J. Ris. & Ap. Mat. Vol. 07 No. 02 (2023) pp. 104-115
Jurnal Riset dan Aplikasi Matematika
e-ISSN: 2581-0154
URL: journal.unesa.ac.id/index.php/jram

# CLUSTER ANALYSIS FOR MEAN-VARIANCE PORTFOLIO SELECTION: A COMPARISON BETWEEN K-MEANS AND K-MEDOIDS CLUSTERING

LA GUBU[1*], EDI CAHYONO[2], ARMAN[3], HERDI BUDIMAN[4], MUH KABIL DJAFAR[5]

[1,2,3,4,5] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Halu Oleo University

*Corresponding author: la.gubu@uho.ac.id

**ABSTRACT**

This paper presents the Mean-Variance (MV) portfolio selection using cluster analysis. Stocks are categorized into various clusters using K-Means and K-Medoids clustering. Based on the Sharpe ratio, a stock from each cluster is chosen to represent that cluster. Stocks with the greatest Sharpe ratio are those that are chosen for each cluster. With the guidance of the MV portfolio model, the optimum portfolio is identified. When there are many stocks included in the formation of the portfolio, we may efficiently create the optimal portfolio using this method. For the empirical study, the daily return of stocks traded on the Indonesia Stock Exchange that are part of the LQ-45 index from August 2022 to January 2023 was used to establish the weight of the portfolio, while the fundamental data of LQ-45 stocks for 2022 were used to build clusters. Using K-Means and K-Medoids clustering, this study's results show that LQ-45 stocks are divided into six groups. Additionally, it is obtained that for risk aversion $\gamma < 15$, portfolio performance with K-Means clustering is better than portfolio performance with K-Medoids clustering. In contrast, for risk aversion $\gamma \geq 15$, portfolio performance with K-Medoids clustering is better than portfolio performance with K-Means clustering.
**Keywords**: cluster analysis, portfolio, return, risk, Sharpe ratio

## 1 Introduction

A fundamental tenet of the design of the mean-variance portfolio model is the use of statistical measures derived from historical data returns, specifically the mean, variance, and covariance [1]. The portfolio model put out by Markowitz expresses the trade-off between portfolio return and risk using the mean and variance of asset returns. This model is described as a conflicting objective optimization problem. In other words, while portfolio risk, which is represented by the variance of returns from various assets, must be minimized, the expected returns of the portfolio must be maximized.

The Markowitz portfolio model has been solved and developed through a number of studies. All of this was done in order to modify the current model to accommodate for changes in financial market variables and capital market practitioners' needs [2]. The effectiveness of choosing the best portfolio in terms of time is one of the areas of portfolio selection study. This makes sense given that there are more conceivable portfolio structures the more stocks that make up a portfolio. By employing cluster analysis to categorize stock data, it is possible to reduce the enormous number of securities that are involved in portfolio selection [3]. A statistical analysis called a cluster analysis seeks to divide items into a number of groups that coincide or differ from one another in terms of certain criteria. Objects in one cluster will have a closer relationship than those in other clusters [4,5].

---

2020 *Mathematics Subject Classification*: 62H30
Tanggal masuk: 13-09-23; direvisi: 04-10-23; diterima: 30-10-23

Cluster analysis has been widely used in recent portfolio selection research. Guan and Jiang [6] proposed a method for optimizing portfolio selection using clustering approaches. This approach uses the clustering technique to separate the stock data into various groups. Then, to build an effective portfolio, stocks were chosen from each group. According to the study's findings, a portfolio comprised of stocks from each cluster has the lowest risk when compared to other portfolios of the same size for a given level of risk. A portfolio optimization approach based on average linkage and single linkage clustering was proposed by Tola et al. [7]. According to the experimental findings, the clustering method can improve the portfolio's dependability in terms of the ratio of predicted to actual risk.

Using clustering methods and fuzzy optimization models, Chen and Huang [8] proposed portfolio optimization. This method divides stock data into groups using clustering techniques, and then uses a fuzzy optimization model to establish the best investment composition for each group.

The Markowitz model and clustering algorithms are the foundation of Nanda et. al.'s [9] investigation of a portfolio selection model. Stock data is divided into categories using the clustering techniques k-means, fuzzy c-means (FCM), and self organizing maps (SOM). Stocks are chosen from the clusters that have formed to create the portfolio. The stock that has the best performance within a cluster is the stock that is chosen within that cluster. However, situations where particular transient macroeconomic factors have an instantaneous impact on market performance are also taken into consideration. The purpose of the portfolio that was created was to reduce risk and evaluate portfolio results against a benchmark index, the Sensex.

The FCM algorithm and the multi-objective genetic algorithm were used to create a portfolio selection model (MOGA) was proposed by Long et. al. [10]. This method involves classifying the stocks into k clusters, selecting m stocks to represent each cluster, and then using these stocks to build an effective portfolio using MOGA. The findings demonstrate how effective the suggested strategy is at creating an effective portfolio.

