Clustering Approach for Portfolio Optimization

Ana Paula dos Santos Gularte[a,c,1], Felipe dos Santos Alves Feitosa[a,b], Vinícius Henrique Pinto[a,b], Vitor Venceslau Curtis[b], Global Customer Service[1,*]

[a] *Department of Aerospace Science and Technology, Aeronautics Institute of Technology (ITA), São José dos Campos, SP - Brazil*
[b] *Computer Science Division, Aeronautics Institute of Technology (ITA), São José dos Campos, SP - Brazil*
[c] *Department of Science and Technology, Federal University of São Paulo (UNIFESP), São José dos Campos, SP - Brazil*

Abstract

This template helps you to create a properly formatted LaTeX manuscript.

Keywords: `Cluster Analysis, Portfolio Optimization, Dimensionality Reduction`

Highlights/Originality

- A novel methodology for reducing dimensionality in the Mean-Variance (MV) model.
- A preliminary step for selecting more stable assets with cluster techniques.
- Portfolio allocation by Hierarchical Risk Parity (HRP) and compare them to the from Quadratic optimization, as represented by Mean-Variance; and (2) traditional risk parity, exemplified by the Inverse-Variance Portfolio (IVP).
- Innovative portfolio stability and risk adjustment evaluation by comparing the HRP, MV, and IVP in-sample and out-of-sample models via Monte Carlo.

*Corresponding author
Email address: `gularte@ita.br (A. Gularte)` (Global Customer Service)

1. Introduction

Portfolio optimization through the selection and proper allocation of stocks is one of the most recurrent issues in modern financial research, as it is desirable to generate profits in the most diverse market scenarios to minimize losses during downturns (Kalayci et al., 2019), (Elton et al., 2013), (Prado, 2018), (Bodnar et al., 2017). The classic models are widely used, such as the ones proposed by Markowitz (1952), particularly the Mean-Variance (MV) - which achieves an optimal allocation of a financial portfolio through diversification by measuring the expected risk of assets for a target return.

The Mean-Variance model was the genesis for numerous research studies that extended Markowitz (1952, 1959) work and supplemented many insights into portfolio formation. Here we refer to some of these studies that address the challenges encountered when using portfolio optimization in practice, including boundary and cardinality constraints, transaction costs, and the sensitivity in estimates of expected returns and covariances (Tobin, 1958), (Sharpe, 1963), (Merton, 1969), (Ruiz-Torrubiano & Suarez, 2010), (Tu & Zhou, 2010), (Brown & Smith, 2011), (Li et al., 2013),(Jurczenko, 2015), (Cesarone & Tardella, 2016), (Bodnar et al., 2017), (Prado, 2018), (Wang et al., 2020), (Shimizu & Shiohama, 2020), among others. Such examples do not necessarily indicate that the risk-return optimization theory is flawed. Rather, it means that the classical framework must be modified when used in practice to achieve reliability, stability, and robustness concerning model estimates and errors.

These extensions further confirm that the MV model plays a significant role in portfolio management. During the investment decision-making process, it would be untenable to apply only complex portfolio optimization methods without the input of high-quality assets, a step before portfolio allocation; however, few researchers perform preliminary asset selection (Wang et al., 2020), (Deng & Min, 2013). Tayali (2020) warns that many assets in a portfolio can have ramifications due to the curse of dimensionality and high transaction costs.

In this context, preliminary asset selection is fundamental for portfolio man-

agement, and a cardinality constraint overcomes this obstacle that will impose an upper bound on the number of assets. Although this type of optimization problem has advantages over its relaxation, the model inherits computational difficulties because the cardinality constraint transforms the problem into a mixed-integer program of an NP-complete class, as evidenced in the research of (Khan et al., 2020), (Ruiz-Torrubiano & Suarez, 2010).

Because of the reported difficulty, an alternative approach to the cardinality constraint is to reduce the size of the problem before moving on to the optimal portfolio allocation. Reduction, in this case, is a decision to select certain assets from a larger universe. In related literature, recent developments in machine learning have brought significant opportunities for integrating clustering methods as a size reduction or preprocessing tool for the MV model. The results of these studies suggest greater efficiency for the output of the portfolio optimization model by the clustering algorithm having the potential to minimize further the measured risk in the MV model and the ability to improve portfolio reliability in terms of the ratio of predicted to realized risk (Tola et al., 2008), (Tayali & Tolun, 2018; Tayali, 2020). As a powerful replacement for the cardinality constraint in the MV model, clustering methods not only satisfy asset selection and portfolio diversification but also increase the reliability of the portfolio, which is affected by errors in the sample mean and standard deviation estimators of returns (Ren, 2005), (Tola et al., 2008). Still, in this sense, methods based on dimension reduction try to preserve, in lower dimensional representations, the information present in the original data set. This feature is present in both linear and nonlinear methods; the latter, a result of development in our research adopted the recent approach called Uniform Manifold Approximation and Projection (UMAP), which estimates a high dimensional data topology and uses these features to build a low dimensional representation, accurately preserving both local and global data structure relationships, with shorter execution time (McInnes et al., 2018, 2020), (Becht et al., 2019), (Dorrity et al., 2020). The numerical experiments present in Lopes & Machado (2021), Pealat et al. (2022) highlight the feasibility and effectiveness of UMAP

in processing data in complex systems such as the financial market.

Kalayci et al. (2019) conducted a comprehensive review of the literature devoted to mean-variance portfolio optimization in recent articles from the last two decades and evidenced that machine learning algorithms represent 12% of the solutions applied to the MV problem, with the K-means technique prominently in the most present subcategory of research; therefore they have not yet reached an adequate level of maturity. Automated asset clustering methods for diversification purposes are recent innovations and present a gap for future developments since the explicit knowledge of the MV model concerning performance measurement is still limited (Marvin, 2015), (Tayali, 2020). For example, Ren (2005), in portfolio construction, used clustering methods to group highly correlated stocks and then used these clusters to run the mean-variance (MV) portfolio. Marvin (2015), meanwhile, proposed a clustering approach with an alternative measure to correlation similarity, which proved robust in times of crisis, resulting in high-performing portfolios tested in pre and post-crisis periods. Paiva et al. (2019) proposed an investment decision model that uses Support Vector Machines (SVMs) to classify assets and combines them with the mean-variance (MV) model to form an optimal portfolio; according to the results, the classifier showed higher discriminatory power, converging positively to a lower cardinality, with a daily average of seven assets in the portfolio. Tayali (2020) incorporates three cluster analysis methods into the mean-variance portfolio optimization model for the pre-selection of assets. A representative stock is selected taken from each cluster that forms a set of medoids to make up the input subset of the MV problem; the results show that using the clustering method with the Euclidean distance pattern significantly improves the portfolio selection and allocation process of the optimization model. Wang et al. (2020) studied a hybrid method combining a recurrent Long Short-Term Memory (LSTM) neural network with the MV model, which optimizes portfolio formation in combination with asset pre-selection. This research has shown that merging machine learning methods in the asset pre-selection stage with the MV optimization model can provide a new perspective for research in finance.

Two stages divide the essence of this research: asset dimension reduction and portfolio optimization. The first stage involves integrating the UMAP method in the dimensional transformation of the time series into a new input for the clustering models. Then we apply hybrid algorithms to cluster the assets using the methods: i) K-means, ii) Partition Around Medoids (PAM), and iii) Agglomerative Hierarchical Clustering to maximize the objective function composed of fundamental and technical data, and finally to compose the input subset of the MV problem. In the second stage, we use the assets pre-selected in the previous step and perform the allocation with the MV model, comparing the results of the optimal portfolio with two other models: i) Inverse-Variance Portfolio (IVP), and ii) Hierarchical Risk Parity (HRP). A backtesting framework follows in each continuous window of the investment horizon via Monte Carlo simulation and careful analysis of the portfolio's financial performance with out-of-sample data. As a result of this process, we sought to reduce the number of assets to circumvent the cardinality constraint and provide better input inputs to the optimization models. Also, in this sense, to offer investors more stable and diversified portfolios with better Sharpe ratios in the out-of-sample results (León et al., 2017), (Prado, 2016), (Prado, 2018).

The remainder of this paper is structured as follows: we review the Modern Portfolio Theory literature from the overview of Mean-Variance (MV), Inverse-Variance Portfolio (IVP) and Hierarchical Risk Parity (HRP) methods for portfolio optimization in Section2. Section 3 reviews the Uniform Manifold Approximation and Projection (UMAP) method, which acts as a dimensionality reducer, followed in hybrid form by the K-means, Partition Around Medoids (PAM) and Agglomerative Hierarchical Clustering) clustering models for asset pre-selection. We describe our methodology in detail, i.e., data description, selection of input variables, and architecture of the proposed model in Section 4. We present the results in each step developed and explain the dimensionality reduction and clustering methods used for asset pre-selection, the portfolio cardinality, and the financial performance of portfolio optimization in the three proposed scenarios in Section 5. Finally, we present our conclusions and recom-

mend future research in Section 6.

2. Modern Portfolio Theory

In Markowitz (1952, 1959) we find the introduction of what became known in the literature as Modern Portfolio Theory (MPT) through the proposition of the classic Mean-Variance mathematical model (Elton et al., 2013). Approached by quadratic programming, the Markowitz (1952) model is a particular case where $b = 0$, where $b \in R^n$ and $\frac{1}{2}Q$, where $Q \in R^{n \times n}$ is a covariance matrix of returns (which is always at least positive semi-definite), where σ_{ij} is the covariance between assets (i) and (j). The objective function f of the form: $f : R^n \to R$ is equal to the variance σ^2, defined as the risk. The model solves an optimization problem in two-dimensional space in an attempt to optimize a portfolio through the relationship between the variables return (mean) and volatility (variance), finding the weights w_i of the assets in portfolio p that minimize the variance σ_p^2 given a higher average return expectation of the (E_p) or maximize (E_p) at a lower expected variance level. Thus, under various risk expectations, there are different optimal portfolios, suggesting generation of a "Efficient Frontier"(Markowitz, 1952). The objective function Eq.1 minimizes the risk of a portfolio using the covariance matrix of the estimated asset returns, where w represents the allocation rates of N available assets in the market.

$$min\ \sigma_p^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i \cdot w_j \cdot \sigma_{ij} \leq \sigma^2 max \tag{1}$$

Subject to the constraints:

$$E_p = \sum_{i=1}^{N} w_i \cdot m_i \tag{2}$$

$$\sum_{i=1}^{N} w_i = 1 \tag{3}$$

The constraint Eq. 2 defines the portfolio's expected return, where m_i is the average return on assets (i) subject to a maximum risk $\sigma^2 max$ in Eq. 1. The

equality constraint Eq. 3 comes from the need for the sum of all asset fractions to equal 1, i.e., indicates that we want to use the entire amount available in investments.

