

## Time Series Clustering of GARCH (1,1) Model Using Modified Piccolo Distance

Vemmi Nastiti Lestari<sup>1</sup>, Abdurakhman<sup>2\*</sup>, Dedi Rosadi<sup>3</sup>

Department of Mathematics, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

Corresponding Author Email: [rachmanstat@ugm.ac.id](mailto:rachmanstat@ugm.ac.id)

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/mmep.120114>

### ABSTRACT

**Received:** 11 October 2024

**Revised:** 4 December 2024

**Accepted:** 10 December 2024

**Available online:** 25 January 2025

#### Keywords:

*GARCH, hierarchical clustering, K-means clustering, Piccolo distance, stock portfolio*

As investors select stocks more efficiently to form portfolios that meet their objectives, they should be grouped based on their time-varying risk levels. One of the clustering approaches is time series clustering with a model-based approach using hierarchical and K-means clustering algorithms. The distance calculation is based on the estimated parameters of the model used; in this case, the GARCH (1,1) model is used. This paper proposes a modified Piccolo distance that uses the absolute value between two GARCH (1,1) models, which is a development of the Manhattan distance. The modified Piccolo distance improves robustness to outliers and simplifies calculations, resulting in more accurate and efficient time series cluster analysis. Applying hierarchical and K-means clustering with modified Piccolo distance will be compared with other model-based distance modifications for clustering applied to simulated data and case studies using stock data incorporated in the Indonesia Stock Exchange. A measure of cluster validity is calculated using the C index. From the simulated data and case studies, it is found that clustering with Piccolo distance modification and other distance modifications between two GARCH (1,1) models produce clusters with a small C index, both for simulated data and case studies. A small C index value in the clustering results indicates good clustering quality, where the clusters formed have high similarity and are well separated from others. Furthermore, the clusters formed will be considered in making a good portfolio, so it is expected to reduce the risk in the stock portfolio.

## 1. INTRODUCTION

A financial portfolio or a collection of assets is an attractive investment to an investor. Forming an optimal portfolio is a challenge in mathematical modeling to achieve the desired goal: minimizing risk or maximizing profits. Therefore, the problem in the portfolio is determining the suitable composition for each asset so that the investor's goals are achieved. One of the development strategies in compiling a portfolio is the selection of assets that are incorporated. The more diverse the assets in the portfolio, the more it can be diversified so that it can be considered to reduce the risk that can occur in the portfolio. The risk in question is the risk of the same movement, which refers to the scenario where the price decline in one asset will also lead to a similar movement in other assets in the portfolio, thereby increasing the overall risk. This is known as systematic risk and cannot be diversified away. Therefore, several studies on portfolios and stock selection strategies have been carried out. One is by grouping stocks using clustering techniques such as those conducted by Tola et al. [1], where clustering techniques can increase portfolio reliability regarding the ratio between risk predictions. Specifically, the clustering used is time series clustering. Time series clustering is an unsupervised learning method that groups data into groups based on their degree of similarity [2], so observations in one group tend to be more similar (based on predetermined criteria) than in other groups

[3]. Aghabozorgi et al. [4] described three methodologies for clustering time series: shape-based, feature-based, and model-based. Shape-based algorithms usually use conventional clustering methods that fit static data, while the distance measure has been modified to a measure suitable for time series. Feature-based algorithms transform raw time series data into feature vectors with reduced dimensionality. These extracted feature vectors are subsequently analyzed using conventional clustering algorithms. Conversely, model-based methods transform raw time series data into parameters associated with a specific parametric model for each time series. Following this transformation, an appropriate model distance is selected, and conventional clustering algorithms are then applied to the derived model parameters.

The clustering algorithm is subsequently applied to the data utilizing a distance measure. Selecting an appropriate distance measure that considers the dependencies between time series is important. Various types of distance measures are utilized in time series clustering, including model-free, model-based, and complexity methods [5]. The model-free approach uses the similarity of the values of two-time series at a specific point in time to calculate how close they are. A possible approach is to compare the two-time series' autocorrelation function (ACF) [6-9]. In addition, in clustering based on raw data, the most commonly used distances are Euclidean distance and Dynamic Time Warping (DTW) [10-12]. Model-based approaches consider that some model or mixture of

underlying probability distributions generates each time series. Model-based approaches, for instance, presume that every time series adheres to an ARMA model. Piccolo [13] used the Euclidean distance between the AR model parameters to compute the distance between two-time series. A chi-square test statistic was assigned to compare two time series using the AR model parameters [14, 15]. Caiado et al. [7] compared and expressed the periodograms of two-time series in terms of Euclidean distance. The complexity-based approach compares the level of complexity of the time series. The similarity of the two-time series depends on measuring the information shared by the two-time series. It is independent of specific time series features or knowledge of the underlying model. Complexity can be considered in two ways: computing each time series' complexity and comparing them to one another [16-18] or giving complexity a weighting function [19].

Kim and Kim [20] have discussed comparing clustering methods for time series with model-free, model-based, and complexity-based approaches. They have been applied to power consumption time series as power consumption profiles. Further to the fact that the variability of financial time series data is not constant, the Generalized Autoregressive Heteroscedasticity (GARCH) model is a suitable choice to capture this variability [21]. GARCH is used because it can overcome the excessive parameters of the ARCH model when applied to time series data, thus ensuring compliance with the principle of parsimony. The GARCH model can categorize stocks into groups based on their volatility. The model uses a positive parameter directly related to the overall variance. As the parameter value increases, the variance value also increases. An essential component of the GARCH model is the grouping of stocks based on time-varying variance. The model that will be of particular interest is the GARCH (1,1) model, which combines the corrected mean return at the time (t-1) and its variance. The research reveals that the GARCH (1,1) model is highly effective for accurately modeling time series data and is the most widely used model for this purpose [22]. This is by parsimony, which effectively reduces the number of parameters in the ARCH model and generalizes them to other parameters [23].