The literature study leads to the conclusion that there are three stages that must be completed in order to maximize both time efficiency and the number of stocks that will construct the optimum portfolio. Stocks are first divided into various groupings. The second step is choosing the stocks that will make the optimal portfolio. The proportion of each stock in the optimum portfolio is determined in the third stage. We employ K-Means and K-Medoids clustering in this paper as our contribution to group stocks into several clusters. Our second contribution is creating an optimum portfolio by choosing representative stocks for each cluster according to the Sharpe ratio.

The rest of the paper is organized as follows. We provide the research methodology in Section 2. Results and Discussion are provided in Section 3. Lastly, we provide Conclusions in Section 4.

## 2 Research Methods

### 2.1 Data Source

There are two kinds of data used in this study. The first is the fundamental data stocks of LQ-45 index on 2022 that was obtained through the https://indopremier.com. There are 13 fundamental data measures for each stock taken, namely:

1. Share Out.
2. Market Capital.
3. Total Aset.
4. Total Equity.
5. Revenue.
6. Net Profit.

7. Earning Before Interest, Taxes, Depreciation, and Amortization (EBITDA).
8. Earning Per Share (EPS).
9. Price Earning Ratio (PER),
10. Book Value Per Share (BVPS).
11. Price to Book Value Ratio (PBV).
12. Return on Asset (ROA).
13. Return on Equipment (ROE)

These data are used for clustering LQ-45 stocks.

The second data used in this study is the daily closing price data for LQ-45 stocks for the period August 2022 – January 2023. This data is used to determine portfolios weight.

### 2.2 Portfolio Mean-Variance

In the financial world, the term return becomes a very important part because it can be used to identify the actual price situation. First, for investors, returns clearly describe price changes. Second, for practitioners, returns are theoretically and empirically more attractive in describing statistical properties, for example stationarity and events related to price changes [11].

Let $p_{it}$ is the price of the *i*-th asset at time *t* and it is assumed that there is no distribution of profits (dividends). Return of *i*-th asset for one period, namely from time $t-1$ to $t$ is [12]:

$$r_{it} = \frac{p_{it} - p_{i(t-1)}}{p_{i(t-1)}} = \frac{p_{it}}{p_{i(t-1)}} - 1 \tag{1}$$

Expected return of *i*-th asset is the average of $r_{it}$ for all $t$, that is

$$E[\boldsymbol{r}_i] = \frac{1}{n}\sum_{t=1}^{n} r_{it} = r_i \tag{2}$$

where $n$ is the time period.

Suppose that an investor wants to invest in *m* assets, $r_i$ is the *i*-th asset return, where $i = 1, \dots, m$, $\boldsymbol{r}' = (r_1, \dots, r_m)$ represents the return of each asset in the portfolio. Furthermore, suppose that the stocks return data of portfolio is presented in Table 1.

**Table 1:** Portfolio Return Data

| | $t_1$ | $t_2$ | ... | $t_j$ | ... | $t_n$ |
|---|---|---|---|---|---|---|
| $\boldsymbol{r}_1$ | $r_{11}$ | $r_{12}$ | ... | $r_{1j}$ | ... | $r_{1n}$ |
| $\boldsymbol{r}_2$ | $r_{21}$ | $r_{22}$ | ... | $r_{2j}$ | ... | $r_{2n}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{r}_k$ | $r_{k1}$ | $r_{k2}$ | ... | $r_{kj}$ | ... | $r_{kn}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\boldsymbol{r}_m$ | $r_{m1}$ | $r_{m2}$ | ... | $r_{mj}$ | ... | $r_{mn}$ |

or in the form of a matrix that table can be presented as

$$\boldsymbol{r} = \begin{pmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \vdots \\ \boldsymbol{r}_k \\ \vdots \\ \boldsymbol{r}_m \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1j} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2j} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{k1} & r_{k2} & \cdots & r_{kj} & \cdots & r_{kn} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mj} & \cdots & r_{mn} \end{pmatrix}$$

The expected value of $\boldsymbol{r}$ is:

$$E(\boldsymbol{r}) = \begin{bmatrix} E(\boldsymbol{r}_1) \\ \vdots \\ E(\boldsymbol{r}_m) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix} = \boldsymbol{\mu} \tag{3}$$

and the covariance matrix is

$$\Sigma = \mathrm{E}[(\boldsymbol{r} - \boldsymbol{\mu})(\boldsymbol{r} - \boldsymbol{\mu})'] = E\left( \begin{bmatrix} (r_1 - \mu_1) \\ \vdots \\ (r_p - \mu_m) \end{bmatrix} [(r_1 - \mu_1) \quad \cdots \quad (r_m - \mu_m)] \right)$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix} \tag{4}$$