In the next moments in MPT's history, in a context reported by (Prado, 2016) where return forecasts do not obtain sufficient accuracy, quadratic optimization models such as MV opened possibilities for the emergence of new portfolio allocation methodologies, especially those that had in the covariance matrix its main modeling object, among them stands out the Risk Parity (Jurczenko, 2015). Moreover, the 2008 global crisis popularized the Risk Parity financial model from a theoretical and practical standpoint. Within this context, the model seeks to diversify risk rather than capital among assets, restricting them so that they can contribute equally to the overall volatility of the portfolio, being less sensitive to errors in parameter estimates, such as those related to the covariance matrix Asness et al. (2012); Qian (2005), however, it does not avoid problems of instability in the face of a high number of conditions (Bailey & Prado, 2013).

The Inverse-Variance Portfolio (IVP) is a classic framework comparable to Risk Parity that assigns each asset a weight w_i inversely proportional to its variances σ_i^2. Thus, smaller w_i are given for high variance assets and larger w_i for low variance assets, demonstrating that the weights obtained via MV are generally proportional to the inverse variance σ_i^2 Clarke et al. (2013), these being calculated by Eq. 4:

$$w_i = \frac{\sigma_i^{-2}}{\sum_{i=1}^{N} \sigma_i^{-2}} \tag{4}$$

In addition to the traditional approaches cited above, Prado (2016, 2018) introduced Hierarchical Risk Parity (HRP), which applies state-of-the-art mathematics, including graph theory and unsupervised machine learning, involved in this study. The methodology uses the IVP in a quasi-diagonal matrix by dividing the weights into bisections of a subset, and this conceptually means

that HRP iteratively calculates the weights of groups of similar assets using an inverse variance. Therefore, this portfolio optimization approach breaks down into three steps.In the following, we describe each step in more detail.

- **Tree Clustering**

  The first step consists in the construction of hierarchical structure of clusters, which is based on the correlation coefficient ρ_{ij} between the assets of a matrix $A = N \times N$, where $\rho_{ij} = \frac{COV(X_1,X_2)}{\sqrt{V(X_1)V(X_2)}}$. From ρ_{ij} there is the definition of the measure d_{ij}, defined by $d_{ij} = d[X_i, X_j] = \sqrt{\frac{1-\rho_{ij}}{2}}$, such that $(X_i, X_j) \subset B \to R \in [0, 1]$, where B is Cartesian product of the items in $\{1, ..., i, ..., N\}$. Thus, we have a space D of Euclidean distances between distances, which is $\widetilde{d}_{ij} = \widetilde{d}[D_i, D_j] = \sqrt{\sum_{i=1}^{N} (d_{n,i} - d_{n,j})^2}$, such that $(D_i, D_j) \subset B \to R \in [0, \sqrt{N}]$.

  Finally, single-linkage agglomerative nesting is applied to arrange the clusters between columns (i^*, j^*), seeking at each iteration of the algorithm until all N-1 clusters obey the objective function $(i^*, j^*) = argmin_{i \neq j}^{(i,j)} \widetilde{d}_{ij}$. The clusters formed can be identified using a Lopez dendrogram from (Prado, 2016).Thus, the algorithm recursively combines the portfolio assets into clusters and updates the distance matrix until a single group is formed.

- **Quasi-Diagonalization**

  In the second step, the quasi-diagonalization is the reordering of rows and columns of the covariance matrix V_i for that the most similar assets, which present the highest covariances V_i among themselves, are ordered along its diagonal, forming the subset $L_i \in L$ (list of all assets). The clusters are part of this organization, the order of its creation is in the first step, and the inverse-variance allocation is optimal for a diagonal covariance matrix (Prado, 2016, 2018).

- **Recursive Bisection**

  As the last step of the HRP, the assignment of asset weights is called

recursive bisection, which comprises the following procedure:

Step 1: The assignment of top-down weights w for each asset:

$$w_n = 1, \ \forall n \ = \ \{1, ..., N\} \tag{5}$$

Step 2: Recursively the subset L_i must be split into two new subsets $L_i^{(1)}$ and $L_i^{(2)}$, i. e., into two sub-clusters.

Step 3: For each subset $L_i^{(1)}$ and $L_i^{(2)}$, the respective variances $\widetilde{V}_i^{(1)}$ and $V_i^{(2)}$ are calculated, using the inverse (bottom-up) allocation weights, since it is a quasi-diagonal matrix.

Step 4: Then, a weighting factor or also called by Prado (2016) as split factor α_i, such that $\alpha_{(i)} \in [0, 1]$:

$$\alpha_i \quad = 1 - \frac{\widetilde{V}_i^{(1)}}{\widetilde{V}_i^{(1)} + \widetilde{V}_i^{(2)}} \tag{6}$$

In this procedure, the quasi-diagonal top-down matrix is used to divide the weights in inverse proportion to the cluster variance (Prado, 2016, 2018).

Step 5: Using α_i, we recalculate the asset weights of $L_i^{(1)}$ and $L_i^{(2)}$):

$$w_n = \alpha_i * w_n, \ \forall n \ \in L_i^{(1)} \tag{7}$$

$$w_n = (1 - \alpha_i) * \ w_n, \forall n \ \in L_i^{(2)} \tag{8}$$

Step 6: The procedures returns to (2) until all weights are assigned to the assets, that is, $\mid L_i \mid = 1, \forall L_i \in L$.

Problems of instability in portfolio performance of quadratic optimizers are present in recent studies by Chen & Yuan (2016), Prado (2016, 2018) and Bnouachir & Mkhadri (2019) that suggest a methodology that produces less risky out-of-sample portfolios compared to the traditional risk of the parity method and MV, and more stable portfolios. In the numerical example of Prado

(2016), the performance of the out-of-sample portfolio is evaluated through Monte Carlo simulation, improving the Sharpe by about 31.3% (for a broader discussion of in-sample vs. out-of-sample performance see Bailey et al. (2014)). Hierarchical clustering generates robust, diversified, and better risk-adjusted out-of-sample performance portfolios compared to traditional portfolio optimization techniques (Raffinot, 2017), (Burggraf, 2021). Although its features are attractive, the empirical literature on out-of-sample portfolio performance compared to the HRP approach is still very scarce (Burggraf, 2021).

3. Dimensionality Reduction and Clustering Methods

We briefly review the machine learning techniques selected for this study. We have chosen three algorithms that represent important classes of grouping. Each algorithm employs a different criterion, according to the type of feature they represent: Agglomerative Hierarchical Clustering, which has a concept of chaining the data, K-means and PAM fall into the category of compactness with different approaches; the first is an iterative algorithm that minimizes the sum of the distances of each pattern to the centroid of each cluster, overall groups; while the second seeks to reduce the sum of dissimilarities, which guarantees the compactness property of the objects, is less sensitive to outliers and noisy data. In addition to UMAP, a graph learning algorithm for dimensional data reduction.

Authors McInnes et al. (2018) developed UMAP for dimension reduction, classified as a k-neighbor-based graph learning algorithm for unsupervised learning problems. It is scalable, practical, computationally unconstrained in embedding dimension, and applied to real-world data. The mathematical construct comprises multiple learning techniques, topological data analysis, and fuzzy logic; see (McInnes et al., 2020, 2018) for details on the mathematical foundations.

The algorithm has five main hyper-parameters: 1) n, the number of neighbors; 2) d, embedding dimension; 3) $min - dist$, a minimum distance between

close points in the embedding space; 4) $n-epochs$, the number of training epochs to use when optimizing the low dimensional representation; and 5) *metric*, the distance metric to be used to calculate the distances. These, in turn, determine whether the application will produce a local or global low-dimensional data structure. By varying its parameters, the model can reduce a high dimensional dataset to a low dimensional dataset preserving both local and global structures without compromising the topological structure of the original dataset while being computationally fast and robust compared to methods in the same class (McInnes et al., 2018), (Becht et al., 2019).

The operation of UMAP boils down to two major steps. The first is learning the manifold structure by finding the nearest neighbors, followed by constructing the neighbor graph, and the second is finding a low-dimensional representation of the manifold structure using the minimum distance and minimizing the cost function.

Let be the input dataset $X = \{x1, ..., xN\}$, with a distance metric, $d(v_{i1}, v_{jk}) \rightarrow \Re \geq 0$, between pairs of the objects v_i and v_j, such that $i, j = 1, ..., N$, and the number of neighbors to be computed denoted by k, which will be computed the K nearest neighbors of v_i under the metric d. The parameter ρ_i and α_i are computed, for each datapoint v_i. The parameter ρ_i represents a non-zero distance from v_i to its nearest neighbor and is given by Eq. 9.

$$\rho = min\left\{d(v_i, v_{ij}) \mid d(v_i, v_{ij}) > 0\right\}, \tag{9}$$

and set α_i as the value such that the constant must satisfy the condition determined using binary search by Eq.10.

$$\sum_{j=1}^{k} exp\left(\frac{-max(0, d(v_i, v_{ij}) - \rho_i)}{\alpha_i}\right) = \log_2(k) \tag{10}$$

The selection of the ρ_i parameter derives from the local connectivity constraint of the manifold, i.e., each v_i connects to at least one other point in the dataset with an edge of weight 1. In practical terms, this is significant as it improves representation on high-dimensional data where different algorithms of the same class suffer from the curse of dimensionality (McInnes et al., 2020).

The algorithm constructs a joint probability distribution p_{ij} to measure similarity between v_i and v_j, such that similar (dissimilar) objects are assigned a (lower) probability

$$p_{ij} = p_{j|i} + p_{i|j} - p_{j|i}p_{i|j}, \tag{11}$$

$$p_{j|i} = \left\{ \begin{array}{ll} exp\left[\frac{-max(0,d(v_i,v_{ij})-\rho_i}{\alpha_i}\right], & j \neq 1 \\ 0, & j = i \end{array} \right\}, \tag{12}$$

where $p_{ij} = p_{ji}, p_{ii} = 0, \sum_i {}_j p_{ij} = 1 \; and \; \sum_j p_{j|i} = 1, \forall i, j.$

In the second stage, the UMAP computes the similarities between each pair of points in the space Q

$$q_{ij} = q_{j|i} + q_{i|j} - q_{j|i}q_{i|j}, \tag{13}$$

$$q_{i|j} = \left\{ \begin{array}{ll} exp\left[1 + a \mid\mid t_i - t_j \mid\mid^{2b}\right]^{-1}, & j \neq 1 \\ 0, & j = i \end{array} \right\}, \tag{14}$$

where $q_{ij} = q_{ji}, q_{ii} = 0, \sum_i {}_j q_{ij} = 1 \; and \; \sum_j q_{j|i} = 1, \forall i, j.$

The algorithm determines the constants $a, b \in \Re$ given the desired separation between close points, $\delta \in \; \Re^+$, in the embedding space Q

$$\left[1 + a \mid\mid t_i - t_j \mid\mid^{2b}\right]^{-1} \approx \left\{ \begin{array}{ll} 1, & t_i - t_j \leq \delta \\ exp\left[-(t_i - t_j) - \delta\right], & t_i - t_j > \delta \end{array} \right. . \tag{15}$$

UMAP then optimizes the layout of the data representation in the low dimensional space while attempting to minimize the cross-entropy C between the distribution of points in P and Q

$$C = \sum_{i \neq j} \left[p_{ij} \ln \frac{p_{ij}}{q_{ij}} - (1 - p_{ij}) \ln \frac{1 - p_{ij}}{1 - q_{ij}} \right]. \tag{16}$$

Then the algorithm optimizes the layout of the data in a low-dimensional space by minimizing the error between the two topological representations.

