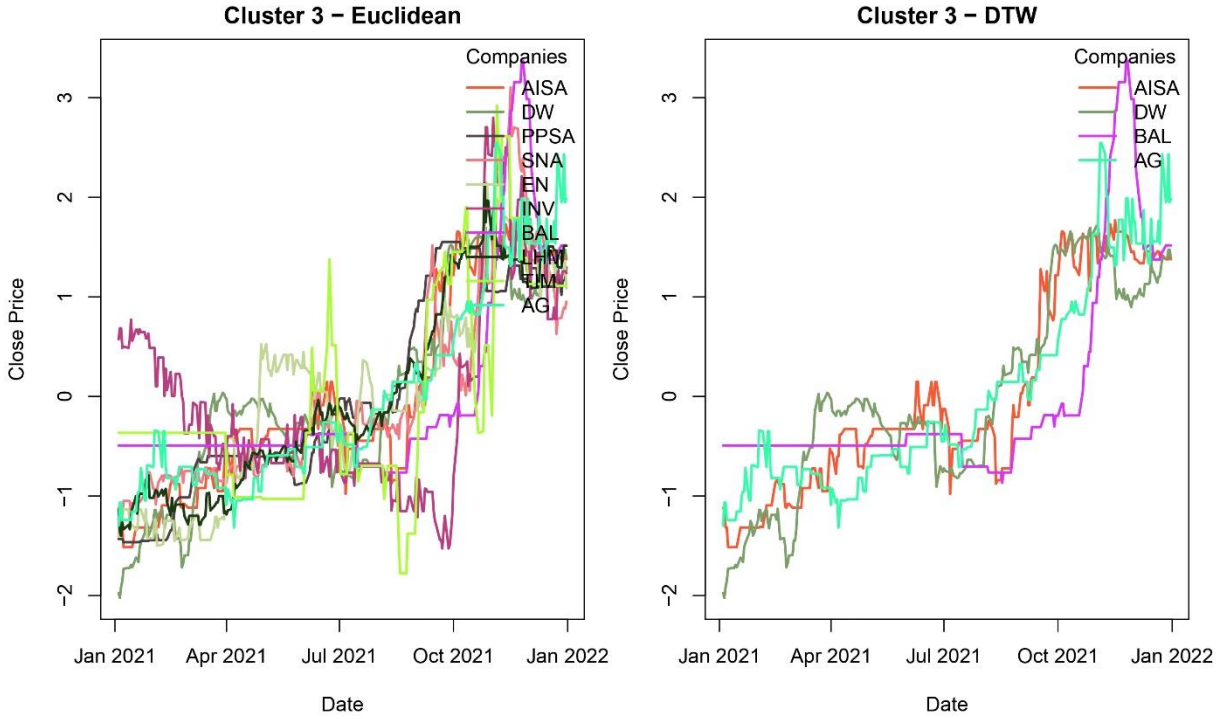

Figure 9. Cluster 3 for both eu clustering and dtw clustering

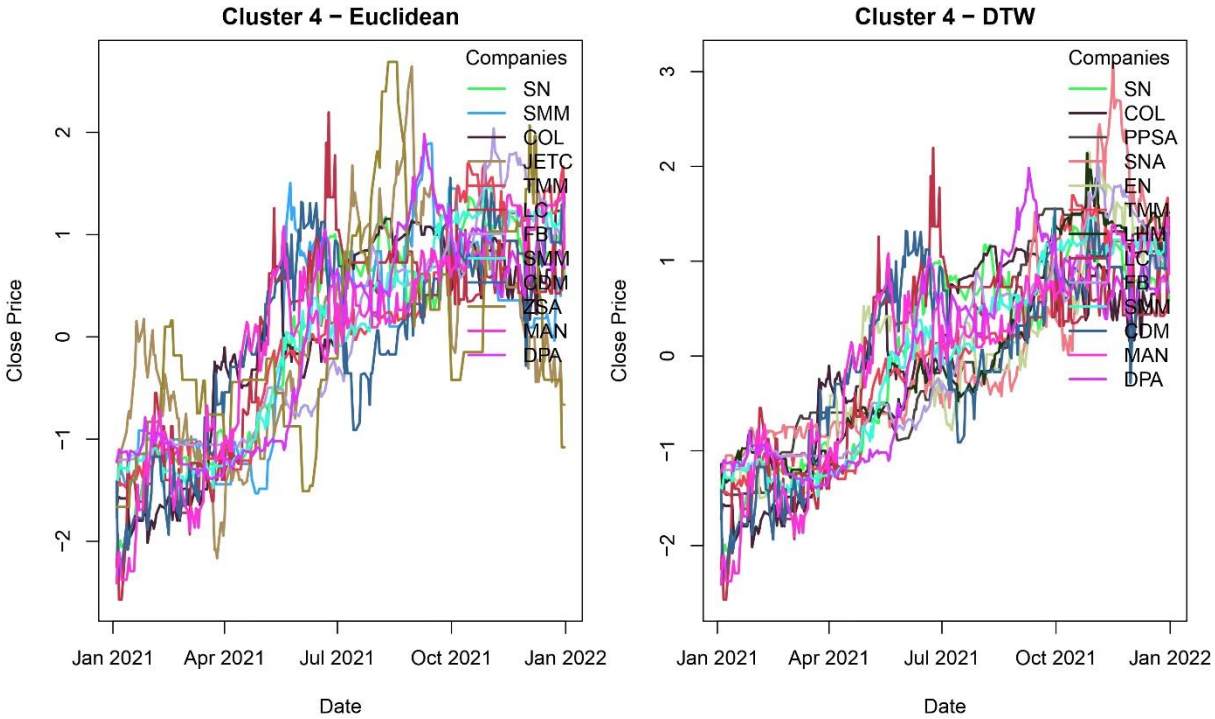

Figure 10. Cluster 4 for both eu clustering and dtw clustering

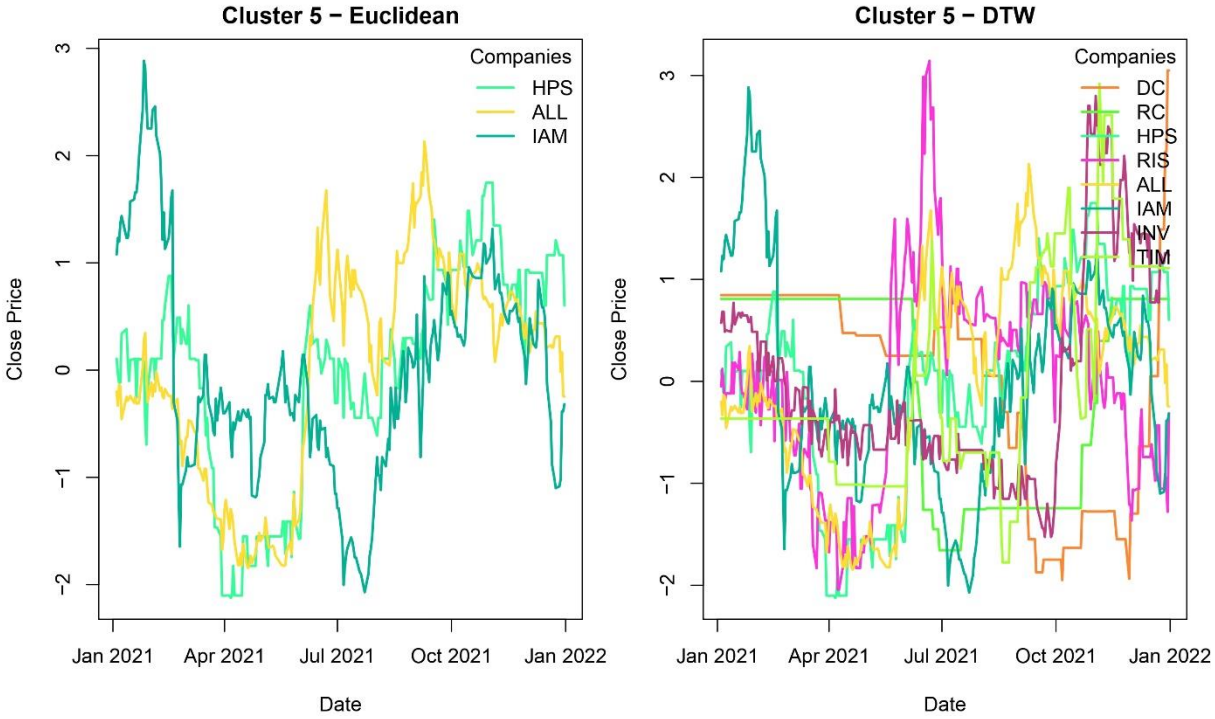

Figure 11. Cluster 5 for both eu clustering and dtw clustering

By utilizing clustering algorithms on data from Moroccan financial series, researchers have discovered interesting revelations regarding patterns of numbers and arrangements. Unique clustering patterns, which allow for interpretation of Moroccan financial market dynamics, have been unraveled by applying Euclidean as well as Dynamic Time Warping (DTW) cluster techniques.