The portfolio return is the weighted average of the return of each asset in the portfolio [13], that is:

$$R_p = w_1 r_1 + \cdots + w_m r_m = \boldsymbol{w}'\boldsymbol{r} \tag{5}$$

where $w_i$, $i = 1, 2, \ldots, m$ states the proportion (weight) of capital invested in the *i*-th asset which are formulated as

$$\boldsymbol{w}' = (w_1, \ldots, w_m) \text{ and } \boldsymbol{r} = (r_1, \ldots, r_m) \tag{6}$$

In investing on financial assets, investors are assumed to invest all of their capital in assets, so that $\sum_{i=1}^{m} w_i = 1$. In addition, investors are also faced with an element of uncertainty, therefore investors can only estimate the amount of expected return and the probability that the actual results will deviate from the expected results (risk). The expected return rate of a portfolio is the expected value of the portfolio return, namely:

$$\boldsymbol{\mu}_p = E\big(R_p\big) = E(\boldsymbol{w}'\boldsymbol{r}\,) = \boldsymbol{w}'\boldsymbol{\mu} = \textstyle\sum_{i=1}^{m} w_i\, \mu_i \tag{7}$$

While the portfolio variance is:

$$\boldsymbol{\sigma}_p^2 = Var\,\big(R_p\big) = Var\,(\boldsymbol{w}'\boldsymbol{r}) = \boldsymbol{w}'\Sigma\,\boldsymbol{w} = \sum_{i=1}^{m}\sum_{j=1}^{m} w_i w_j\, \sigma_{ij} \tag{8}$$

where $\sigma_{ij}$ is the covariance of asset *i* and asset *j*.

The mean and variance approach are the cornerstone of Markowitz's portfolio theory, where mean denotes expected return level and variance denotes the risk level. Markowitz's portfolio theory is hence also referred to as the mean-variance (MV) model. In order to choose and create the best portfolios, this approach stresses efforts to maximize expected return and minimize risk (variance). By resolving the following optimization issue, the mean-variance portfolio can be created [3]:

$$\max_{w} \boldsymbol{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\boldsymbol{w}'\Sigma\boldsymbol{w} \tag{9}$$

and satisfy the condition

$$\boldsymbol{w}'\boldsymbol{e} = 1 \tag{10}$$

where $\boldsymbol{w}$ stands for portfolio weight, $\boldsymbol{\mu}$ is the mean vector, Σ is the covariance matrix, $\boldsymbol{e}$ is the column matrix where all elements are 1 and $\gamma \geq 0$ is the risk aversion parameter, which is a relative measure of risk aversion. According to [3], the solution of issue (9) and satisfy condition (10) is

$$\boldsymbol{w}(\boldsymbol{\mu}, \Sigma) = \frac{1}{\gamma}(\Sigma^{-1} - \Sigma^{-1}\boldsymbol{e}(\boldsymbol{e}'\Sigma^{-1}\boldsymbol{e})^{-1}\boldsymbol{e}'\Sigma^{-1})\boldsymbol{\mu} + \Sigma^{-1}\boldsymbol{e}(\boldsymbol{e}'\Sigma^{-1}\boldsymbol{e})^{-1} \tag{11}$$

Equation (11) shows that the optimal portfolio ($\boldsymbol{w}$) is determined by the inputs of $\boldsymbol{\mu}$ and Σ.

### 2.2 Stocks Clustering

Because there are so many stocks on the market, determining the investment proportion of each share is difficult. Data mining techniques must be used to address this. One of the data mining techniques available is cluster analysis. Cluster analysis is a statistical technique that attempts to classify objects into groups with similar or dissimilar properties. Each group in this analysis is homogeneous among members, or the variation of objects in the formed group is as small as possible. Objects in one cluster are more similar to those in another.

In the literature, there are numerous cluster techniques. In this study, K-Means and K-Medoids clustering will be used. The two clustering methods will be discussed briefly in the following sections.

#### 2.2.1 K-Means Clustering

K-Means cluster analysis is a cluster partition analysis method that aims to group data into two or more groups. In accordance with the characteristics of cluster partition analysis, in this method, each data must enter into one of the clusters, and it is possible for each object belonging to a certain cluster at one stage of the process to move to another cluster at the next stage.

K-Means cluster analysis partitions *n* objects into k groups or clusters. The value of *k* is predetermined where $k < n$. Each cluster has a mean (average) of the objects in a cluster which is called the centroid (cluster center). The allocation of objects into a cluster is based on the distance between the object and the closest cluster mean. The K-Means clustering method starts with determining the desired *k* value, then generating the k centroid (mean) of the initial cluster that is chosen randomly. Then the objects are allocated to the cluster with the nearest centroid where in the new cluster formed the new centroid is iteratively calculated. According to Jain and Dubes [14], cluster analysis using the K-Means method is generally carried out with the K-Means algorithm as follows:

1. Choose the desired k number of clusters.
2. Set the initial k centroids $(c_1, \dots, c_k)$ randomly.
3. Calculate the distance between each object $(x_j)$ and each centroid $(c_i)$.
4. Group objects by closest distance to centroid $(c_i)$.
5. Calculate the new centroid point, namely the mean of each cluster that has been formed.
6. If the centroid changes, return to step 3. The iteration continues until the centroid does not change or until the objects do not move clusters anymore. The final centroid is obtained $(c_1, \dots, c_k)$ where each object has been grouped into k clusters based on the closest distance to the centroid of each cluster.