UMAP uses the Laplacian Graph and assigns initial low-dimensional coordinates, then proceeds with optimization using gradient descent

$$\frac{\partial C}{\partial t_i} = \sum_j \left[\frac{2ab\left[d(t_i,t_j)\right]^{2\,(b-1)}}{1+a\left[d(t_i,t_j)\right]^{2b}} p_{ij} - \frac{2b}{\left[d(t_i,t_j)\right]^2 \left(1+a\left[d(t_i,t_j)\right]^{2b}\right)} (1-p_{ij}) \right] (t_i - t_j). \tag{17}$$

James McQueen first proposed k-means in 1967 (MacQueen, 1967) and is one of the most widely used partitional clustering models, owing its popularity to its relative simplicity and ability to scale large datasets. The mathematical basis of K-means works by, as a first step, choosing k random data points to be the initial centroids. Then, the data is moved from one cluster to another to improve the value of the clustering criterion and calculate the minimum distance between the data points and the k centroids over all groups to associate each data point with the nearest centroid iteratively. The squared error E for a cluster containing k clusters is the sum of the variation within the clusters, where i = sample i, j = centroid j, illustrated by Equation 18.

$$E = min(d(x_i, x_j) \tag{18}$$

Then generate new centroids with their new cluster composition using:

$$C_j = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{19}$$

Where C_j is the centroid of cluster j for a data set with N samples, and it is necessary to repeat these steps until they find a convergence or adopt a stopping criterion.

The Euclidean distance is the most common metric adopted to calculate the distance between two points (Lorena et al., 2021). Given by equation 20 where x_i and x_j are two objects represented by vectors in $\Re^d$ space, and x_i^l and x_j^l are elements of these vectors, which correspond to the values of the l coordinate (attributes). Shown below:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{d} (p_i^l - q_i^l)^2} \tag{20}$$

Another clustering algorithm we used was PAM, proposed in 1987 by Kaufman and Rousseeuw (Kaufman & Rousseeuw, 1987). The PAM method is more robust because it minimizes the sum of the dissimilarities and does not rely on an initial assumption for the centers of the clusters, as with K-means. In addition, it is less sensitive to outliers and noisy data, as alluded to earlier. The algorithm has two phases (Kaufman & Rousseeuw, 1990):

- **Buil**: works by choosing k random data points as medoids, the first object corresponds to the pattern where the sum of the dissimilarities between all patterns is minimal. Subsequent objects are selected so as to minimize the objective function as much as possible. The function is given by:

  $$\sum_i = 1^N d(i, m(i), \tag{21}$$

  Where N is the total data, i is the object in the data set, $m(i)$ is the closest medoid to object i and $d(i, m(i)$ is the dissimilarity between i and $m(i)$.

- **Swap**: this phase seeks the improvement of the medoid set by swapping objects between them by calculating the new value of the objective function; then, it is necessary to check if the cost of swapping the randomly selected non-medoid data item is less than zero, and then calculate the new set of k-medoids. And is necessary to repeat these last two steps until the values of the medoids do not change or adopt a stopping criterion (Bhat, 2014).

Finally, the third algorithm we use is Agglomerative Hierarchical Clustering (AHC), which chains the data together and aims to identify a clustering hierarchy in the dataset. The agglomerative (bottom-up) approach starts with n objects in k clusters where level 1 presents n clusters of an object, and level n

presents a cluster with all objects that form the sequence of partitions grouping the clusters successively. Thus, when clustering two objects at some level, they remain part of the same group at higher levels, building a hierarchy of clusters (K.Sasirekha, 2018), (Duda et al., 2000). This strategy facilitates exploration of the data at different levels of granularity and easy use of any form of similarity or distance, and it further allows the use of any attribute.The main steps of the agglomerative algorithm are:

Step 1: Create the similarity matrix between the clusters $S_{n \cdot n}$.

Step 2: Calculate the dissimilarity $D(C_k, C_l)$ between all available k, l cluster pairs using a linkage method.

Step 3: Calculate the dissimilarity $D(C_k, C_l)$ between all available k, l cluster pairs using a linkage method.

Step 4 Combine the cluster pair C_k and C_l, generating a single cluster C_{kl}.

Step 5: If there is more than one cluster remaining, return to step 2. If not, end.

In addition to the calculated distance between the elements, it is necessary to define a linking method to translate the distance between the clusters described in step 2 of the algorithm. We use Ward's method (Ward Jr., 1963) from this paper's various types of linking strategies. The chosen method minimizes the total variation within the cluster after merging two clusters into one. At each step, ponder the union of all possible cluster pairs, and the two clusters whose fusion results in a minimum increase in information loss are combined (Nielsen, 2016). Ward defines the minimum loss in terms of the sum of squares of errors (ESS), given by the equation 22. It is also the most suitable method for quantitative variables and the only clustering strategy in the research of Tayali (2020) that returned a gain for the investor.

$$Ward\ linkage : d(k,l) = \sqrt{\frac{2n_k n_l}{n_k + n_l}}\ ||\ \overline{b} - \overline{a}\ ||_2 \tag{22}$$

As one of the main parameters of clustering techniques is the cluster number K and for its specification, we adopted two measures of internal validation, the Silhouette Coefficient and the Davies-Bouldin Index, justified in subsection 5.1.

The Silhouette Coefficient (Rousseeuw, 1987) aims to evaluate a partition present in the structure of the data set. The silhouette measure is based on comparing the disparity within each cluster and between clusters obtained by implementing a clustering algorithm. For a sample i, $s(i)$ is given by:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{23}$$

Where:

$a(i)$: average dissimilarity of i to all other objects of A,

$d(i, C)$: average dissimilarity of i to all objects of C,

$b(i)$: minimum $d(i, C)$

And as output will be given an average $-1 \leq s(i) \leq 1$, where $s(i) = 1$ indicate a sample perfectly assigned to its cluster, and an $s(i) = 1$ indicates a sample incorrectly alocated.

The Davies-Boudin index (Davies & Bouldin, 1979) is a metric to evaluate the partitions of the clustering models that, unlike other methods, can be supervised or unsupervised. This index calculates the relationship between intragroup dispersion and intergroup similarity, and the selection of the optimal number of clusters will be the one with values closest to zero for well-partitioned clusters. The Davies-Bouldin Index $\widetilde{R}$ for a data set with N samples is given by:

$$\widetilde{R} = \frac{1}{N} \sum_{i=1}^{N} R_i \tag{24}$$

Where:

$R_i \equiv R_{ij}$, $i \neq j$

R_{ij} : $\frac{S_i + S_j}{M_{ij}}$

M_{ij}: distance between centroids i and j

S_i: Dispersion of cluster i

S_j: Dispersion of cluster j

4. Methodology

4.1. Data description

The implementation of the algorithm framework used Python 3.7.13 language and Google Colaboratory cloud storage service. The full code of the solution, for reproducibility purposes, has been made publicly available at: **https://github.com/ComputerFinance/ESWA**.

Economatica provided the data for this research, the largest firm of financial information on the Brazilian stock market, Latin America, and the United States of America (USA). The data consist of twelve fundamentalist and eight technical indicators calculated from the information in the companies' financial statements and daily closing asset prices of companies listed in two different indices. The Ibovespa, the most important indicator of the average performance of Brazilian stock market prices traded on the B3 (an acronym for Brazil, "Bolsa," "Balcão"), formed by a hypothetical portfolio of the stocks with the highest trading volume in recent months, at the beginning of 2022, was composed of 93 stocks from 90 companies (B3, 2022). The Standard and Poor's 500 (S&P 500) is widely considered the best indicator of USA large-cap stocks that gathers the 500 leading companies and represents approximately 80 percent of the available market capitalization (S&PGlobal, 2022). The time chosen was from January 1st, 2016, to December 31st, 2021 (i.e., six years of data resulting in 1504 scenarios of daily changes in the indices). In evaluating the performance of the optimized portfolios, we use the closing asset price data from the Yahoo Finance API for the same period as the indicators dataset.

4.2. Input variable selection

The variables in this study comprised twelve fundamentalist and eight technical indicators calculated from the information in the companies' financial statements and daily closing prices. Table A.5 and A.6 in Appendix A shows the main characteristics and the formulas available on Economatica's platform (Economatica, 2022).

The objective of these indicators in the proposed framework is to optimize the objective function 31 and score the stocks in each cluster. This approach broadly compares the economic and financial situation and capital structure to find stable companies with good growth potential for higher returns. Once the assets have been ranked, only those designated with the highest objective function value are eligible to participate in the next step.

Even before the indicators are used as input attributes for the first method of modeling called UMAP, we perform a pre-processing on the dataset in order to minimize the influence of possible problems, such as: noise and missing values, among others. Each step is described below:

- **Cleaning**

  The criterion adopted was to use each index's most recent stock portfolio composition, whose historical values were complete from January 2016 to December 31, 2021—resulting in 74 stocks from Ibovespa and 464 from the S&P500, totaling 538 stocks.

- **Integration**

  With the daily closing prices values, we calculate the logarithmic return in which each asset (i) has a closing value on day t given by V_t^i. Thus, we denote the return between day $t-1$ and day t by $R_t^i = \ln(V_t^i/V_{t-1}^i)$. Each asset has a vector of daily returns $R^i = [R_1^i R_2^i ... R_N^i]^T$, where N is the total number of days analyzed in a given period. We denote the average return of the asset (i) by $m_i = (R_1^i + R_2^i + + R_N^i)/N$, that is, the average vector of all daily returns.