Using the gap statistic approach, we were able to ascertain that five clusters provide the best representation of the underlying patterns in the dataset. We then used the Partitioning Around Medoids (PAM) technique for Euclidean distance and the DTW Barycenter Averaging (DBA) algorithm for DTW distance to perform clustering utilizing both Euclidean and DTW distance metrics.

The clustering results showed both parallels and contrasts between the Euclidean and DTW clustering approaches, with each cluster exhibiting intriguing patterns. For example, in Cluster 1, stocks like DH, DC, RC, ST, MT, and AN were discovered by the Euclidean clustering approach, while a comparable collection of stocks like DH, ST, MT, and AN was found by the DTW clustering method. This implies that the two approaches identify similar underlying patterns inside this cluster, highlighting the similarities in price movements.

In Cluster 2, the Euclidean clustering method identified stocks like RIS, CS, RDS, RM, and MS, while the DTW clustering method revealed a slightly different set of stocks, including SMM, CS, JETC, MD, RDS, RM, ZSA, and MS. This disparity indicates that the DTW method captures nonlinear similarities in the time series data, which may not be fully captured by the Euclidean method.

Similarly, in Cluster 3, both clustering methods identified distinct sets of stocks, albeit with some overlap. While the Euclidean method highlighted stocks such as AISA, DW, PPSA, SNA, EN, INV, BAL, LHM, and TIM, the DTW method emphasized a subset of these stocks, including AISA, DW, BAL, and AG. This suggests that the DTW method may provide a more refined understanding of nonlinear similarities within this cluster.

In Cluster 4, the clustering methods revealed differences in the composition of stocks identified. The Euclidean method identified stocks like SN, SMM, COL, JETC, TMM, LC, FB, CDM, ZSA, MAN, and DPA, whereas the DTW method highlighted a slightly different set of stocks, including SN, COL, PPSA, SNA, EN, TMM, LHM, LC, FB, SMM, CDM, MAN, and DPA. These variations underscore the importance of employing multiple clustering methods to capture the diverse structures present in the data comprehensively.

Lastly, in Cluster 5, both clustering methods identified similar sets of stocks, albeit with slight variations. While the Euclidean method detected HPS, ALL, and IAM, the DTW method highlighted DC, RC, HPS, RIS, ALL, IAM, INV, and TIM. This suggests that both methods recognize comparable patterns in this cluster, emphasizing similarity in price trends.

Overall, the results demonstrate the effectiveness of employing both Euclidean and DTW clustering methods in uncovering meaningful patterns within the Moroccan stock market data. While the Euclidean method emphasizes similarity in price trends, the DTW method offers a more nuanced understanding of nonlinear similarities in the time series data. The things make people comprehend the market's behavior better and they are important to both investors and financial analysts.A

However, it is vital to select the right distance metrics when conducting clustering studies; a comparison between Euclidean and DTW clustering methods has shown this in practice. The complex temporal connections found in financial time series data may not be sufficiently captured by the Euclidean distance metric, which is based on basic vertical distances between data points. The DTW distance metric, on the other hand, is better suited to capture minute similarities in stock price movements since it allows for nonlinear alignments between time series, even when the time series' forms are similar but temporally displaced.

This study emphasizes how important it is to use sophisticated clustering algorithms, such DTW clustering, to get beyond the drawbacks of conventional approaches and learn more about the underlying patterns of financial time series data. By combining the Euclidean and DTW clustering techniques and contrasting the outcomes, we offer a thorough analysis that improves our comprehension of the dynamics of the Moroccan stock market.

To sum up, the results of this investigation highlight how crucial it is to use appropriate clustering methods, such DTW clustering, when examining financial time series data. The knowledge gathered from this study helps to build stronger investment plans and decision-making procedures, which in turn helps investors and financial analysts deal with the intricacies of the stock market more skillfully.