The K-Means algorithm attempts to find the cluster's center, namely the centroid $(c_1, \dots, c_k)$ so that the sum of the squares of the distances between objects $(x_j)$ and each centroid $(c_i)$ in the cluster is minimum (within-cluster sum of squares, WSS).

$$d = \sum_{i=1}^{k} \left[ \min_{j=1,\dots,n} D(X_j, C_j) \right]^2 \tag{12}$$

The distance function D in this case is used to measure the object's distance to the centroid. Thus, function D can be written as follows:

$$r_i = E[r_{it}] = \frac{1}{n} \sum_{t=1}^{n} r_{it} \tag{13}$$

where

$X_j$ is the j-th object, $j = 1, \dots, n$

$n$ denotes the number of variables.

$x_{jz}$ is the $z$-th observation in the $j$-th object, where $z = 1, \dots, l$

$l$ indicates the number of observations in each object.

and

$C_i$ is the centroid of the $i$-th cluster, $i = 1, \dots, k$

$k$ indicates the number of clusters. $C_i = \{c_{i1}, \dots, c_{il}\}$

$c_{iz}$ is the centroid (mean) of the $z$-th observation of all cluster members in the $i$-th cluster, $i = 1, \dots, l$.

$c_{iz}$ is calculated using the following formula:

$$c_{iz} = \frac{\sum_{j=1}^{n_i} x_{jz}}{n_i} \tag{14}$$

$n_i$ is the number of objects located in the i-th cluster.

In step 4, the K-Means method allocates objects to each cluster based on the distance between the object and the centroid of each existing cluster. Objects are included in the cluster with the centroid that is closest to the object. Allocation can be expressed as follows:

$$a_{ji} = \begin{cases} 1, & if\ d = \min\{d(X_j, C_i)\} \\ 0, & \text{others} \end{cases} \tag{15}$$

where $D(X_j, C_i)$ is the Euclidean distance of the $j$-th object to the $i$-th cluster centroid. $a_{ji}$ is the membership value of the $j$-th object in the $i$-th cluster. If a $j$-th object is a member of an $i$-th cluster, then the value $a_{ji} = 1$, otherwise $a_{ji} = 0$.

### 2.2.2 Clustering K-Medoids

K-Medoids cluster analysis, like K-Means cluster analysis, divides a data set into k groups or clusters [15]. Each cluster in the K-Medoids clustering is represented by a cluster data point. These points are referred to as cluster medoids. A medoid is a cluster entity with the smallest average difference between it and all other cluster members. This is the cluster's epicenter. These entities can be viewed as representative examples of cluster members that may be useful in certain situations. Remember that the center of a cluster is determined by the average value of all cluster data points when using k-means cluster analysis. K-Medoid is a reliable substitute for K-Means cluster analysis. Because the medoid is used as the cluster's center, the algorithm is less sensitive to outliers than K-Means. The most common cluster analysis approach for K-Medoids is Partitioning Around Medoids (PAM) [15].

The PAM algorithm searches a data set of observations for k representative objects or medoids. After locating a collection of k medoids, clusters are formed by assigning each observation to the closest medoid. The objective function is then calculated by swapping each selected medoid and non-medoid data point. The sum of all object dissimilarities to their nearest medoid is the objective function. The exchange phase tries to improve cluster analysis efficiency by swapping selected (medoids) and unselected objects. If the objective function can be reduced by swapping selected objects with unselected objects, a swap is performed. This process is repeated until the objective function can no longer be deduced. The goal is to find k representative objects that have the fewest observable differences from their nearest representative objects.

Kaufman and Rousseeuw [15] describe the steps of the PAM cluster analysis algorithm in detail as follows:

1. Choose $k$ objects to be medoids, or use these objects as medoids if they are provided;
2. If the dissimilarity matrix is not available, calculate it;

3. Assign each object to the medoid closest to it;
4. Determine whether one of the cluster objects reduces the average dissimilarity value; if so, choose the entity that reduces the most as the medoid for this cluster;
5. Return to step 3 if at least one medoid has changed; otherwise, stop the algorithm.