Otranto [24] introduced a metric to measure the dissimilarity between two GARCH (1,1) models and used it in clustering. The distance metric between two GARCH (1,1) models, called Piccolo distance, is suitable for comparing time series data. Therefore, the novelty in this study is the development of Otranto, which modifies the Piccolo distance by replacing it with the absolute value between two GARCH (1,1) models, which is the development of the Manhattan distance. The modified Piccolo distance is proposed as an alternative to overcome the distance formula's limitations in handling outliers in time series data. Replacing the distance formula with the absolute value between two GARCH (1,1) models improves robustness to outliers and simplifies calculations compared to measures involving squares or other powers. Furthermore, clustering is performed on the GARCH (1,1) model using the modified Piccolo distance and compares the distance with the distances based on model-based approaches [7, 13, 15, 25], and cosine distance. Clustering is categorized into two primary types: hierarchical clustering and partitional clustering [26]. This study utilizes both algorithms for its analysis. In hierarchical clustering, the number of clusters is not predetermined. This algorithm initiates by treating each time series as an individual cluster, then gradually merging the closest clusters until all the data is

incorporated into one large cluster. Next, complete linkage is used to calculate the distance between groups, where the method is based on maximum distance. The K-means algorithm, the simplest and most commonly used clustering method, is employed to group data due to its fast and efficient computation time. This algorithm uses a centroid-based partitioning method to divide observations into several K clusters, with a predetermined value of K, which is the number of clusters to be created by the algorithm. The selection of the optimal number of clusters and the accurate identification of centroids determines the effectiveness of the K-means clustering process. These factors are essential for achieving precise and reliable clustering results [27, 28].

The structure of this paper is as follows. Section 2 introduces the modified Piccolo distance between GARCH (1,1) models and some model-based distance measures between two GARCH (1,1) models. Section 3 presents the hierarchical and K-means clustering algorithm and validity clustering measures. Section 4 outlines the results of the clustering analyses based on simulated data, and a case study applied to some stock data listed on the Indonesia Stock Exchange. The conclusions and limitations of the study are given in Section 5.

## 2. DISTANCE BETWEEN TWO GARCH (1,1) MODELS

The model of primary focus is the GARCH (1,1) model, which combines the corrected mean return at a point in time and its variance. Research indicates that the GARCH (1,1) model is highly effective for accurately modeling time series data and is the most popular model adopted for time series [22]. This is done by parsimony, which effectively reduces the number of parameters in the ARCH model and generalizes them to other parameters [23]. This section will explore the theoretical basis of the distance between two GARCH (1,1) models.

### 2.1 GARCH (1,1) model

The GARCH (1,1) model is widely recognized as one of the foremost models for analyzing various time series data [20]. This model adheres to the principle of parsimony by effectively reducing the number of parameters present in ARCH models and generalizing them into alternative parameters [22]. According to Bollerslev [23], the GARCH (1,1) model is considered the most suitable for characterizing financial data. Its simplicity, involving only two parameters, facilitates the calculation of the distance between two GARCH (1,1) models, thereby enhancing its utility in practical applications.

It is known that  $\varepsilon_t = r_t - \mu_t$  the mean corrected log return,  $\varepsilon_t$  adheres to the GARCH (1,1) model [21] if

$$\varepsilon_t = \nu_t \sigma_t \quad (1)$$

with

$$\sigma_t^2 = \gamma + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \quad (2)$$

where,  $\nu_t \sim IID(0,1)$ ,  $\gamma > 0$ ,  $\alpha_1 \geq 0$ ,  $\beta_1 \geq 0$  and  $(\alpha_1 + \beta_1) < 1$ . Assumed  $E(\varepsilon_t | F_{t-1}) = 0$ ,  $Var(\varepsilon_t | F_{t-1}) = E(\varepsilon_t^2 | F_{t-1}) = \sigma_t^2$ .

## 2.2 Parameter estimation of GARCH (1,1) model

Parameter estimation of  $\gamma$ ,  $\alpha_1$  and  $\beta_1$  in the GARCH (1,1) model requires modeling the regression such that

$$y_t = \tau_0 + \tau_1 y_{t-1} + \varepsilon_t, t = 1, \dots, T \quad (3)$$

with

$$\varepsilon_t = v_t \sigma_t = v_t \sqrt{h_t}, h_t = \gamma + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} \quad (4)$$

The parameter vector  $\tilde{\theta}$  is expressed as

$$\tilde{\theta} = (\tau_0, \tau_1, \gamma, \alpha_1, \beta_1)' = (\tilde{\tau}', \tilde{\delta}')'$$

with

$$\tilde{\tau} = \begin{bmatrix} \tau_0 \\ \tau_1 \end{bmatrix}, \tilde{\delta} = \begin{bmatrix} \gamma \\ \alpha_1 \\ \beta_1 \end{bmatrix}$$

In general, Eq. (3) is an AR (1) model with  $\tau_0$  as a constant and  $\tau_1$  autoregressive coefficients, while the GARCH (1,1) model is according to Eq. (2), which is rewritten in Eq. (4). For this reason, because the GARCH (1,1) model is used, the parameters estimated according to Eq. (4) are  $\gamma$ ,  $\alpha_1$  and  $\beta_1$ . To estimate the parameters of  $\gamma$ ,  $\alpha_1$  and  $\beta_1$  begins by finding the likelihood function on:

$$f(\varepsilon_t | F_{t-1}) = \frac{1}{\sqrt{2\pi h_t}} e^{-\frac{1}{2} \frac{\varepsilon_t^2}{h_t}} \quad (5)$$