- **Transformation**

  We performed the normalization by the amplitude of the variables because the limits of values of distinct attributes are very different, thus avoiding that one attribute predominates over the other in the dimensional reduction process. We adopted the normalization by re-scaling, also called min-max, so the upper and lower limits are 1 and 0, respectively, illustrated such that: $\widetilde{X} = \frac{x - x_{min}}{x_{max} - x_{min}}$, where x is the current value of the

attribute, x_{min} the smallest value of the attribute subtracted from each value. Then each resulting value is divided by the difference between the smallest, X_{min}, and the largest, X_{max} original attribute values, and $\widetilde{X}$ is the standardized vector.

4.3. Performance measures

The measures for portfolio performance evaluation presented below are among the most popular in portfolio management for long-term horizons (Almahdi & Yang, 2017), (Hochreiter, 2007),(Wang et al., 2020), (Metaxiotis & Liagkouras, 2012), (Kalayci et al., 2019), (Silva et al., 2015), (Georgantas et al., 2021). We compare the three optimization methods (MV, IVP, and HRP) with these measures to validate the portfolio's financial performance using in-sample and out-of-sample data. To evaluate portfolio diversification, we use the Herfindahl-Hirschman index (HHI), which is among the most suitable indices in the economic context and provides a more consistent measurement of portfolios with small exposures (Chen et al., 2013), (Dincer, 2018).

- **Herfindahl-Hirschman index**

  The indicator measures portfolio diversification. The calculation consists of squaring each company's market share, that is, the allocation of each asset in the portfolio. Where w_i is the allocation rate of the company i expressed as an integer, not a decimal, and N is the number of investments in the portfolio. When it shows low values, it indicates more diversified portfolios, and the higher the ratio, the higher the concentration level (see also the discussions in Slime (2016)). The HHI ranges from 10,000 to 0, and the scales adopted in this work are defined according to the US Department of Justice and the Federal Trade Commission Justice & Commission (2010) and assume three variations:

  (i) **Non-concentrated markets**: HHI below 1500;

  (ii) **Moderately concentrated markets**: HHI between 1500 and 2500;

  (iii) **Highly concentrated markets**: HHI above 2500.

It is obtained by Eq. 25:

$$HHI = \sum_{i=1}^{N} w_i^2 \tag{25}$$

- **Cumulative Portfolio Return**

  A portfolio return refers to how much an investment portfolio gains or loses over a given period. Portfolios aim to deliver returns based on the investment strategy's stated objectives and the risk tolerance of the type of investor targeted by the portfolio. Denote the portfolio return as Eq. 26:

$$E = \sum_{i=1}^{N} X_i.\mu_i \tag{26}$$

  Where:

  E: average portfolio return.

  N: is the number of assets.

  μ_i: is the average return of asset (i).

  X_i: fraction of asset (i) in the portfolio.

- **Annualized Volatility**

  All human endeavors involve uncertainty and risk, present in any sector of economic activity; what makes them different comes from their nature, origin, potentialities, social, economic, environmental, and political effects according to each kind of activity (Olson & Wu, 2020), (Hampton, 2015), (Gularte, 2021). Volatility in modern finance theory is associated with the risk of fluctuations in prices and rates charged by the market. It measures the dispersion of portfolio returns and indicates greater risk as the deviations around the mean increase; it generally refers to the amount of uncertainty related to the size of changes in the investment portfolio's value (see also the discussions in Holton (1992)). The n-day volatility calculation uses a series of $n + 1$ day closing prices $\{d0, d1, d2, d3, ..., dn\}$ in Eq. 27:

$$V = \sqrt{\frac{\sum (Si - Sm)^2}{n * PPA}} \tag{27}$$

Where:

Si = neperian logarithm of $(di/di-1)$ $i = 1...n$

Sm = mean of $S1, S2, S3, ..., Sn$

PPA: Periods per year and is worth 1 for annual closures.

- **Sharpe ratio**

  The Sharpe ratio is widely used and was developed based on mean-variance optimization; therefore, its measure of risk is the standard deviation of returns. By using a risk-adjusted measure, the objective is to maximize the Sharpe ratio in Eq. 28:

$$SR = \frac{E(R_t)}{\sigma} \tag{28}$$

  Where:

  $E(R_t)$: is the mean of the returns.

  σ: is the standard deviation of returns.

- **Maximum drawdown**

  A maximum drawdown (MDD) is a risk metric that measures the maximum cumulative loss from a peak to the following bottom in the portfolio. It is an indicator of the risk of loss over a historical period. It is calculated as follows in Eq. 29):

$$MDD_T = \frac{trough\,value - peak\,value}{peak\,value} \tag{29}$$

  Where:

  T: the chosen period within which the MDD is to be measured.

  $peak\,value$: Points at which the asset value reaches a level never went since the beginning of the period.

  $trough\,value$: is the point (after the peak) where the asset value reaches its minimum level. Each peak has, therefore, its respective trough value; i.e., the trough value of the Peak (X) is the minimum point comprised in the period that starts at the Peak (X) and ends when a new Peak is reached (the new peak is the point where the asset value exceeds the value of Peak (X).

- **Beta Index**

  Beta calculates the stock fluctuations and the benchmark in each period and determines the portfolio's volatility relative to the market portfolio, which is around 1.0. A portfolio ranking is formed by how much it deviates from this value; portfolios with a beta above 1 are more exposed to market risk, and those below 1 are less exposed. To determine in in Eq. 30:

$$Beta = \frac{\sigma 2(stock\ fluctuations, benchmark\ fluctuations)}{DP(benchmark\ fluctuations)} \tag{30}$$

  where:

  $\sigma 2$: is covariance

  DP: standard deviation

4.4. Proposed model

To perform dimensionality reduction and asset pre-selection, we integrate the UMAP, K-means, PAM, and Agglomerative Hierarchical Clustering methods to maximize the objective function formed of fundamental and technical data, finally compose the input subset of the optimization problem. We calculated the MV portfolio, managed by the classical Markowitz model measures: mean and covariance matrices compared to the IVP and HRP optimization models and the performance analysis of the in-sample and out-of-sample portfolios. It is worth clarifying that we do not use simulations of investment decision-making and risk-free assets. Figure 1 presents the model framework.

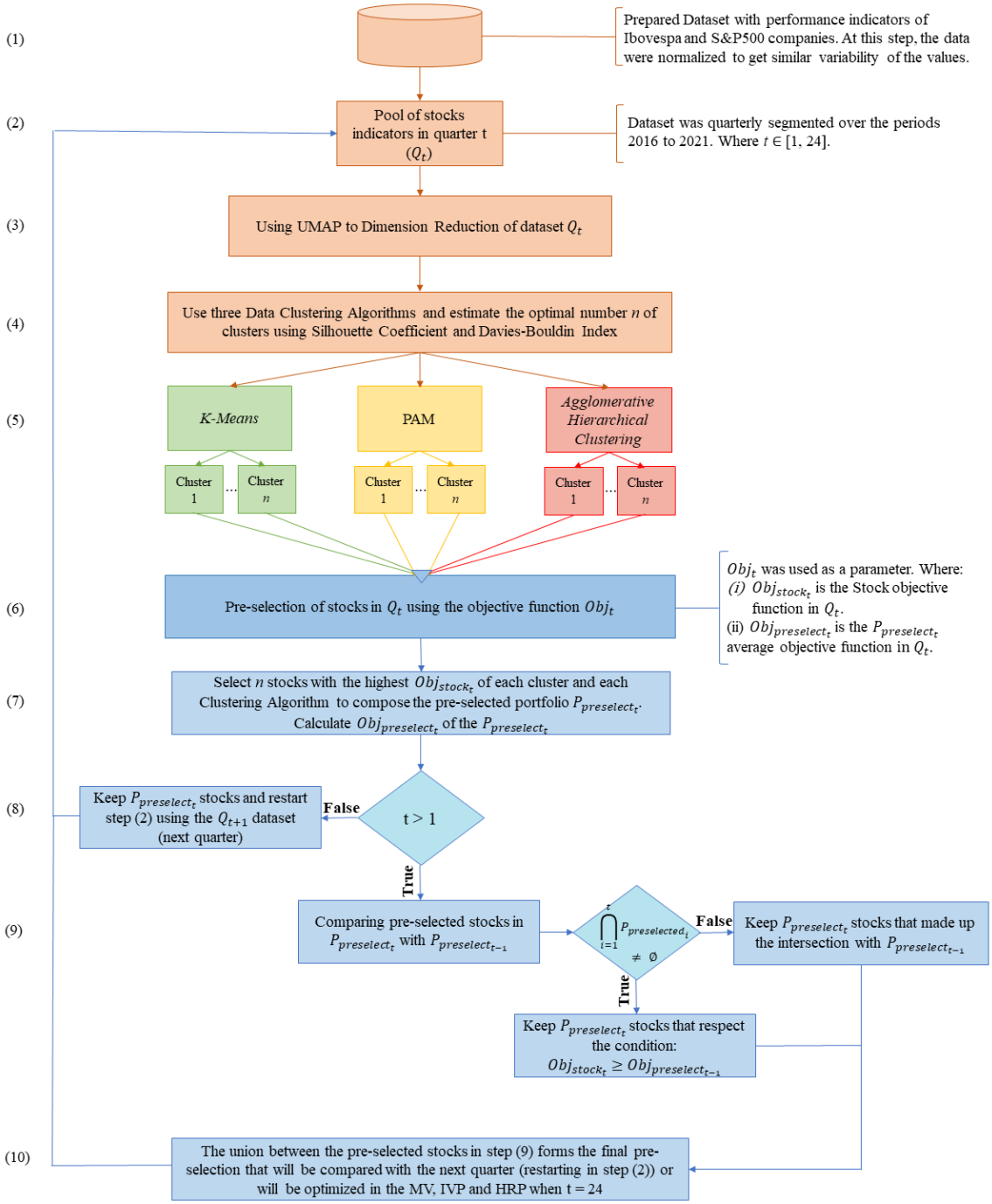

Figure 1: Flowchart for Pre-selection stocks using the clustering approach. Source: Own authorship, 2022.

The first and second steps collected the time series data of 538 stocks that composed the Ibovespa and S&P500 indexes from the digital platform of Economatica. We selected 20 fundamentalist and technical indicators over the analysis period from January 1st, 2016, to December 31st, 2021, comprising 1.504 trading days. We subdivided the dataset into quarters designated by Q_t, where $t \in [1, 24]$, which, in addition to the indicators of each stock in quarter t, presented the quarterly average stock price as an additional feature. In the following procedures, they were processed one by one along with the pre-selection steps. Our implementation considered the Python programming language for preprocessing and normalization of the data and machine learning algorithms.

For the reduction of dimensionality applied in the Q_t sets and demonstrated in (3), was use the UMAP (Uniform Manifold Approximation and Projection) model, which in this study has an as main objective to perform a transformation in the m-dimensional database of input data from cluster models to an n-dimensional database, with n¡m, to optimize the distance calculations between the vectors performed by clusters. We use the Silhouette Coefficient and Davies-Bouldin Index to obtain the optimal number of groups formed by the model in (4). This step is detailed in section 5, in particular Algorithm 1.