### 2.3 Sharpe Ratio

Following the formation of the clusters, the performance of each stock in each cluster is evaluated using the Sharpe ratio (*SR*). The Sharpe ratio, also known as the Sharpe index, is a measure of the excess return in assets per unit of risk. The Sharpe ratio measures how well investors are compensated by the return on assets for the risk they are taking. The Sharpe ratio is calculated by dividing the difference between the stock return r and the risk return free rate ($r_f$) by the standard deviation of stock returns ($\sigma$), or it can be written as [16]:

$$SR = \frac{r - r_f}{\sigma} \tag{16}$$

The higher a stock's Sharpe ratio, the better its performance.

By replacing stock return r with portfolio return ($r_p$) and standard deviation of stock ($\sigma$) portfolio risk ($\sigma_p$), the Sharpe ratio can also be used to measure portfolio performance.

### 2.4 Procedures

The following procedures were used to conduct this study:

1. Collect fundamental data stocks of LQ-45 index for 2022 through https://indopremier.com.
2. Collect daily closing price data stocks of LQ-45 index for the period August 2022 - January 2023 through https://finance.yahoo.com.
3. Collect data Bank Indonesia rate for 2022 through https://www.bi.go.id
   Bank Indonesia rate data is used as the risk return free rate.
4. Calculate the return, risk and Sharpe ratio for each stock based on the data obtained in step 2.
5. Clustering LQ-45 stocks using K-Means and K-Medoids clustering based on the data obtained in step 1.
   The number of clusters in the study was 6 clusters for each clustering method.
6. Choose a representative stock for each cluster for both clustering methods.
   Stocks with the highest Sharp ratio were chosen to represent the clusters.
7. Determine the covariance matrix of the portfolio's stocks for two clustering methods.
8. Determine the return and risk of portfolios.
9. Determine weight of the portfolios formed
10. Determine the performance of the portfolio constructed using both clustering methods.
    The performance of the two formed portfolios is measured using the Sharpe ratio

## 3 Results and Discussion

### 1.1. Clustering Results

The cluster analysis used in this study was K-Means and K-Medoids clustering. Using *cluster* package in R 3.6.1, it was found that LQ-45 stocks were clustered into 6 clusters using K-Means and K-Medoids clustering, as shown in Table 2 and Table 3.

**Table 2:** Cluster of Stocks with K-Means

| Cluster | Stocks | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | BBTN | BFIN | BMRI | BRIS | BRPT | BUKA | CPIN | |
| 2 | ADRO | AMRT | ANTM | ARTO | ASII | BBCA | BBNI | BBRI |
| 3 | INCO | INDF | INDY | INKP | INTP | ITMG | JPFA | |
| 4 | EMTK | ERAA | EXCL | GOTO | HMSP | HRUM | ICBP | |
| 5 | TBIG | TINS | TLKM | TOWR | TPIA | UNTR | UNVR | WIKA |
| 6 | KLBF | MDKA | MEDC | MIKA | MNCN | PGAS | PTBA | SMGR |

**Table 3:** Cluster of Stocks with K-Medoids

| Cluster | Stocks | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ADRO | AMRT | ANTM | ARTO | ASII | BBCA | BBNI | BBRI | BBTN |
| 2 | BFIN | BMRI | BRIS | BRPT | BUKA | CPIN | EMTK | ERAA | EXCL |
| 3 | GOTO | HMSP | HRUM | ICBP | INCO | INDF | INDY | INKP | |
| 4 | INTP | ITMG | JPFA | KLBF | MDKA | MEDC | MIKA | | |
| 5 | MNCN | PGAS | PTBA | SMGR | TBIG | TINS | | | |
| 6 | TLKM | TOWR | TPIA | UNTR | UNVR | WIKA | | | |

Following the formation of these clusters, the Sharpe ratio is computed for each stock in each cluster produced by the two clustering methods. The risk return free rate used in the Sharpe ratio calculation is the Bank Indonesia rate for 2022, which is 4.0% per year.

Using K-Means clustering, in cluster 1, when compared to the other stocks in the cluster, BMRI has the best performance, as indicated by the cluster's highest Sharpe ratio of 0.09566. As a result, BMRI stock was chosen to represent Cluster 1. Then, in cluster 2, AMRT stock with Sharpe ratio of 0.12326 represent the cluster. And so on, MEDC shares with a Sharpe ratio of 0.17017 represent Cluster 6.

Clustering using K-Medoids, on the other hand, discovered that in cluster 1, AMRT stock has outperformed the other stocks in the cluster, as indicated by the highest Sharpe ratio in the cluster, which is 0.12326. As a result, AMRT stock is used to represent Cluster 1. Then, in cluster 2, BMRI stock represents cluster 2 with a Sharpe ratio of 0.09566. And so forth, UNVR stock with a Sharpe ratio of 0.00133 belongs to Cluster 6.

If we pay further attention, it turns out that there are four stocks which besides being cluster representation in the K-Mean clustering are also cluster representative in the K-Medoids clustering. These stocks are BMRI, AMRT, UNVR, and MEDC.