The likelihood function for the  $t$  observation and sample size expressed with  $T$  is denoted by  $L_t$ , then:

$$L_t = -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln h_t - \frac{1}{2} \frac{\varepsilon_t^2}{h_t} \quad (6)$$

Furthermore, it is derived from  $\tilde{\delta}$ , obtained:

$$\frac{\partial L_t}{\partial \tilde{\delta}} = \frac{1}{2h_t} \frac{\partial h_t}{\partial \tilde{\delta}} \left( \frac{\varepsilon_t^2}{h_t} - 1 \right) = \frac{1}{2} \left( \frac{1}{h_t} \right) \tilde{b}_t v_t \quad (7)$$

The iteration form is derived from the modified Newton-Raphson method as follows:

$$\tilde{\delta}_{i+1} = \tilde{\delta}_i - \left[ \sum_{t=1}^T v_t^2 \left( \frac{1}{2} \frac{\tilde{b}_t}{h_t} \right) \left( \frac{1}{2} \frac{\tilde{b}_t}{h_t} \right)' \right]^{-1} \sum_{t=1}^T \frac{1}{2} \left( \frac{1}{h_t} \right) \tilde{b}_t v_t \quad (8)$$

with

$$\tilde{b}_t = \frac{\partial h_t}{\partial \tilde{\delta}} \frac{\varepsilon_t^2}{h_t^2}, v_t = \frac{\partial h_t}{\partial \tilde{\delta}}$$

## 2.3 Distance between two GARCH (1,1) models

The distance between two GARCH (1,1) models is calculated based on an extension of the ARMA (1,1) model distance proposed by Piccolo [13]. The time series model with  $t = 1, 2, \dots, T$  is as follows:

$$\text{Model 1: } y_{1,t} = \mu_1 + \varepsilon_{1,t}$$

$$\text{Model 2: } y_{2,t} = \mu_2 + \varepsilon_{2,t}$$

where,  $\varepsilon_{1,t}$  and  $\varepsilon_{2,t}$  are errors with zero mean and time-varying variance. It is assumed that the variances  $h_{1,t}$  and  $h_{2,t}$  adhere to two distinct and independent GARCH (1,1) structures, as in Eq. (4),

$$\begin{aligned} \sigma_{1,t}^2 &= h_{1,t} = \gamma_1 + \alpha_1 \varepsilon_{1,t-1}^2 + \beta_1 h_{1,t-1} \\ \sigma_{2,t}^2 &= h_{2,t} = \gamma_2 + \alpha_2 \varepsilon_{2,t-1}^2 + \beta_2 h_{2,t-1} \end{aligned}$$

with  $\gamma_i > 0, 0 < \alpha_i < 1, 0 < \beta_i < 1, (\alpha_i + \beta_i) < 1 (i = 1, 2)$ . Further,  $V_t$  be a zero-mean ARMA invertible process and  $F$  is a class of invertible ARMA processes. It is known that if  $V_t \in F$ , then, there exists a constant  $\pi_i$  such as

$$\sum_{j=1}^{\infty} |\pi_j| < \infty$$

and

$$V_t = \sum_{j=1}^{\infty} \pi_j V_{t-j} + \varepsilon_t, \text{ with } \varepsilon_t \sim WN(0, \sigma^2)$$

The distance between two processes  $V_{1t}, V_{2t} \in F$  is defined as stated by Piccolo and then called Piccolo distance [13].

$$\begin{aligned} d^2(V_{1t}, V_{2t}) &= \sum_{j=1}^{\infty} (\pi_{1j} - \pi_{2j})^2 \\ d &= \left[ \sum_{j=1}^{\infty} (\pi_{1j} - \pi_{2j})^2 \right]^{\frac{1}{2}} \end{aligned} \quad (9)$$

with  $\pi_{1j}$  and  $\pi_{2j}$  are the coefficients of the two AR processes. From Eq. (9), the distance of two GARCH (1,1) models is as follows:

$$\begin{aligned} d &= \left( \sum_{j=0}^{\infty} (\alpha_1 \beta_1^j - \alpha_2 \beta_2^j)^2 \right)^{\frac{1}{2}} \\ d &= \left( \frac{\alpha_1^2}{1 - \beta_1^2} + \frac{\alpha_2^2}{1 - \beta_2^2} - \frac{2\alpha_1 \alpha_2}{1 - \beta_1 \beta_2} \right)^{\frac{1}{2}} \end{aligned} \quad (10)$$

### 2.3.1 Modified Piccolo distance

Modified Piccolo distance is the new distance that is different from the Piccolo distance in the GARCH model (see Eq. (10)) developed by Otranto [24]. The background of the modified Piccolo distance is based on the Manhattan distance, where the calculation method for the distance space applies the concept of absolute difference. The squared distance is generally the most used distance metric, as in Eq. (9). Still, it has some disadvantages compared to the absolute distance, which is sensitive to outliers. Since the difference of the data is squared, if there is extreme data in one dimension, the overall value of the squared distance can increase significantly, so the outlier value will greatly affect the result of the distance calculation. This is one of the motivations to modify the Piccolo distance using absolute values. Then, the distance between the two processes  $V_{1t}, V_{2t} \in F$  is defined as