4. CONCLUSION

In summary, this study looked at time series data analysis using clustering techniques inside the context of the Moroccan stock market. We found unique grouping patterns by applying both Euclidean and DTW distance metrics, providing important insights into the dynamics of the market. The results highlight how crucial it is to choose suitable distance metrics and sophisticated clustering techniques in order to identify meaningful patterns in time series data related to finance [17].

By comparing the outcomes of the DTW and Euclidean clustering techniques, this study adds to our understanding of the dynamics and structure of the Moroccan stock market. Particularly, the DTW method worked well for identifying nonlinear patterns that conventional methods might overlook, and the Euclidean method

identified businesses with comparable price fluctuations. By spotting small market patterns and possible investment opportunities, these insights can help financial analysts and investors make better judgments [18].
This study does have certain drawbacks, though. First off, the results can't be applied to other markets with distinct features due to the Moroccan stock market's exclusive emphasis. Secondly, the intricacy and diversity of the market might not be adequately captured by examining just 50 businesses. In addition, some potentially useful metrics might have been missed due to the sole use of DTW and Euclidean distance measurements [19].

By investigating clustering approaches across a wider range of financial markets and incorporating a larger, more diversified set of organizations, future research should solve these constraints. Moreover, adding more distance metrics and using more sophisticated clustering algorithms may improve the accuracy and resilience of the outcomes. Increasing the scope to incorporate more pertinent data sources, such sentiment research or macroeconomic indicators, may yield a more thorough picture of market dynamics. Deeper understanding of the dynamic nature of stock market activity would also be provided by longitudinal studies that monitor changes over time [20].

Future research endeavors can build upon these discoveries and make even greater contributions to the field of financial time series analysis by tackling these restrictions and broadening the study focus. In addition to advancing scholarly understanding, this will provide professionals in the finance and investing industries useful advice [21].