Table 4 and Table 5 show the representation of each cluster in the two clustering methods in full detail.

**Table 4**: Stock Representation of Cluster with K-Means Clustering

| Cluster | Representation | Return | Risk | Sharpe Ratio |
|---|---|---|---|---|
| 1 | BMRI | 0.00184 | 0.01607 | 0.09566 |
| 2 | AMRT | 0.00351 | 0.02596 | 0.12326 |
| 3 | INCO | 0.00147 | 0.02263 | 0.05118 |
| 4 | EXCL | 0.06468 | 0.82019 | 0.07849 |
| 5 | UNVR | 0.00033 | 0.01690 | 0.00113 |
| 6 | MEDC | 0.00698 | 0.03922 | 0.17017 |

**Table 5:** Stock Representation of Cluster with K-Medoids Clustering

| Cluster | Representation | Return | Risk | Sharpe Ratio |
|---|---|---|---|---|
| 1 | AMRT | 0.00351 | 0.02596 | 0.12326 |
| 2 | BMRI | 0.00184 | 0.01607 | 0.09566 |
| 3 | ICBP | 0.00108 | 0.01411 | 0.05471 |
| 4 | MEDC | 0.00698 | 0.03922 | 0.17017 |
| 5 | SMGR | 0.00125 | 0.01915 | 0.04907 |
| 7 | UNVR | 0.00033 | 0.01690 | 0.00113 |

## 3.2 The Comparison of Portfolios Performance

The MV portfolio model is used in this study to determine the optimal portfolio. The initial step is to calculate portfolio weightings for different levels of risk aversion $\gamma$. Stocks that represent each cluster for the two clustering methods are used, as shown in Table 2 and 3. Table 6 and Tables 7 show the portfolio weights that resulted from the two clustering methods.

**Table 6:** Portfolio Weight with K-Means Clustering

| $\gamma$ | BMRI | AMRT | INCO | EXCL | UNVR | MEDC |
|---|---|---|---|---|---|---|
| 0.5 | -1.97841 | 6.11290 | -1.81195 | 0.17791 | -8.04430 | 6.54386 |
| 1 | -0.81753 | 3.10331 | -0.84211 | 0.08891 | -3.83737 | 3.30478 |
| 2 | -0.23708 | 1.59852 | -0.35720 | 0.04441 | -1.73390 | 1.68524 |
| 5 | 0.11118 | 0.69565 | -0.06625 | 0.01771 | -0.47181 | 0.71352 |
| 10 | 0.22727 | 0.39469 | 0.03074 | 0.00881 | -0.05112 | 0.38961 |
| 15 | 0.26597 | 0.29437 | 0.06307 | 0.00585 | 0.08911 | 0.28164 |
| 20 | 0.28531 | 0.24421 | 0.07923 | 0.00437 | 0.15923 | 0.22766 |
| 25 | 0.29692 | 0.21411 | 0.08893 | 0.00348 | 0.20130 | 0.19527 |
| 30 | 0.30466 | 0.19405 | 0.09539 | 0.00288 | 0.22934 | 0.17367 |

**Table 7:** Portfolio Weight with K-Medoids Clustering

| $\gamma$ | AMRT | BMRI | ICBP | MEDC | SMGR | UNVR |
|---|---|---|---|---|---|---|
| 0.5 | 5.87584 | -1.00589 | -3.72310 | 6.96901 | -0.76188 | -6.35397 |
| 1 | 2.99140 | -0.41307 | -1.68873 | 3.49734 | -0.28993 | -3.09701 |
| 2 | 1.54918 | -0.11665 | -0.67154 | 1.76150 | -0.05395 | -1.46854 |
| 5 | 0.68384 | 0.06119 | -0.06123 | 0.72000 | 0.08764 | -0.49145 |
| 10 | 0.39540 | 0.12048 | 0.14221 | 0.37283 | 0.13483 | -0.16575 |
| 15 | 0.29925 | 0.14024 | 0.21002 | 0.25711 | 0.15057 | -0.05719 |
| 20 | 0.25118 | 0.15012 | 0.24393 | 0.19925 | 0.15843 | -0.00291 |
| 25 | 0.22233 | 0.15605 | 0.26427 | 0.16453 | 0.16315 | 0.02966 |
| 30 | 0.20310 | 0.16000 | 0.27784 | 0.14139 | 0.16630 | 0.05138 |

In the portfolio weighting using the K-Mean clustering results as presented in Table 6, it can be seen that for risk aversion $\gamma = 0.5$, the stock weight with the highest Sharpe ratio, namely AMRT stock has the largest weight, which is 6.11290, while the stock weight with the lowest Sharpe ratio, namely UNVR stocks has the smallest weight of -8.04430 (short selling). When $\gamma < 15$, as the value of risk aversion rises $\gamma$, so will the weight of each stock in the portfolio and the weight of all stocks making up the portfolio becomes positive when $\gamma = 15$. The portfolio weighting using the K-Medoids clustering results received the same fate, as shown in Table 7.