$$d = \sum_{j=1}^{\infty} |\pi_{1j} - \pi_{2j}| \quad (11)$$

with  $\pi_{1j}$  and  $\pi_{2j}$  are the coefficients of the two AR processes, like the Piccolo distance for the GARCH model by Otranto [24]. From Eq. (11), the distance between two GARCH (1,1) models, which is a modification of the Piccolo distance, is shown below:

$$\begin{aligned} d &= \sum_{j=1}^{\infty} |\alpha_1 \beta_1^{j-1} - \alpha_2 \beta_2^{j-1}| \\ &= \sum_{j=0}^{\infty} |\alpha_1 \beta_1^j - \alpha_2 \beta_2^j| = \left| \alpha_1 \sum_{j=0}^{\infty} \beta_1^j - \alpha_2 \sum_{j=0}^{\infty} \beta_2^j \right| \quad (12) \\ &= |\alpha_1(\beta_1^0 + \beta_1^1 + \beta_1^2 + \dots) - \alpha_2(\beta_2^0 + \beta_2^1 + \beta_2^2 + \dots)| \\ &= \left| \frac{\alpha_1}{1 - \beta_1} - \frac{\alpha_2}{1 - \beta_2} \right| \end{aligned}$$

### 2.3.2 Distance measure by Caiado

Two-time series each fit the GARCH (1,1) model, with  $\pi_1 = (\alpha_1, \beta_1)$  and  $\pi_2 = (\alpha_2, \beta_2)$  as the parameter estimation vectors. Furthermore,  $S_1$  and  $S_2$  are estimates of the variance-covariance matrix, respectively. Caiado and Crato [29] defined the distance between volatilities of time series data as

$$d = (\pi_1 - \pi_2)' S^{-1} (\pi_1 - \pi_2) \quad (13)$$

where,  $S = S_1 + S_2$ . This distance captures all the stochastic structure of a process's conditional variance and offers a solution for comparing time series data of unequal lengths [29].

### 2.3.3 Maharaj distance

Maharaj et al. [14, 30] proposed a hypothesis test that has practical applications in evaluating whether there is a significant difference between two-time series. This test is designed to assess whether the generation processes of the two-time series differ significantly. The corresponding test statistic is defined as follows:

$$d = \sqrt{T} (\pi_1 - \pi_2)' S^{-1} (\pi_1 - \pi_2) \quad (14)$$

with  $T$  is the length of the series,  $\pi_1 = (\alpha_1, \beta_1)$  and  $\pi_2 = (\alpha_2, \beta_2)$  as the parameter estimation vectors.  $S = \sigma_1^2 R_1^{-1} + \sigma_2^2 R_2^{-1}$ , where  $\sigma_1^2$  and  $\sigma_2^2$  are the variances of the white noise processes for each series,  $R_1$  and  $R_2$  the samples covariance matrices of both series.

The null hypothesis ( $H_0$ ) is rejected at the  $\alpha$  significance level when the statistic  $d > \chi^2(k)$ , where  $\chi^2(k)$  denotes the  $(1 - \alpha)$  th quantile of the chi-square distribution with  $k$  degrees of freedom. If the null hypothesis is rejected, it indicates that the series  $y_{1,t}$  and  $y_{2,t}$  generating processes are significantly different [14].

### 2.3.4 Distance cosine

Cosine distance is a quantitative measure of dissimilarity between two vectors and is calculated as the complement of the cosine similarity value. Cosine similarity assesses the degree of similarity between two vectors by evaluating the cosine of the angle formed between them. The Euclidean dot product formula can determine this cosine value for two non-zero vectors.

$$\pi_1 \cdot \pi_2 = \|\pi_1\| \|\pi_2\| \cos(\theta)$$

with  $\pi_1 = (\alpha_1, \beta_1)$  and  $\pi_2 = (\alpha_2, \beta_2)$  as the parameter estimation vectors. The cosine similarity,  $\cos(\theta)$  is represented using a dot product given two  $n$ -dimensional vectors,  $\pi_1$  and  $\pi_2$ .

$$\begin{aligned} \text{cosine similarity} &= \cos(\theta) = \frac{\pi_1 \cdot \pi_2}{\|\pi_1\| \|\pi_2\|} \\ \cos(\theta) &= \frac{\sum_{i=1}^n \pi_{1i} \pi_{2i}}{\sqrt{\sum_{i=1}^n \pi_{1i}^2} \cdot \sqrt{\sum_{i=1}^n \pi_{2i}^2}} \quad (15) \end{aligned}$$

where,  $\pi_{1i}$  and  $\pi_{2i}$  are the  $i$ th components of vectors  $\pi_1$  and  $\pi_2$ , respectively. The cosine distance between two vectors  $\pi_1$  and  $\pi_2$  is defined as:

$$\text{cosine distance}(\pi_1, \pi_2) = 1 - \cos(\theta) \quad (16)$$

with  $\cos(\theta)$  is the cosine similarity calculated according to Eq. (15).

## 3. CLUSTERING ALGORITHM

This section will explore the theoretical basis of the clustering algorithm.

### 3.1 Hierarchical clustering

Hierarchical clustering is a method utilized in unsupervised machine learning to identify clusters of observations within a dataset. This technique does not necessitate the specification of a predefined number of clusters, as is the case with K-means clustering. A principal application of hierarchical clustering is consolidating groups exhibiting similar volatility structures. The methodology involves merging the two closest clusters based on a defined distance measure. Furthermore, two clusters with a notably low distance measure may also be combined. The distance used in this context is calculated between two GARCH (1,1) models, as discussed in Section 2.3.