In this way, in (5), the clustering models K-means, Partition Around Medoids (PAM) and Agglomerative Hierarchical Clustering were applied to generate the clusters from the Q_t set so that, soon after, in (6) pre-selection of the stocks of each cluster and each model according to the Eq. 31 below:

$$Obj_t = max(\frac{CR+QR+NPM+ROA+ROE+IRR+SR}{DTEB+EVTE+DTEQ+PTE+PTB+PFCF+V+VaR+MDD+B}) \quad (31)$$

Where:

$CR = Current\,ratio$

$QR = Quick\,ratio$

$NPM = Net\,Profit\,Margin$

$ROA = Return\,on\,Assets$

$ROE = Return\,on\,Equity$

$IRR = Internal\,Rate\,of\,Return$

$SR = Sharpe\,ratio$

$DTEB = Debt - to - EBITDA$

$EVTE = Enterprise\,Value - to - EBITDA$

$DTEQ = Debt - to - Equity$

$PTE = Price - to - Earning$

$PTB = Price - to - Book$

$PFCF = Price - to - Free\,Cash\,Flow$

$V = Volatility$

$VaR = Value\,at\,Risk$

$MDD = Maximum\,Drawdown$

$B = Beta$

Through the objective function Ob_t two other pre-selection parameters were calculated: (i) Obj_{stock_t} is the Stock objective function in Q_t and (ii) $Obj_{preselect_t}$ is the $P_{preselect_t}$ average objective function in Q_t, where $P_{preselect_t}$, refers to the pre-selected portfolio in step (7). Because it is an iterative process of selection of assets that occurs quarterly, in step (8) when $t = 1$, step (2) was returned with the set referring to Q_{t+1} so that in (9) with the two pre-selections consecutive stocks $P_{preselect_{t-1}}$ and $P_{preselect_t}$ it was possible to find among them the intersection stocks existing in both, which would be kept on $P_{preselect_t}$. The stocks that did not composed the intersection should respect the Obj condition $Obj_{stock_t} \geq Obj_{preselect_{t-1}}$, seeking to have pre-selected stocks that would progressively improve the average objective function of the portfolio over the quarters. Finally, in step (10) we have that the resulting $P_{preselect_t}$ is the union between these two previous conditions, starting again at (2) until $t = 24$ when we have the $P_{preselect_t}$ to $MeanVariance$ (MV), $Inverse - VariancePortfolio$ (IVP) and $HierarchicalRiskParity$ (HRP).

5. Experiments and results

5.1. Results analysis in the first stage: asset pre-selection

The types of experiments performed are divided into three big steps, the first being the testing and validation of the UMAP dimensionality reduction model, then the implementation of cluster models with the development of an iterative method for selecting the number of clusters parameterized in the models, and finally, the parameterization and evaluation of the results obtained from the pool of stocks generation model.

Then, the first step is to apply the UMAP dimensionality reduction model. UMAP is a dimensionality reduction model that can, by varying its main parameters, reduce a high-dimensional database to a low-dimensional (generally two or three dimensions) dataset focusing on a global or local aspect of the data. We used dimensionality reduction models as a pre-processing step for clustering models application to improve the distance relationships of the data points that will be calculated in the models and to upgrade the computational performance of the models. Furthermore, the clustering models present different distributions of elements between clusters depending on the application or not of a dimensionality reduction model as a data pre-processing step.

In this application was tested the K-Means clustering model with and without the application of UMAP in pre-processing step. For both cases, the selection of the number of clusters used clustering evaluation techniques (the same method used in Algorithm 1, that following we describe each step in more detail), and the method indicated some clusters for the dataset without dimensionality reduction of 2, and 5 for the dataset with the application of UMAP in pre-processing step. And the experiment presented the following results in Table 1.

Table 1: UMAP in the data preprocessing step (applied in the K-means method).

Cluster Id	**Distribution of stocks per cluster**
K-Means results with UMAP in preprocessing step	
0	22,3%
1	26,7%
2	21,9%
3	14,1%
4	15,0%
K-Means results without UMAP in preprocessing step	
0	98,9%
1	1,1%

The experiment in table 1 illustrates a more even distribution of assets among clusters and better identifying groups. Furthermore, because of UMAP's non-linear approach to capturing the underlying geometric structure of the data faithfully, the overall topology of the dataset is more accurately preserved even at low dimensions, both local and global structures.

UMAP has five main parameters: the number of neighbors, the minimum distance, the target number of dimensions, the number of epochs and the distance metric. We use the standard parameters for the application, with the number of neighbors equal to 15, a minimum distance equal to 0.1, a target number of dimensions equal to 2, a number of epochs equal to 500 (in the standard parameters, the number of epochs is equal to 500 for small datasets - with less than or equal to 10000 samples -, and equal to 200 for large datasets - greather than 10000 samples), and the distance metric used is the Euclidean metric. Figure 2 has presented the output of the application of UMAP for the last quarter of 2021.

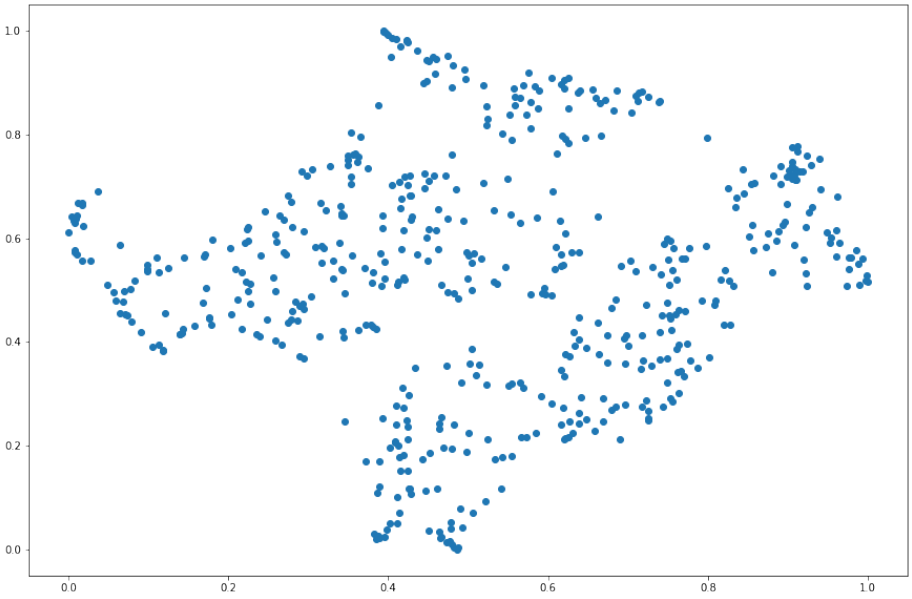

Figure 2: UMAP output. Source: Own authorship, 2022.

After the data set dimensional reduction phase, the K-means, PAM, and Hierarchical Agglomerative Clustering algorithms had the output of UMAP as input data for processing. We find several indexes in the literature for specifying and evaluating the most appropriate number of clusters present in the data structure. Although there is no consensus on which measures to use, Jain & Dubes (1988) summarize the main aspects related to the use of an evaluation index, and among them stand out the internal validation measures that have in their structure the degree to which an obtained partition is justified, both for compactness and separability of the data. Therefore, we adopted the Silhouette Coefficient and the Davies-Bouldin Index, detailed in section 3.

We apply the method at each quarter in the dataset iteratively to select the optimal number of clusters defined by K in each of the algorithms, K-means, PAM, and Hierarchical Agglomerative Clustering. We calculate the optimal K in each continuous window using the internal validation measures of clustering

as explained earlier and detailed in section 3. The algorithm will obtain the cluster number that maximizes the Silhouette Coefficient and minimizes the Davies-Bouldin Index by generating a score. Then we average between these values to determine the optimal cluster number once the clustering method is applied with a value of $K = \{2, ..., 20\}$.

Figure 3 showed the result of the application of the models for the last quarter of 2021 and the number of clusters equal to 5:

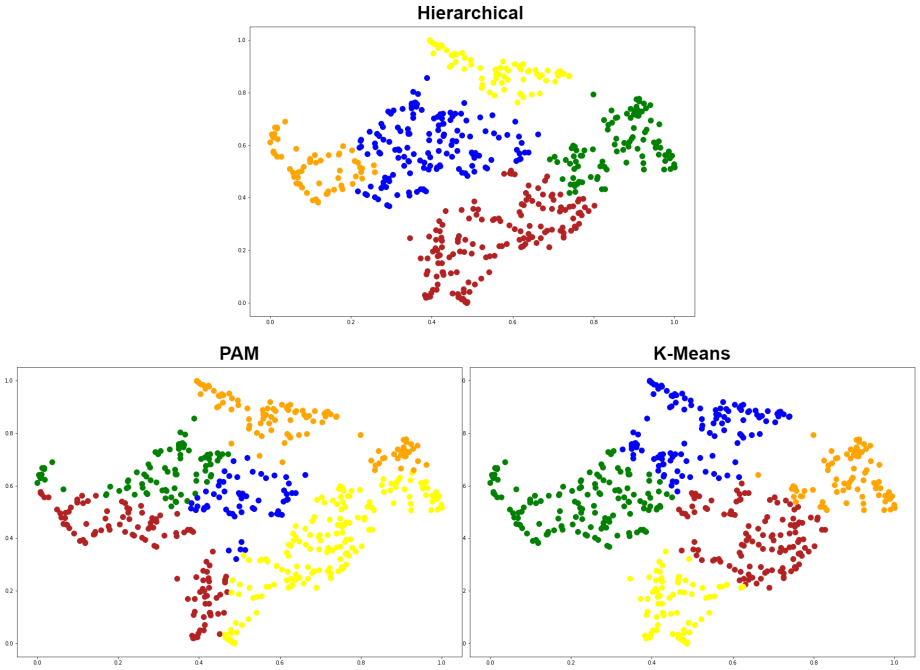

Figure 3: Clustering models outputs. Source: Own authorship, 2022.

And as the last step, the number n of stocks was selected as mentioned in steps 6 to 9 in the Figure 1 in section 4.4. The goal was to find a subset based on the target number of the entire universe of assets, reducing its size and soon becoming the input data for the MV, IVP, and HRP problems. After splitting the financial time series dataset into a cluster structure equal to 5, selecting the n number of 2 representative asset from each cluster for the three clustering

methods will be the one with the largest values of the objective function 31.