Based on weights, mean vector, and covariance matrix of the stocks that construct the portfolio, then return, risk, and Sharpe ratio of the two portfolios can be determined as presented in Table 8 and Table 9.

**Table 8:** Return, Risk, and Sharpe Ratio of Portofolio with K-Means Clustering

| $\gamma$ | Return | Risk | Sharpe Ratio |
|---|---|---|---|
| 0.5 | 0.06971 | 0.36891 | 0.18814 |
| 1 | 0.03572 | 0.18469 | 0.19173 |
| 2 | 0.01872 | 0.09281 | 0.19840 |
| 5 | 0.00852 | 0.03840 | 0.21391 |
| 10 | 0.00512 | 0.02133 | 0.22578 |
| 15 | 0.00399 | 0.01631 | 0.22577 |
| 20 | 0.00342 | 0.01414 | 0.22037 |
| 25 | 0.00308 | 0.01301 | 0.21333 |
| 30 | 0.00286 | 0.01235 | 0.20630 |
| 50 | 0.00240 | 0.01133 | 0.18488 |

| | | | |
|---|---|---|---|
| 100 | 0.00206 | 0.01087 | 0.16143 |
| 1000 | 0.00176 | 0.01072 | 0.13523 |

**Table 9:** Return, Risk, and Sharpe Ratio of Portfolio with K-Medoids Clustering

| $\gamma$ | Return | Risk | Sharpe Ratio |
|---|---|---|---|
| 0.5 | 0.06037 | 0.34313 | 0.17505 |
| 1 | 0.03096 | 0.17172 | 0.17848 |
| 2 | 0.01625 | 0.08616 | 0.18500 |
| 5 | 0.00742 | 0.03529 | 0.20160 |
| 10 | 0.00448 | 0.01905 | 0.21905 |
| 15 | 0.00350 | 0.01412 | 0.22605 |
| 20 | 0.00301 | 0.01192 | 0.22657 |
| 25 | 0.00272 | 0.01076 | 0.22381 |
| 30 | 0.00252 | 0.01007 | 0.21968 |
| 50 | 0.00213 | 0.00897 | 0.20287 |
| 100 | 0.00183 | 0.00846 | 0.18024 |
| 1000 | 0.00157 | 0.00829 | 0.15207 |

Portfolio performance with K-Mean clustering as presented in Table 8 can be seen that for $\gamma = 0.5$ the Sharpe ratio is 0.18814, then increases to 0.19173 for $\gamma = 1$, increases to 0.19840 for $\gamma = 2$, continues to increase to 0.22578 for $\gamma = 10$. Sharpe ratio starts to decrease for $\gamma = 15$, namely to 0.22577, decreases again to 0.22037 for $\gamma = 20$, decreases to 0.20630 for $\gamma = 30$, continues to decrease to 0.18488 for $\gamma = 50$, decreases to 0.16143 for $\gamma = 100$ and finally becomes 0.13523 for $\gamma = 1000$.

On the other hand, the resulting portfolio performance with K-Medoids clustering is as presented in Table 9, for $\gamma = 0.5$ the Sharpe ratio is 0.17505, then increases to 0.17848 for $\gamma = 1$, increases to 0.18500 for $\gamma = 2$, continues to increase to 0.22657 for $\gamma = 20$. The Sharpe ratio starts to decrease for $\gamma = 25$, becomes 0.22381, decreases again to 0.21968 for $\gamma = 30$, continues to decrease to 0.20287 for $\gamma = 50$, decreases to 0.18024 for $\gamma = 100$ and finally becomes 0.15207 for $\gamma = 1000$. In general, it can be concluded that risk aversion $\gamma < 15$, portfolio performance with K-Means clustering is better than portfolio performance with K-Medoids clustering. In contrast, for risk aversion $\gamma \geq 15$, portfolio performance with K-Medoids clustering is better than portfolio performance with K-Means clustering.

## 4 Conclusions

This paper shows how to integrate clustering techniques into portfolio management and develop a system for obtaining the best portfolio. By using clustering techniques, Stocks in similar categories can be easily grouped together to form a cluster. When choosing stocks, this can save a significant amount of time. To build the portfolio, the best performing stocks from each cluster are chosen as cluster representatives. The findings revealed that 45 stocks on the Indonesia Stock Exchange included in the LQ-45 index were grouped into six clusters using K-Means and K-Medoids clustering. Using the MV portfolio model, the stocks representing each cluster are then combined to form a portfolio. Portfolio performance is compared by combining both clustering techniques and the MV portfolio model.

The research results show that for risk aversion $\gamma < 15$, portfolio performance with K-Means clustering is better than portfolio performance with K-Medoids clustering. In contrast, for risk aversion $\gamma \geq 15$, portfolio performance with K-Medoids clustering is better than portfolio performance with K-Means clustering.