The distance between two GARCH models can be employed to cluster  $n$  time series into a homogeneous group with a similar structure. In agglomerative hierarchical clustering, all-time series commences in individual clusters, which are then recursively merged at various levels based on similarity. Meanwhile,  $T - 1$  merging steps are reported to correspond with  $T$  observations. Measuring the dissimilarity between the clusters is necessary to figure out the merging order. Several measures exist, including complete, single, and average linkage. For this study, complete linkage has been selected to define the dissimilarity between clusters  $A$  and  $B$  as follows:

$$d(A, B) = \max_{i \in A, j \in B} d_{i,j} \quad (17)$$

where,  $d_{i,j}$  is the distance between observation  $i$  and  $j$  in clusters  $A$  and  $B$  according to Eqs. (10)-(15).

### 3.2 K-means clustering

K-means clustering is a practical algorithm for partitioning a given data set into  $K$  distinct clusters. This method enhances data organization by ensuring high similarity among objects within the same cluster (high intra-class similarity) while

maintaining low similarity among objects in different clusters (low inter-class similarity). In the K-means clustering process, each cluster is characterized by a centroid, which is determined by calculating the average of the data points assigned to that cluster. Using the mean formula between the objects, the algorithm allowed objects to be clustered according to the nearest centroid. Let  $y = C_1 \cup C_2 \cup \dots \cup C_K$  and  $C_i \cap C_j = \emptyset$ . Clusters are determined by

$$\arg \min_c \sum_{i=1}^K \sum_{y_j \in C_i} \|y_j - C_i\|^2 \quad (18)$$

where,  $C_i$  is the center of the cluster. The K-means clustering process resembles that of the EM algorithm. At the outset, each object is randomly allocated to a cluster according to the cluster centers.

### 3.3 Cluster validity measure

The evaluation of clustering algorithm results is fundamentally based on cluster validation. Several validation indices have been developed to assess the quality of clusters, with the C-index serving as a significant validation tool [31, 32]. The C-index is categorized as an internal validity index that aims to define and identify the most effective partitioning of a set of  $n$  objects. This process utilizes unlabelled feature vectors or dissimilarity matrix data [33]. This measure considers the ratio between the total observed within-cluster distances and the total minimum and maximum distances possible for the same number of clusters. The calculation of the C-index is detailed as follows:

$$C = \frac{S_w - S_{min}}{S_{max} - S_{min}} \quad (19)$$

where,  $S_w$  denotes the total distance between all pairs of items within the same cluster. Here,  $n$  represents the number of pairings while  $S_{min}$  indicates the total of all object pairs' lowest distances. In contrast,  $S_{max}$  refers to the sum of the highest distances across all possible pairs. The C-index's properties evaluate how well a cluster's data points are grouped. The C-index value ranges from 0 to 1. A C-index value close to 0 indicates that the resulting cluster is close to optimal clustering. C-index values close to 1 indicate that the clustering results are less than optimal. The reason for choosing the C-Index lies in its simplicity and effectiveness. Unlike other validity indices (e.g., silhouette score, Dunn index), the C-index focuses only on the distance within clusters without considering the inter-cluster distance, which is particularly useful for assessing the quality of clustering based on parameter estimates (e.g., in GARCH models). Since the modified Piccolo distance is used in this study, the C-index effectively captures cluster compactness based on this distance measure, ensuring consistency in evaluation.

## 4. DATA ANALYSIS

### 4.1 Simulation

The simulated data consists of 15 time series clustered into 3 clusters. Each cluster consists of 5 time series with 100 points in each time series. The following provides a detailed

description of each cluster.

Cluster 1. GARCH (1,1) Model, with

$$h_t = 0.005 + 0.1\varepsilon_{t-1}^2 + 0.1h_{t-1}$$

Cluster 2. GARCH (1,1) Model, with

$$h_t = 0.1\varepsilon_{t-1}^2 + 0.8h_{t-1}$$

Cluster 3. GARCH (1,1) Model, with

$$h_t = -0.005 + 0.9\varepsilon_{t-1}^2 + 0.01h_{t-1}$$

The steps of the clustering process include:

1. Generate time series using the GARCH (1,1) model with the conditions described in the simulation data above. Repeat for each model (Cluster 1, 2, and 3) with 5 time series each.
2. Calculate the distance between time series using different distances based on the estimated GARCH (1,1) parameters according to Eqs. (10)-(16).
3. Implementation of clustering algorithm
  - a. Hierarchical Clustering
    - i. Use complete linkage as the linkage method according to Eq. (17).
    - ii. Input the distance between time series calculated in Step 2 into the hierarchical clustering algorithm.
  - b. K-means Clustering
    - i. Initialize the initial centroid using 3 random time series from the dataset (since the number of clusters is 3).
    - ii. Assign each time series to the cluster with the smallest distance from the centroid.
    - iii. Update the centroid to show the average of the estimated parameters of each cluster.
    - iv. Repeat the above steps until it converges.
4. Evaluate the validity of clustering results using the C-index validity measure.

Figure 1 shows the average line of each cluster. Table 1 shows the results of 5 distance measures with hierarchical complete linkage and K-means clustering algorithms. The complete linkage method based on Eq. (17) minimizes the variance between clusters in hierarchical clustering. As for K-means, the initial centroid initialization process is randomly selected from the existing dataset. Furthermore, the C-index is used to measure the validity of the cluster.