Below is presented the method built to generate the pool of pre-selected stocks (steps 3 to 6 of Figure 1, in the 4.4 section):

Algorithm 1 Generate the pool of stocks $P_{preselect_i}$

1: $QuartersList \Leftarrow$ all quarters between 1Q2016 and 4Q2021
2: $StocksTable \Leftarrow$ table with the stocks and their indicators
3: **for** $quarter_i < QuartersList$ **do**
4: $\quad StocksApplication \Leftarrow StocksTable$ with $quarter_i$ filtered
5: $\quad$ Apply UMAP in $StocksApplication$
6: $\quad$ **for** $ClusterNum$ in range(0, 20): **do**
7: $\quad\quad$ Apply K-Means using $ClusterNum$ as number of clusters
8: $\quad\quad$ Calculate Silhouette Coefficient and Davies-Bouldin Index
9: $\quad$ **end for**
10: $\quad a \Leftarrow cluster_num$ with the maximum Silhouette Coefficient
11: $\quad b \Leftarrow cluster_num$ with the minimum Davies-Bouldin Index
12: $\quad ClusterNumParam \Leftarrow \frac{(a+b)}{2}$
13: $\quad$ Apply K-Means with $ClusterNumParam$ as number of clusters
14: $\quad$ Apply PAM with $ClusterNumParam$ as number of clusters
15: $\quad$ Apply Hierarchical with $ClusterNumParam$ as number of clusters
16: $\quad StocksNum \Leftarrow 2$
17: $\quad P_{preselect_i \Leftarrow StocksNum}$ of each cluster of each clustering model
18: **end for**

As a result, the method selected a set of 7 stocks from the S&P500 index and 1 from IBOV index with an average Sharpe Ratio of 5.82 and Ob_t of 0.28, out of a universe of 538 stocks with an average Sharpe Ratio of 1.73 and Ob_t of 0.07, the selected stocks are shared among five industry classes (according to the classification of business establishments by type of economic activity - North American Industry Classification System (NAICS).

Researchers agree that keeping a high number of different assets in the port-

folio is not realistic for individual investors. Many focus on ten or fewer assets, as evidenced in the works of Almahdi & Yang (2017) and Ruiz-Torrubiano & Suarez (2010) with five assets and Wang et al. (2020) with ten assets, given that a very high number of assets are difficult to manage and can incur high transaction costs. In recent research Paiva et al. (2019) proposed a model to form the optimal portfolio in which the classifier showed higher discriminatory power, converging positively to a lower cardinality, with a daily average of seven assets in the portfolio. These researches reinforce the results obtained in our experiments that reached a final basket of 8 stocks, thus significantly reducing the problem's computational complexity. In Fig. 4, we show the average quarterly cardinality of the portfolios. The high dispersion of results is due to the different performance of the stocks in each quarter and, consequently, different outputs in the UMAP.

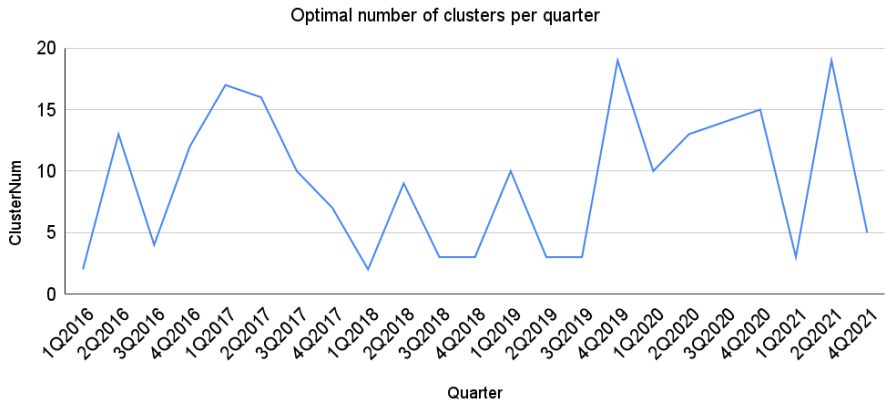

Figure 4: Optimal number of clusters per quarter.

The next section describes a detailed illustration of the assets that formed the portfolio for input to the optimization models and discusses the results of the section 4.3 in-sample and out-of-sample performance metrics.

5.2. Results analysis in the second stage: optimal portfolio formation

After the pre-selection of assets presented in subsection 5.1, an analysis of their daily log returns between January 2016 and December 2021 allowed the construction of a correlation matrix. The matrix presented in Figure 5 was reorganized into clusters using the Hierarchical Tree Clustering (Ward linkage agglomerative method) and Quasi-diagonalization. This process then serves as input to the real task of asset allocation, resulting in a reordering that groups similar assets placed together and dissimilar assets further apart, which helps the investor make more meaningful asset allocation decisions and build more diversified portfolios.Prado (2016, 2018). The positive correlation of the assets during the study period is in a range of [0.14;0.51]; another analysis from a dendrogram obtained through the correlations is that at $d \leq 0.89$ shows two clusters that group respectively three assets (BRAP4, PFE, and VRTX) and the other five (NLOK, NVR, ANET, NVDA, and VRSN).In addition, between $0.70 \leq d \leq 0.79$, we see the presence of two other more delimited clusters compared to the previous interval — in the composition of the first of them are the assets ANET, NVDA and VRSN, companies that, except for VRSN (Professional, Scientific and Technical Services), are from the same economic sector of Manufacturing Industry. The composition of the second of them is the assets PFE and VRTX, both from the Manufacturing Industry economic sector.

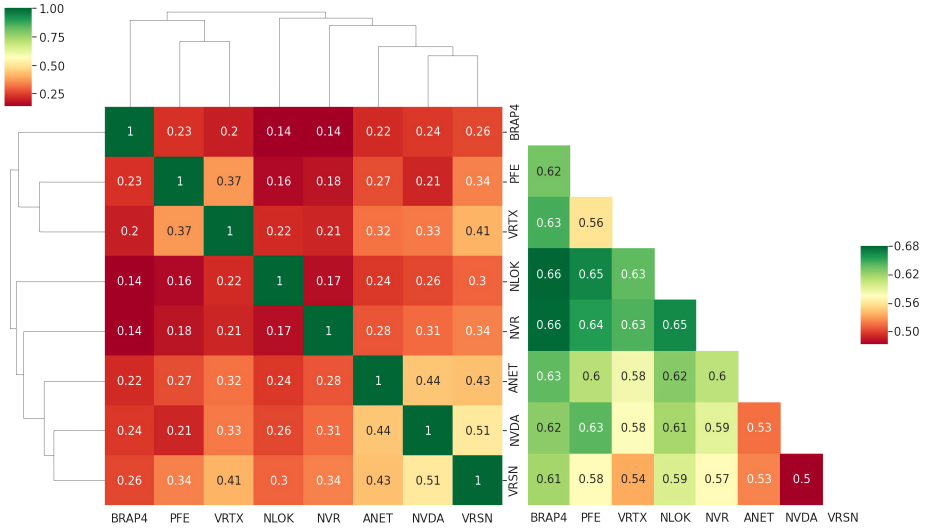

Figure 5: Correlation and Distance Matrices. Source: Own authorship, 2022.

Next, through an MST (Minimum Spanning Tree) generated by eliminating the links formed between nodes by correlations lower than 0.25, we observed that, reciprocally, its topology reinforced the aspect described in the distance matrix and dendrogram. That is, that the assets form two clusters with emphasis on the VRSN and NVDA companies that between them are the nodes with the closest proximity to one of the clusters and between the companies PFE and VRTX present in the second cluster. The MST of Figure 6, whose size of the nodes is proportional to the size of the annualized returns and the green color signifies a positive performance in the period, also allowed the calculation of the Degree Centrality (C_D). The asset with the highest C_D was VRSN (1.0), followed by ANET and NVDA (0.71). In parallel, BRAP4 and NLOK had the fewest links in the network and, therefore, the lowest C_{Ds} (0.14 and 0.28, respectively).

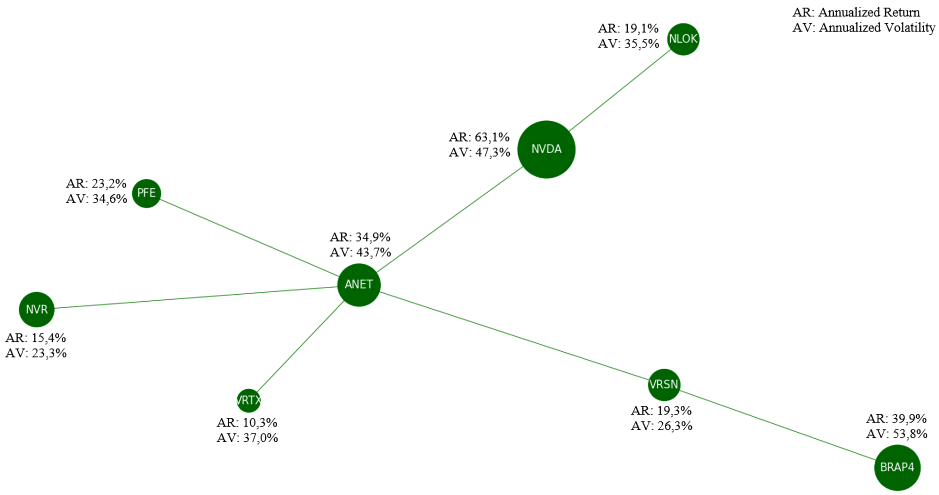

Figure 6: Minimum Spanning Tree (MST). Source: Own authorship, 2022.

The construction of optimized portfolios applying the classic Mean-Variance (MV) model, the Inverse-Variance Portfolio (IVP), and the Hierarchical Risk Parity (HRP) generated weighted portfolios presented in Table 2. Among the characteristics obtained from the three methods studied, we have that MV concentrated the portfolio weights in four assets (93.1%) and assigned weights between 0.6% and 2.9% for the other assets in the portfolio, facts also expressed by the HHI index (2828), demonstrating a high portfolio concentration. IVP and HRP were the methods that most evenly distributed the weights among the assets, showing a moderate concentration HHI of 1639 and 1710, respectively. The highlight of the HRP is that the weights assigned respected the structure of the set formed between the assets, i.e., assets that share close distances d among themselves had their weights distributed in a less concentrated way than the MV.