ANNEX

```
# Load necessary libraries
library(cluster)  # For clustering algorithms
library(dtwclust)  # For Dynamic Time Warping based clustering
library(ggplot2)  # For data visualization
library(factoextra)  # For visualizing clustering results

# Define unique colors for plots
unique_hex <- c(
  "#FF5733", "#33FF57", "#5733FF", "#FF8833",
  "#57FF33", "#33AAFF", "#33FF99", "#48233C",
  "#FF33DD", "#AA33FF", "#FFDD33", "#8833FF",
  "#7EA16B", "#00B295", "#4E4144", "#F67987",
  "#CBBEEE", "#C3D898", "#BE3E82", "#DD33FF",
  "#AD8E5C", "#F33F54", "#1C3A13", "#33FF22",
  "#c33149", "#D6C7AE", "#E1BD3D", "#B19CE8",
  "#AAFF33", "#33FFDD", "#33FFAA", "#FF9933",
  "#336699", "#9A8732", "#FF33DD", "#DD33FF",
  "#33AAFF"
)

# Read data from the CSV file
stock_data <- read.csv(file = "stock_data.csv")

# Convert Date column to Date type
stock_data$Date <- as.Date(stock_data$Date, format = "%m/%d/%Y")

# Replace NA values with the mean of the respective column
stock_data[is.na(stock_data)] <- sapply(stock_data, function(x) mean(x, na.rm = TRUE))

# Normalize the data
stock_data[, -1] <- scale(stock_data[, -1])

# Extract close prices for all companies
close_prices <- t(stock_data[, -1])  # Transpose and exclude the Date column
```

```
# Define the abbreviations vector
abbreviations <- c("AISA", "SN", "DH", "DC", "RC", "SMM", "HPS", "COL", "RIS", "CS", "ALL", "ST",
"DW", "IAM", "PPSA", "SNA", "MT", "EN", "INV", "BAL", "JETC", "TMM", "LHM", "AN", "LC", "MD",
"RDS", "FB", "TIM", "SMM", "AG", "RM", "CDM", "ZSA", "MAN", "DPA", "MS")

# Replace rownames of close_prices matrix with abbreviations
rownames(close_prices) <- abbreviations

# Generate a palette of distinguishable colors
color_palette <- unique_hex

# Determine the optimal number of clusters using the gap statistic method
# Visualize the number of clusters using the gap statistic method
fviz_nbclust(close_prices, pam, method = "gap_stat") +
  labs(title = "")

# According to the graph, 5 is the optimal number of clusters
k <- 5

# Perform clustering using Euclidean distance
eu_cluster <- tsclust(close_prices, type = "partitional", k,
                      distance = "Euclidean", centroid = "pam", seed = 1234,
                      trace = TRUE)

# Perform clustering using DTW distance
dtw_cluster <- tsclust(close_prices, type = "partitional", k,
                       distance = "dtw_basic", centroid = "dba", seed = 1234,
                       trace = TRUE)

# Get cluster assignments for Euclidean clustering
eu_assignments <- eu_cluster@cluster

# Get cluster assignments for DTW clustering
dtw_assignments <- dtw_cluster@cluster

# Get the number of clusters
num_clusters <- max(eu_assignments)

# Set up a multi-page layout for plots
pdf("Norm_cluster_comparison_plots.pdf", width = 10, height = 6)

# Create plots for each cluster
for (cluster_num in 1:num_clusters) {
  # Identify stocks assigned to the current cluster for Euclidean clustering
  eu_cluster_stocks <- which(eu_assignments == cluster_num)

  # Identify stocks assigned to the current cluster for DTW clustering
  dtw_cluster_stocks <- which(dtw_assignments == cluster_num)

  # Calculate the minimum and maximum y-axis values for Euclidean clustering
  eu_ymin <- min(close_prices[eu_cluster_stocks, ])
  eu_ymax <- max(close_prices[eu_cluster_stocks, ])

  # Calculate the minimum and maximum y-axis values for DTW clustering
  dtw_ymin <- min(close_prices[dtw_cluster_stocks, ])
  dtw_ymax <- max(close_prices[dtw_cluster_stocks, ])

  # Set up a multi-panel layout for plots in the cluster
```

```
  par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))

  # Plot for Euclidean clustering
  plot(stock_data$Date, close_prices[eu_cluster_stocks[1], ], type = 'l',
       xlab = 'Date', ylab = 'Close Price', main = paste('Cluster', cluster_num, '- Euclidean'),
       xlim = c(min(stock_data$Date), max(stock_data$Date)), ylim = c(eu_ymin, eu_ymax), col =
color_palette[eu_cluster_stocks[1]], lwd = 2)  # Adjust line color and width

  # Plot additional stocks for Euclidean clustering
  for (i in 2:length(eu_cluster_stocks)) {
    lines(stock_data$Date, close_prices[eu_cluster_stocks[i], ], type = 'l', col =
color_palette[eu_cluster_stocks[i]], lwd = 2)
  }

  # Add legend for Euclidean clustering
  legend("topright", legend = rownames(close_prices)[eu_cluster_stocks], title = "Companies", bty = "n",
col = color_palette[eu_cluster_stocks], lwd = 2)

  # Plot for DTW clustering
  plot(stock_data$Date, close_prices[dtw_cluster_stocks[1], ], type = 'l',
       xlab = 'Date', ylab = 'Close Price', main = paste('Cluster', cluster_num, '- DTW'),
       xlim = c(min(stock_data$Date), max(stock_data$Date)), ylim = c(dtw_ymin, dtw_ymax), col =
color_palette[dtw_cluster_stocks], lwd = 2)  # Adjust line color and width

  # Plot additional stocks for DTW clustering
  for (i in 2:length(dtw_cluster_stocks)) {
    lines(stock_data$Date, close_prices[dtw_cluster_stocks[i], ], type = 'l', col =
color_palette[dtw_cluster_stocks[i]], lwd = 2)
  }

  # Add legend for DTW clustering
  legend("topright", legend = rownames(close_prices)[dtw_cluster_stocks], title = "Companies", bty = "n",
col = color_palette[dtw_cluster_stocks], lwd = 2)
}

# Close the PDF device
dev.off()


# Evaluation Metrics for Clustering Algorithms

# Function to calculate Silhouette Score
silhouette_score <- function(data, cluster_labels) {
  silhouette_values <- silhouette(cluster_labels, dist(data))
  mean(silhouette_values[, 3])
}

# Function to calculate Rand Index
rand_index <- function(cluster_labels1, cluster_labels2) {
  rand.index(cluster_labels1, cluster_labels2)
}

# Calculate Silhouette Scores
eu_silhouette <- silhouette_score(close_prices, eu_assignments)
dtw_silhouette <- silhouette_score(close_prices, dtw_assignments)

# Calculate Rand Index
```

```
# Assuming that we use one set of assignments as the true labels for comparison
eu_rand <- rand_index(eu_assignments, dtw_assignments)
dtw_rand <- rand_index(dtw_assignments, eu_assignments)

# Print the results
cat("Silhouette Score (Euclidean):", eu_silhouette, "\n")
cat("Silhouette Score (DTW):", dtw_silhouette, "\n")

cat("Rand Index (Euclidean vs. DTW):", eu_rand, "\n")
```

REFERENCES

[1] G. E. P. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control. Holden-Day, 1970.