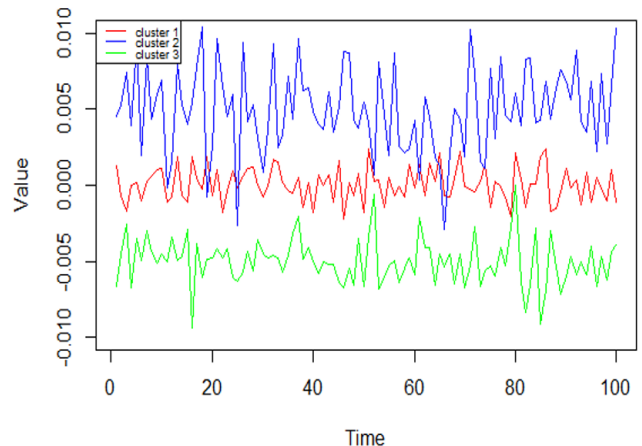


Figure 1. Plot of each cluster

**Table 1.** Clustering with simulation data

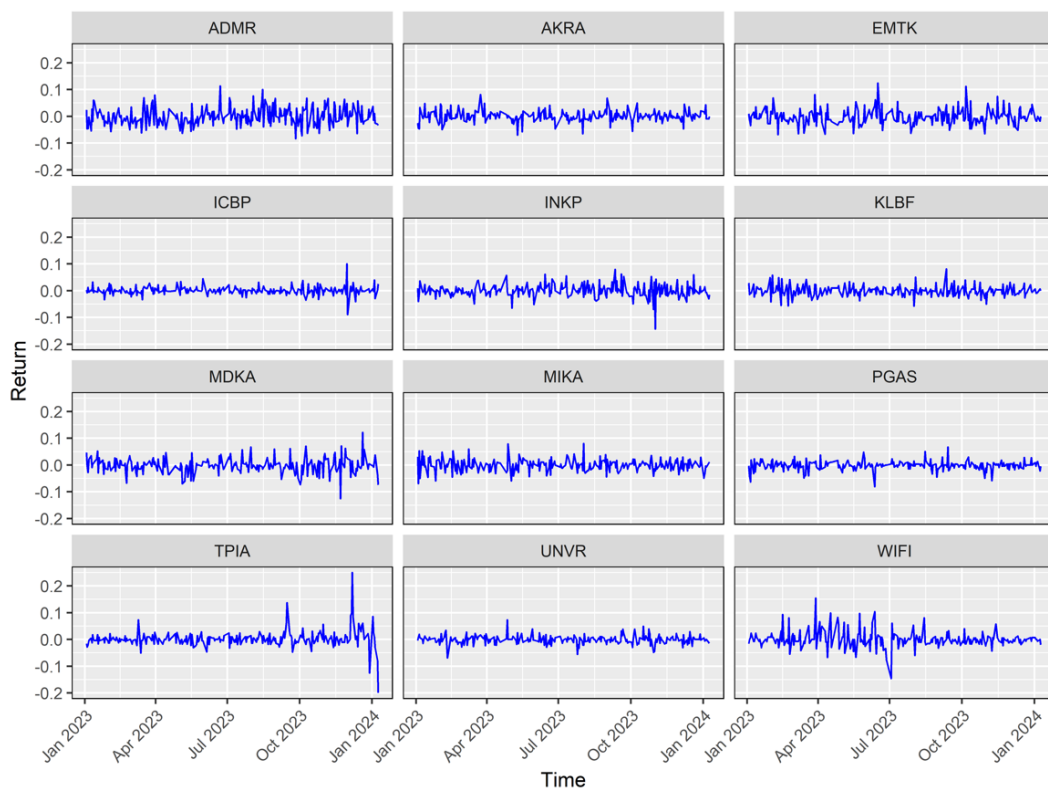
Algorithm	Distance	C-Index
Hierarchical	Piccolo	0.578118
	Modified Piccolo	0.578118
	Caiado	0.450013
	Maharaj	0.450013
	Cosine	0.450013
	Piccolo	0.578118
K-means	Modified Piccolo	0.455615
	Caiado	0.007818
	Maharaj	0.001686
	Cosine	0.001686

Table 1 shows that the best-performing hierarchical clustering with small c-index values is obtained using Caiado, Maharaj, and cosine distances. Then, in K-means clustering, Maharaj and cosine distance gave the best performance. However, when compared to the Piccolo distance, the

modified Piccolo distance produces a smaller c-index for K-means clustering.

#### 4.2 Stock data analysis

The case study used data from twelve stocks listed on the Indonesia Stock Exchange using daily closing prices taken from Yahoo Finance. These stocks include PT Adaro Minerals Indonesia Tbk (ADMR), PT AKR Corporindo Tbk (AKRA), PT Elang Mahkota Teknologi Tbk (EMTK), PT Indofood CBP Sukses Makmur Tbk (ICBP), PT Indah Kiat Pulp and Paper Tbk (INKP), PT Kalbe Farma Tbk (KLBF), PT Merdeka Copper Gold Tbk (MDKA), PT Mitra Keluarga Karyasehat Tbk (MIKA), State Gas Company (PGAS), PT Chandra Asri Pacific Tbk, PT Unilever Indonesia Tbk (UNVR), and PT Solusi Sinergi Digital Tbk from January 3<sup>rd</sup>, 2023 until January 9<sup>th</sup>, 2024. The first step is to calculate the return, and then each stock return is estimated using the GARCH (1,1) model according to Eq. (4).