In an extension of the analysis of portfolio weights by methods, in Figure 7, we see that Manufacturing Industry was the sector with the largest number of assets in the portfolios, occupying the largest in the three optimizations, respectively MV (49.4%), IVP (50.9%), and HRP (53.7%). Thus, for the other

Table 2: Portfolio's Optimized

	MV	**IVP**	**HRP**		
Stock				**Economic Sector**	**Index**
PFE	44,8%	26,5%	30,0%	Manufacturing Industry	S&P500
VRSN	20,4%	20,8%	17,1%	Professional, Scientific and Technical Services	S&P500
NLOK	14,1%	11,4%	12,4%	Information	S&P500
VRTX	2,9%	10,4%	11,4%	Manufacturing Industry	S&P500
NVR	13,8%	12,0%	11,2%	Construction	S&P500
ANET	1,1%	7,5%	7,0%	Manufacturing Industry	S&P500
BRAP4	2,3%	5,0%	5,6%	Management of Companies and Enterprises	Ibovespa
NVDA	0,6%	6,4%	5,3%	Manufacturing Industry	S&P500
HHI	**2828**	**1639**	**1710**		

sectors that had three assets, MV was the method that concentrated the weights on them the most (80.5%), IVP concentrated the least (57.8%) and HRP distributed the weights of these sectors more equally in the portfolio, totaling 61.3% of the total allocation.

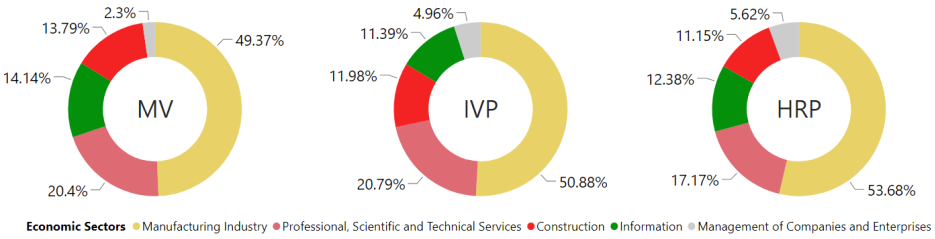

Figure 7: Portfolio weights MV, IVP and HRP by Economic Sector. Source: Own authorship, 2022.

Figure 8 presents the performance of the portfolios optimized by MV, IVP and HRP in relation to their benchmarks. Using a preliminary visualization of the portfolios' weights in the three methods, we performed a performance analysis of the portfolios formed between January 2016 and December 2021. This analysis helped us compare indicators such as Cumulative Returns, Annual Volatility, Sharpe Ratio, Max Drawdown, and the Beta of these portfolios

concerning to the reference indexes.

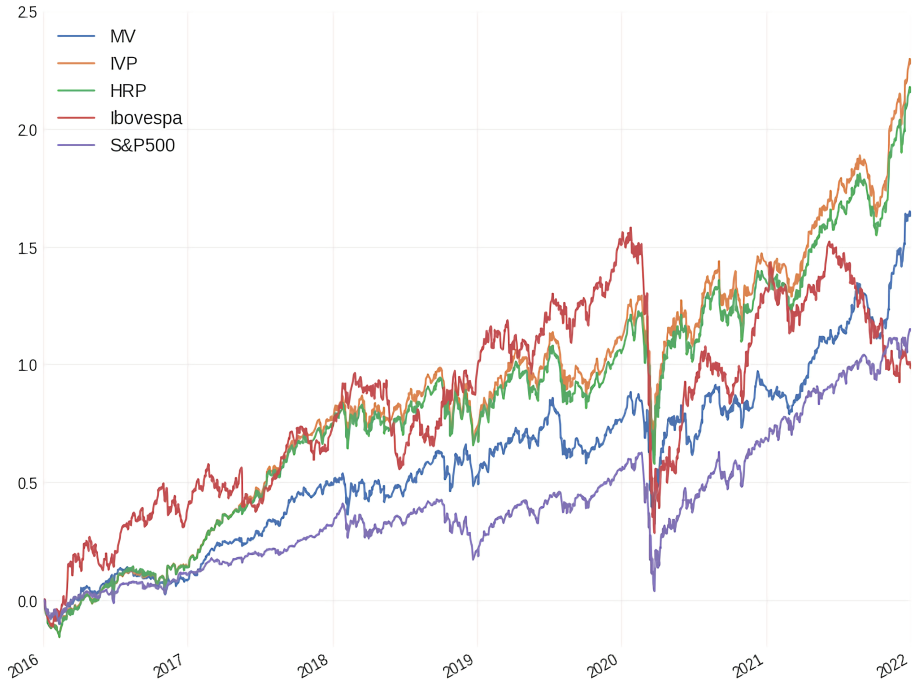

Figure 8: Portfolio Performance and Benchmarks. Source: Own authorship, 2022.

In Table 3, we also summarize these analyzed indicators detailed in subsection 4.3 .

Table 3: Portfolio Performances

	MV	**IVP**	**HRP**
Cumulative Returns	164,8%	228,8%	216,9%
Annual Volatility	18,7%	22,8%	22,1%
Sharpe Ratio	1,01	1,14	1,11
Max Drawdown	-29,9%	-29,6%	-29,1%
Beta S&P500	0,79	0,91	0,89
Beta Ibovespa	0,34	0,40	0,39

As noted in the performance indicators, the three portfolios outperformed

the S&P500 and Ibovespa in terms of accumulated return, once the S&P500 performed with gains of 114.61% and Ibovespa 100.03% in the period. In common, the portfolios presented annual volatilities in the 20% level, with IVP being the portfolio with the highest risk and MV the lowest. Thus, reflecting the accumulated returns and volatilities, the IVP portfolio had the best Sharpe Ratio (1.14), followed by HRP (1.11) and MV (1.01). In Figure 8, we also observe that higher daily losses in the period occurred on a similar day in the portfolios, closely related to the increase in restrictive measures associated with the Covid-19 pandemic on a global scale, thus affecting them in drawdowns close to 30%, which took 149 trading days to recover in the MV portfolio, 111 days for the IVP and 117 for the HRP.

In order to test the stability of the in-sample and out-sample optimizations, we performed Monte Carlo simulations in which we generated synthetic returns from our empirical covariance matrix using the multivariate normal distribution. We constructed 10000 simulated portfolios whose variances we distributed into histograms obtained based on 260 observations (equivalent to one-year frequency) both in-sample and out-sample. In Table 4, we present the average of the in-sample and out-sample variance distributions.

Table 4: Average Variances of the In-sample and Out-sample distributions

	σ^2_{MV}	σ^2_{IVP}	σ^2_{HRP}
In-sample	3,39%	3,99%	3,91%
Out-Sample	3,52%	4,04%	3,97%

Figure 9 shows the distributions generated by Monte Carlo simulations for the three methods and the two samplings.

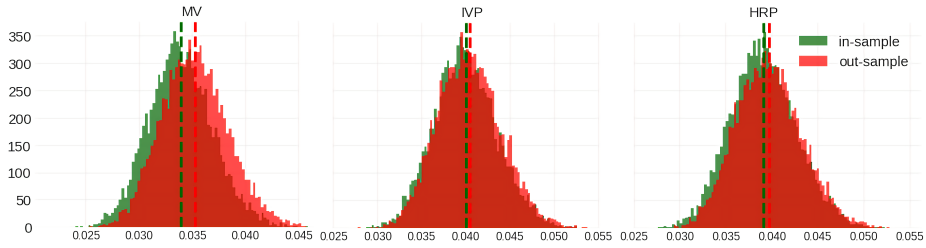

Figure 9: Monte Carlo Simulations of In-sample and Out-of-sample variances distributions. Source: Own authorship, 2022.

The simulations showed us that the in-sample and out-sample distributions of the HRP-optimized portfolios were more stable, especially when compared to the variance distributions obtained through MV. The good performance of the HRP occurs when we notice that its variance distributions are very similar, having an average in-sample of 3.91% and 3.97% out-sample. On the other hand, in MV, although the averages in-sample 3.39% and out-sample 3.52% variances were smaller than those of HRP, we saw in Figure 9 that the distributions were more shifted among themselves, which is also we justified as a result of the greater difference between these same averages. Therefore, this experiment reinforces the results obtained by Prado (2016, 2018) about the greater stability of HRP in the face of MV (represented by the Critical-Line Algorithm in the original experiment) and Classical Risk Parity.

6. Conclusions

Acknowledgments

The authors thank the anonymous associate editor and referees for their helpful comments. Thanks to the Voluntary Scientific and Technological Initiation Program that provided the generation of insights to improve this research. We thank Economatica, the largest financial information company on the Brazilian stock market, Latin America, and the USA, which provided access

to the data platform to support this research, as well as support and training in the correct interpretation and use of the information, saving the authors time in handling large volumes of data, allowing productivity gains and better and more comprehensive analyses.

Appendix A. Datasets selected for the experiment.

References

Almahdi, S., & Yang, S. Y. (2017). An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, *87*, 267–279. doi:10.1016/j.eswa.2017.06.023.

Asness, C., Frazzini, A., & Pedersen, L. H. (2012). Leverage aversion and risk parity. *CFA Institute. Reproduced and republished from Financial Analysts Journal with permission from CFA Institute.*, *68*, 47–59. doi:10.2469/faj.v68.n1.1.

B3 (2022). B3 publishes the third preview of ibovespa and other indices. *B3 Hypothetical Portfolios*, .

Bailey, D., Borwein, J., Prado, M. L., & Zhu, Q. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society.*, *61*, 458–471. doi:10.1090/noti1105.

Bailey, D. H., & Prado, M. L. (2013). An open-source implementation of the critical-line algorithm for portfolio optimization. *Algorithms*, *6*, 169–196. doi:10.3390/a6010169.

Becht, E., McInnes, L., Healy, J., Dutertre, C., Kwok, E. W., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*. doi:10.1038/nbt.4314.

Table A.5: Description of Fundamentalist Variables for the Machine Learning model.

Ratio type	**Indicator**	**Description**
Liquidity	Current ratio (CR)	It indicates the financial health of the company, that is, how much it has in available resources, goods, and rights realizable in the short term to meet the total of its short-term debts.
	Quick ratio (QR)	It indicates the company's liquidity situation to pay its debts with cash and trade credits without considering the uncertainty of selling stock.
Debt	Debt-to-Equity (DTEQ)	The share of third-party capital indicates whether or not the company depends on external resources, i.e., how much of the total assets are financed with third-party resources.
	Debt-to-EBITDA (DTEB)	Measures the revenue generated and available to pay off debt before covering interest, taxes, depreciation and amortization expenses.
Profitability	Net Profit Margin (NPM)	Measures the profits available to shareholders after deducting interest and taxes.
	Return on Assets (ROA)	Establish efficiency by asset management full use in the company's operations. Measures the use of capital to make a profit (before interest and income tax).
	Return on Equity (ROE)	Indicator of a company's efficiency in generating profits.
Market	Price-to-Earnings (PTE)	Measures the current asset price relative to the profit per asset.
	Price-to-Book (PTB)	Measures the market valuation of a company concerning its book value.
	Enterprise Value (EV)	It is a company's total value metric, often used as a more comprehensive alternative to stock market capitalization.
	Enterprise Value-to-EBITDA (EVTE)	It is a popular metric used as a valuation tool to compare the value of a company, including debt, to the company's cash earnings minus non-cash expenses. It is ideal for comparisons of companies in the same industry.
	Price-to-Free Cash Flow (PFCF)	It evaluates a company against its free cash flow, the amount it can use to pay off debt, distribute dividends, or reinvest to expand the business.