## 5 Acknowledgment

We thank the Head of the Mathematics Laboratory at FMIPA Universitas Halu Oleo for granting us access to their facilities for the purpose of conducting this research.

## References

[1] H. M. Markowitz, "Portfolio Selection", *Journal of Finance*, vol. 7, pp. 77-91, 1952. https://www.math.hkust.edu.hk/~maykwok/courses/ma362/07F/markowitz_JF.pdf

[2] L. Gubu, D. Rosadi, and Abdurakhman, "Classical Portfolio Selection with Cluster Analysis: Comparison Between Hierarchical Complete Linkage and Ward Algorithm", *in Proc. The Eighth SEAMS-UGM International Conference on Mathematics and Its Applications*, *AIP Conference Proceedings 2192*, pp. 090004-1–090004-7, 2019 https://aip.scitation.org/doi/abs/10.1063/1.5139174

[3] L. Gubu, D. Rosadi, and Abdurrakhman, "Robust Mean-Variance Portfolio Selection Using Cluster Analysis: A Comparison Between KAMILA and Weighted k-mean Clustering", *Asian Economic and Financial Review*, vol. 10, no. 10, pp. 1169–1186, 2020. http://www.aessweb.com/journals/October2020/5002/5155

[4] P. Rai and S. Singh, "A Survey of Clustering Techniques". *International Journal of Computer Applications*, vol. 7, no. 12, pp. 1-5, 2010. https://www.researchgate.net/publication/49586251_A_Survey_of_Clustering_Techniques

[5] R. Xu and D. C. Wunsch, *Clustering*, Hoboken, New Jersey, John Wiley & Sons Inc., 2009. https://www.researchgate.net/publication/245507626_Clustering_Xu_R_and_Wunsch_DC_2008_Book_review

[6] H. S. Guan and Q. S. Jiang, "Cluster Financial Time Series for Portfolio", *in Proc. Int. Conf. on Wavelet Analysis and Pattern Recognition*, ICWAPR, Beijing, China, pp. 851–856, 2007. https://www.semanticscholar.org/paper/Cluster-financial-time-series-for-portfolio-Guan-Jiang/9996584f72dccc607d6df9dc75f054264a75ab31

[7] V. Tola, F. Lillo, M. Gallegati, and R. N. Mantegna, "Cluster Analysis for Portfolio Optimization", *J. Econ. Dyn. Control*, vol. 32, no. 1, pp. 235–258, 2008. https://www.sciencedirect.com/science/article/abs/pii/S0165188907000462

[8] L. H. Chen and L. Huang, "Portfolio Optimization of Equity Mutual Funds with Fuzzy Return Rates and Risks", *Expert Systems with Applications*, vol. 36, pp. 3720-3727, 2009. https://www.sciencedirect.com/science/article/abs/pii/S0957417408001528

[9] R. Nanda, B. Mahanty, and M. K. Tiwari, "Clustering Indian Stock Market Data for Portfolio Management, *Expert Syst. Appl.* vol. 37, no. 12, pp. 8793–8798, 2010. https://www.sciencedirect.com/science/article/abs/pii/S0957417410005300

[10] N. C. Long, N. Wisitponghan, and P. Meesad, "Clustering Stock Data for Multi-Objective Portfolio Optimization", *International Journal of Computational Intelligence and Applications*, vol. 13, no. 2, pp. 1-13, 2014. https://www.worldscientific.com/doi/abs/10.1142/S1469026814500114

[11] Sukono, "Measurement of Value-at-risk (VaR) with Inconstant Volatility and Long Memory Effect", Ph.D. dissertation, Gadjah Mada University, Yogyakarta, Indonesia, 2011.

[12] E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis*, 9th Edition, New York, John Wiley and Sons, Inc., 2014. http://dl.rasabourse.com/Books/Finance%20and%20Financial%20Markets/%5BEdwin_J._Elton%2C_Martin_J._Gruber%2C_Stephen_J._Brow_Modern%20Portfolio%20Theory%20and%20Investment%28rasabourse.com%29.pdf

[13] E. D. Supandi, "Developing of Mean-Variance Portfolio Modeling Using Robust Estimation and Robust Optimization Method", Ph.D. dissertation, Gadjah Mada University, Yogyakarta, Indonesia, 2017.

[14] A. Jain and R. Dubes, *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice Hall, 1988. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1874615

[15] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990. https://www.amazon.com/Finding-Groups-Data-Introduction-Analysis/dp/0471735787

[16] W. F. Sharpe, "The Sharpe Ratio", *The Journal of Portfolio Management*, vol. 21, pp. 49–58, 1994. https://www.scirp.org/(S(czeh2tfqw2orz553k1w0r45))/reference/referencespapers.aspx?referenceid=1451308