**Figure 2.** Plot return of each stock**Table 2.** Clustering with return stock data

Algorithm	Distance	Number of Clusters	C-Index
Hierarchical	Piccolo	2	0.514991
	Modified Piccolo	2	0.547525
	Caiado	2	0.650697
	Maharaj	2	0.650697
	Cosine	2	0.650697
	Piccolo	2	0.452220
K-means	Modified Piccolo	2	0.490864
	Caiado	2	0.001865
	Maharaj	2	0.087182
	Cosine	2	0.001865

**Table 3.** Recapitulation of cluster members for hierarchical clustering with Piccolo and modified Piccolo distance

Cluster 1	Cluster 2
ADMR, EMTK, PGAS, UNVR, ICBP, MDKA, MIKA, WIFI, KLBF, TPIA, INKP	AKRA

**Table 4.** Recapitulation of cluster members for K-means clustering with Caiado and cosine distance

Caiado Distance		Cosine Distance	
Cluster 1	Cluster 2	Cluster 1	Cluster 2
ADMR, PGAS, ICBP, MIKA	EMTK, UNVR, MDKA, WIFI, KLBF, TPIA, INKP, AKRA	EMTK, UNVR, MDKA, WIFI, KLBF, TPIA, INKP, AKRA	ADMR, PGAS, ICBP, MIKA

Figure 2 shows the return plot for each stock. Each stock shows different volatility. It can be seen that TPIA and WIFI stocks show higher volatility spikes compared to others. Table 2 shows the results of 5 distance measures with hierarchical complete linkage and K-means clustering algorithms. Similarly to what was done with the simulated data, in hierarchical clustering, the complete linkage method based on Eq. (17) minimizes the variance between clusters. For K-means, the initial centroid initialization process is randomly selected from the existing data set. Next, the C-index measures the validity of the cluster.

In Table 2, the hierarchical clustering algorithm with the distance that gives the best performance is the small c-index value achieved in Piccolo and modified Piccolo distance. In K-means clustering, Caiado and cosine distance gave the best performance. Furthermore, Table 3 shows the clusters formed for hierarchical clustering with Maharaj, Caiado, and cosine distance, and Table 4 shows K-means clustering with Caiado and cosine distance.

Table 3 shows that two clusters with the same cluster members are obtained for each of the Piccolo and modified Piccolo distances. K-means clustering can be seen in Table 4, where two clusters with the same cluster members are also obtained using Caiado and cosine distances.

The clustering results based on GARCH (1,1) models allow investors to identify asset groups with similar volatility patterns. From the clustering results, two clusters with different volatility patterns are obtained. Stocks in one cluster tend to have similar volatility parameters, providing a framework for selecting assets from different clusters. TPIA and WIFI stocks that show higher volatility (also shown in Figure 2) are in the same group. The results are expected to reduce the risk in portfolio selection. For example, investors can combine stocks from high-volatility clusters with stocks from low-volatility clusters to create a more balanced portfolio. Information from the clusters can also be used to determine the optimal weight of assets in the portfolio. Stocks in lower-risk clusters can be given more weight for conservative investors, while stocks from high-volatility clusters can be included in smaller proportions to increase profit opportunities. Furthermore, the clusters formed can be updated regularly as market volatility patterns change. This helps investors adjust the portfolio to dynamic market conditions, such as increased volatility during periods of economic uncertainty. The ability to cluster assets based on volatility patterns aims to provide analytical tools for investors and portfolio managers. Therefore, these clustering results bridge theoretical models of volatility and practical applications in investment decision-making.

## 5. CONCLUSIONS

This study develops a modified Piccolo distance, based on Manhattan distance and absolute difference, to improve the clustering of time series data modeled with the GARCH (1,1) model. Using hierarchical clustering and K-means algorithms, the proposed distance metric is evaluated on simulated data and stock return data from the Indonesia Stock Exchange (January 3<sup>rd</sup>, 2023 to January 9<sup>th</sup>, 2024). The results show that the modified Piccolo distance consistently produces high-quality clusters, as small C index values indicate. Two distinct clusters are identified for the stock data, which capture differences in volatility patterns. This research shows that the

proposed clustering method can improve portfolio optimization by guiding investment strategies, allowing the identification of homogeneous groups of assets based on risk profiles. For example, clusters with low volatility may be suitable for conservative investors, while clusters with high volatility may suit higher risk and higher return strategies. In addition, it can also improve risk management by grouping assets based on volatility patterns, allowing for more optimized asset allocation and early detection of systemic risk.

Although this research is limited to the GARCH (1,1) model and standard clustering algorithms, it provides a basis for further exploration of more sophisticated models and methods. Future research could extend the modified Piccolo distance to other time series models, such as (EGARCH or higher order GARCH models) or utilize other clustering techniques, such as density-based clustering. This research advances the understanding of time series clustering and offers practical insights for financial modeling.

## REFERENCES

- [1] Tola, V., Lillo, F., Gallegati, M., Mantegna, R.N. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1): 235-258. <https://doi.org/10.1016/j.jedc.2007.01.034>
- [2] Nastiti Lestari, V., Abdurakhman, Rosadi, D. (2024). Robust time series clustering of GARCH (1,1) models with outliers. *Statistics*. <https://doi.org/10.1080/02331888.2024.2426741>
- [3] Javed, A., Lee, B.S., Rizzo, D.M. (2020). A benchmark study on time series clustering. *Machine Learning with Applications*, 1: 100001. <https://doi.org/10.1016/j.mlwa.2020.100001>
- [4] Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y. (2015). Time-series clustering—A decade review. *Information Systems*, 53: 16-38. <https://doi.org/10.1016/j.is.2015.04.007>
- [5] Montero, P., Vilar, J.A. (2015). TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62: 1-43. <https://doi.org/10.18637/jss.v062.i01>
- [6] Bohte, Z., Čepar, D., Košmelj, K. (1980). Clustering of time series. In *Proceeding in Computational Statistics*, pp. 587-593.
- [7] Caiado, J., Crato, N., Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, 50(10): 2668-2684. <https://doi.org/10.1016/j.csda.2005.04.012>
- [8] D'Urso, P., Maharaj, E.A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, 160(24): 3565-3589. <https://doi.org/10.1016/j.fss.2009.04.013>
- [9] Galeano, P., Pena, D.P. (2000). Multivariate analysis in vector time series. *Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo*, 4(4): 383-403.
- [10] Cai, B., Huang, G., Samadiani, N., Li, G., Chi, C.H. (2021). Efficient time series clustering by minimizing dynamic time warping utilization. *IEEE Access*, 9: 46589-46599. <https://doi.org/10.1109/ACCESS.2021.3067833>
- [11] Holder, C., Middlehurst, M., Bagnall, A. (2024). A review and evaluation of elastic distance functions for time series clustering. *Knowledge and Information*