[2] R. F. Engle, "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," Econometrica, vol. 50, no. 4, pp. 987-1007, 1982.

[3] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010.

[4] R. N. Mantegna, "Hierarchical structure in financial markets," The European Physical Journal B, vol. 11, no. 1, pp. 193-197, 1999.

[5] W. Huang, Y. Nakamori, and S.-Y. Wang, "Forecasting stock market movement direction with support vector machine," Computers & Operations Research, vol. 32, no. 10, pp. 2513-2522, 2005.

[6] P. Senin, "Dynamic Time Warping Algorithm Review," Information and Computer Science Department University of Hawaii at Manoa, 2008.

[7] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Springer, 1981.

[8] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Advances in Neural Information Processing Systems, 2002, pp. 849-856.

[9] M. Bolla and M. Banerjee, "Spectral clustering and p-median problem," Mathematical Methods of Operations Research, vol. 67, pp. 481-496, 2008.

[10] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley-Interscience, 1990.

[11] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, pp. 43-49, Feb. 1978.

[12] F. Petitjean, A. Ketterlin, and P. Gancarski, "A global averaging method for dynamic time warping, with applications to clustering," Pattern Recognition, vol. 44, no. 3, pp. 678-693, Mar. 2011.

[13] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: Experimental comparison of representations and distance measures," Proceedings of the VLDB Endowment, vol. 1, no. 2, pp. 1542-1552, Aug. 2008.

[14] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 63, no. 2, pp. 411-423, 2001.

[15] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley-Interscience, 1990.

[16] W. M. Rand, "Objective criteria for the evaluation of clustering methods," Journal of the American Statistical Association, vol. 66, no. 336, pp. 846-850, 1971.

[17] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley-Interscience, 1990.

[18] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, pp. 43-49, Feb. 1978.

[19] G. A. F. Seber, "Multivariate Observations," Wiley-Interscience, 1984.

[20] F. Petitjean, A. Ketterlin, and P. Gancarski, "A global averaging method for dynamic time warping, with applications to clustering," Pattern Recognition, vol. 44, no. 3, pp. 678-693, Mar. 2011.

[21] C. Chatfield, "The Analysis of Time Series: An Introduction," Chapman and Hall/CRC, 2003.

[22] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, 1994, pp. 359-370.

[23] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," Knowledge and Information Systems, vol. 7, no. 3, pp. 358-386, 2005.

[24] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," Data Mining and Knowledge Discovery, vol. 26, no. 2, pp. 275-309, Mar. 2013.

[25] T. Warren Liao, "Clustering of time series data—a survey," Pattern Recognition, vol. 38, no. 11, pp. 1857-

1874, Nov. 2005.
[26] P. Esling and C. Agon, "Time-series data mining," ACM Computing Surveys, vol. 45, no. 1, pp. 1-34, Nov. 2012.
[27] I. D. G. Mateos, G. O. Pérez, M. Graña, and J. García-Sebastián, "A comparison of clustering methods for energy consumption profiling," in 2011 IEEE Symposium on Computational Intelligence Applications In Smart Grid (CIASG), Paris, 2011, pp. 1-8.
[28] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835-850, June 2005.
[29] S. Paparrizos and L. Gravano, "k-Shape: Efficient and accurate clustering of time series," in Proc. 2015 ACM SIGMOD International Conference on Management of Data, 2015, pp. 1855-1870.
[30] K. Avrachenkov, P. V. Berkhin, D. K. S. Tong, and G. Y. Yen, "K-means++: Balancing exploration and exploitation," in Proc. 2016 SIAM International Conference on Data Mining (SDM), 2016, pp. 907-915.
[30] K. Avrachenkov, P. V. Berkhin, D. K. S. Tong, and G. Y. Yen, "K-means++: Balancing exploration and exploitation," in Proc. 2016 SIAM International Conference on Data Mining (SDM), 2016, pp. 907-915.
[31] Casablanca Stock Exchange, "Financial Instruments Data," Casablanca Stock Exchange. [Online]. Available: https://www.casablanca-bourse.com/fr/instruments (accessed April 29, 2024).