- Systems, 66(2): 765-809. <https://doi.org/10.1007/s10115-023-01952-0>
- [12] Bhavani, S.V., Xiong, L., Pius, A., Semler, M., et al. (2023). Comparison of time series clustering methods for identifying novel subphenotypes of patients with infection. *Journal of the American Medical Informatics Association*, 30(6): 1158-1166. <https://doi.org/10.1093/jamia/ocad063>
- [13] Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, 11(2): 153-164. <https://doi.org/10.1111/j.1467-9892.1990.tb00048.x>
- [14] Maharaj, E.A. (1996). A significance test for classifying ARMA models. *Journal of Statistical Computation and Simulation*, 54(4): 305-331. <https://doi.org/10.1080/00949659608811737>
- [15] Maharaj, E.A. (2000). Cluster of time series. *Journal of Classification*, 17(2): 297-314. <https://doi.org/10.1007/s003570000023>
- [16] Li, M., Chen, X., Li, X., Ma, B., Vitányi, P.M. (2004). The similarity metric. *IEEE Transactions on Information Theory*, 50(12): 3250-3264. <https://doi.org/10.1109/TIT.2004.838101>
- [17] Keogh, E., Lonardi, S., Ratanamahatana, C.A. (2004). Towards parameter-free data mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, USA, pp. 206-215. <https://doi.org/10.1145/1014052.1014077>
- [18] Keogh, E., Lonardi, S., Ratanamahatana, C.A., Wei, L., Lee, S.H., Handley, J. (2007). Compression-based data mining of sequential data. *Data Mining and Knowledge Discovery*, 14: 99-129. <https://doi.org/10.1007/s10618-006-0049-3>
- [19] Batista, G.E., Wang, X., Keogh, E.J. (2011). A complexity-invariant distance measure for time series. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, Mesa, Arizona, pp. 699-710. <https://doi.org/10.1137/1.9781611972818.60>
- [20] Kim, J., Kim, J. (2020). Comparison of time series clustering methods and application to power consumption pattern clustering. *Communications for Statistical Applications and Methods*, 27(6): 589-602. <https://doi.org/10.29220/CSAM.2020.27.6.589>
- [21] Otranto, E. (2008). Clustering heteroskedastic time series by model-based procedures. *Computational Statistics & Data Analysis*, 52(10): 4685-4698. <https://doi.org/10.1016/j.csda.2008.03.020>
- [22] Bollerslev, T., Chou, R.Y., Kroner, K.F. (1992). ARCH modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics*, 52(1-2): 5-59. [https://doi.org/10.1016/0304-4076\(92\)90064-X](https://doi.org/10.1016/0304-4076(92)90064-X)
- [23] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3): 307-327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- [24] Otranto, E. (2004). Classifying the markets volatility with ARMA distance measures. *Quaderni di Statistica*, 6: 1-19.
- [25] Piccolo, D. (1989). On a measure of dissimilarity between ARIMA models. In *Proceedings of the ASA Meetings-Business and Economic Stat.* Washington DC.
- [26] Al Kababchee, S.G., Algamal, Z.Y., Qasim, O.S. (2023). Improving penalized-based clustering model in big fusion data by hybrid black hole algorithm. *Fusion: Practice & Applications*, 11(1): 70-76. <https://doi.org/10.54216/FPA.110105>
- [27] Al-Kababchee, S.G.M., Algamal, Z.Y., Qasim, O.S. (2023). Enhancement of K-means clustering in big data based on equilibrium optimizer algorithm. *Journal of Intelligent Systems*, 32(1): 20220230. <https://doi.org/10.1515/jisys-2022-0230>
- [28] Al Radhwani, A.M.N., Algamal, Z.Y. (2021). Improving K-means clustering based on firefly algorithm. *Journal of Physics: Conference Series*, 1897: 012004. <https://doi.org/10.1088/1742-6596/1897/1/012004>
- [29] Caiado, J., Crato, N. (2007). A GARCH-based method for clustering of financial time series: International stock markets evidence. In *Recent Advances in Stochastic Modeling and Data Analysis*, pp. 542-551. [https://doi.org/10.1142/9789812709691\\_0064](https://doi.org/10.1142/9789812709691_0064)
- [30] Maharaj, E.A., D'Urso, P., Caiado, J. (2019). *Time Series Clustering and Classification*. Chapman and Hall/CRC.
- [31] Jain, A.K., Murty, M.N., Flynn, P.J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3): 264-323. <https://doi.org/10.1145/331499.331504>
- [32] Koutroumbas, K., Theodoridis, S. (2008). *Pattern Recognition*. Academic Press.
- [33] Hubert, L., Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, 29(2): 190-241. <https://doi.org/10.1111/j.2044-8317.1976.tb00714.x>