Table A.6: Description of Technical Variables for the Machine Learning model.

Ratio type	**Indicator**	**Description**
Profitability	Log-return Trimestral (R)	It measures the change in asset prices over time and has statistical properties such as stationarity and ergodicity.
	Internal Rate of Return (IRR)	It is a metric used in financial analysis to estimate the profitability of potential investments.
	Sharpe Ratio (SR)	Its risk measure is the standard deviation of returns that it aims to maximize.
Volatility	Bollinger Bands Quarterly (BB)	Provides information about price volatility, defined by a set of trend lines drawn at two standard deviations (positive and negative) from a simple moving average (SMA) of an asset's price, which adjusts to the investor's preferences.
	Volatility (V)	Measures the dispersion of portfolio returns and indicates higher volatility as the deviations from the mean increases.
	Value at Risk (VaR)	It consists of three components: a time, a confidence level, and a loss value (or loss percentage) that predict the largest possible losses.
	Maximum Drawdown (MDD)	A risk metric measures the maximum cumulative loss from a peak to the following bottom in the portfolio.
	Beta (B)	Beta determines the volatility of a portfolio that we calculate based on its deviation from a benchmark, with a beta value of 1.

Bhat, A. (2014). K-medoids clustering using partitioning around medoids for performing face recognition. *International Journal of Soft Computing, Mathematics and Control (IJSCMC)*, *3*, 2–4. doi:10.14810/ijscmc.2014.3301.

Bnouachir, N., & Mkhadri, A. (2019). Efficient cluster-based portfolio optimization. *Communications in Statistics - Simulation and Computation.*, *50*, 3241–3255. doi:10.1080/03610918.2019.1621341.

Bodnar, T., Mazur, S., & Okhrin, Y. (2017). Bayesian estimation of the global minimum variance portfolio. *European Journal of Operational Research*, *256*, 292–307. doi:10.1016/j.ejor.2016.05.044.

Brown, D. B., & Smith, J. E. (2011). Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds. *Management Science*, *57*, 1752–1770. doi:10.1287/mnsc.1110.1377.

Burggraf, T. (2021). Beyond risk parity – a machine learning-based hierarchical risk parity approach on cryptocurrencie. *Finance Research Letters.*, *38*, 101523. doi:10.1016/j.frl.2020.101523.

Cesarone, F., & Tardella, F. (2016). Equal risk bounding is better than risk parity for portfolio selection. *Journal of global optimization*, *68*, 439–461. doi:10.1007/s10898-016-0477-6.

Chen, J., & Yuan, M. (2016). Efficient portfolio selection in a large market. *Journal of Financial Econometrics*, *14*, 496–524. doi:10.1093/jjfinec/nbw003.

Chen, Y., Wei, X., Zhang, L., & Shi, Y. (2013). Sectoral diversification and the banks' return and risk: Evidence from chinese listed commercial bank. *Procedia Computer Science.*, *18*, 1737–1746. doi:10.1016/j.procs.2013.05.342.

Clarke, R., Silva, H., & Thorley, S. (2013). Risk parity, maximum diversification, and minimum variance: An analytic perspective. *The Journal of Portfolio Management*, *39*, 39–53. doi:10.2469/dig.v43.n4.5.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*, 224–227. doi:10.1109/TPAMI.1979.4766909.

Deng, S., & Min, X. (2013). Applied optimization in global efficient portfolio construction using earning forecasts. *The Journal of Investing*, *22*, 104–114. doi:10.3905/joi.2013.22.4.104.

Dincer, H. (2018). Hhi-based evaluation of the european banking sector using an integrated fuzzy approach. *The international journal of cybernetics, systems and management sciences*, *48*, 1195–1215. doi:10.1108/K-02-2018-0055.

Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., & Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, *11*. doi:10.1038/s41467-020-15351-4.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification, 2nd Edition*. Menlo Park, California: A Wiley Interscience Publication.

Economatica (2022). Indicator formulas. Available from: http://confluence.economatica.com/pages/viewpage.action?pageId=11469000#heading-IndicadoresT%C3%A9cnicos.

Elton, E. J., Gruber, M. J., Brown, S. J., & Goetzmann, W. N. (2013). *Modern Portfolio Theory and Investment Analysis - Ninth edition*.

Georgantas, A., Doumpos, M., & Zopounidis, C. (2021). Robust optimization approaches for portfolio selection: a comparative analysis. *Annals of Operations Research.*, *1*, 1–17. doi:10.1007/s10479-021-04177-y.

Gularte, S. S. F. (2021). The relevance of rural as an instrument of of agricultural policy. *Highlight News*, *186*, 52–53.

Hampton, J., J (2015). *Fundamentals of Enterprise Risk Management: How Top Companies Assess Risk, Manage Exposure, and Seize Opportunity - second edition*. New York, NY: American Management Association.

Hochreiter, R. (2007). *An Evolutionary Computation Approach to Scenario-Based Risk-Return Portfolio Optimization for General Risk Measures* volume 4448. Berlin, Heidelberg: Springer Berlin, Heidelberg. doi:10.1007/978-3-540-71805-5_22.

Holton, G. A. (1992). Time: The second dimension of risk. *Financial Analysts Journal*, *48*, 38–45. doi:10.2469/faj.v48.n6.38.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey: Prentice Hall Advanced Reference Series. doi:10.1007/978-3-662-60608-7.

Jurczenko, E. (2015). *Risk-Based and Factor Investing*. ISTE Press, Elsevier, London.

Justice, D., & Commission, F. T. (2010). Horizontal merger guidelines. *The US Department of Justice and the Federal Trade Commission.*, *1*, 1–37.

Kalayci, C. B., Ertenlice, O., & Akbay, M. A. (2019). A comprehensive review of deterministic models and applications for mean-variance portfolio optimization. *Expert Systems with Applications*, *125*, 345–368. doi:10.1016/j.eswa.2019.02.011.

Kaufman, L., & Rousseeuw, P. (1990). Partitioning around medoids (program pam). In *Finding Groups in Data* chapter 2. (pp. 68–125). John Wiley Sons, Ltd. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch2. doi:https://doi.org/10.1002/9780470316801.ch2. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470316801.ch2.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. *Statistical Data Analysis Based on the L1 - Norm and Related Methods, edited by Y. Dodge, North- Holland*, (pp. 405–416). URL: https://wis.kuleuven.be/stat/robust/papers/publications-1987/kaufmanrousseeuw-clusteringbymedoids-l1norm-1987.pdf.

Khan, A. H., Cao, X., Katsikis, P., V. N.and Stanimirovic, Brajevic, I., Li, S., Kadry, S., & Nam, A. Y. (2020). Optimal portfolio management for engineering problems using nonconvex cardinality constraint: A computing perspective. *IEEE Access*, *8*, 57437–57450. doi:10.1109/access.2020.2982195.

K.Sasirekha, P. (2018). Agglomerative hierarchical clustering algorithm- a review. *International Journal of Scientific and Research Publications*, *3*, 2–4.

León, D., Aragón, A., Sandoval, J., Hernández, G., Arévalo, A., & Niño, J. (2017). Clustering algorithms for risk-adjusted portfolio construction. *Procedia Computer Science*, *108*, 1334–1343. doi:10.1016/j.procs.2017.05.185.

Li, T., Zhang, W., & Xu, W. (2013). Fuzzy possibilistic portfolio selection model with var constraint and risk-free investment. *Economic Modelling*, *31*, 12–17. doi:10.1016/j.econmod.2012.11.032.

Lopes, A. M., & Machado, J. A. T. (2021). Dynamical analysis of the dow jones index using dimensionality reduction and visualization. *Entropy*, *23*, 600. doi:10.3390/e23050600.

Lorena, A. C., Facelli, K., Gama, J., Almeida, T. A., & Carvalho, A. C. P. L. F. (2021). *Artificial Intelligence: A Machine Learning Approach. Second Edition*. Rio de Janeiro, RJ, Brazil: LTC.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability* (pp. 281–297).

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*, 77–91. doi:10.1111/j.1540-6261.1952.tb01525.x.

Markowitz, H. (1959). Portfolio selection: efficient diversification of investments. *Cowles Foundation for Research in Economics at Yale University*, .

Marvin, K. (2015). Creating diversified portfolios using cluster analysis. *Independent Work Report Fall*, .

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform manifold approximation and projection for dimension reduction. *Statistics, Machine Learning*, *3*. doi:arXiv:1802.03426.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, *3*, 861. doi:10.21105/joss.00861.

Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case. *The Review of Economics and Statistics*, *51*, 247–257. doi:10.2307/1926560.

Metaxiotis, K., & Liagkouras, K. (2012). Multiobjective evolutionary algorithms for portfolio management: A comprehensive literature review. *Expert Systems with Applications*, *39*, 11685–11698. doi:10.1016/j.eswa.2012.04.053.

Nielsen, F. (2016). *Introduction to HPC with MPI for Data Science*. Springer Cham. doi:10.1007/978-3-319-21903-5.

Olson, D. L., & Wu, D. (2020). *Enterprise Risk Management Models - Third Edition*. Berlin, Heidelberg: Springer Berlin, Heidelberg. doi:10.1007/978-3-662-60608-7.

Paiva, F., Cardoso, R., Hanaoka, G., & Duarte, W. (2019). Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. *Expert Systems with Applications*, *115*, 635–655. doi:10.1016/j.eswa.2018.08.003.

Pealat, C., Bouleux, G., & Cheutet, V. (2022). Improved time series clustering based on new geometric frameworks. *Pattern Recognition*, *124*, 108423. doi:0.1016/j.patcog.2021.108423.

Prado, M. L. (2016). Building diversified portfolios that outperform out of sample. *Journal of Portfolio Management*, *42*, 59–69. doi:10.3905/jpm.2016.42.4.059.

Prado, M. P. (2018). *Advances in Financial Machine Learning*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Qian, E. Y. (2005). Risk parity portfolios: Efficient portfolios through true diversification. *PanAgora Asset Management, Inc.*, *1*, 1–6.

Raffinot, T. (2017). Hierarchical clustering-based asset allocation. *The Journal of Portfolio Management.*, *44*, 89–99. doi:10.3905/jpm.2018.44.2.089.

Ren, Z. (2005). Portfolio construction using clustering methods. *A Thesis submitted to the Faculty of the Worcester Polytechnic Institute. In partial fulfillment of the requirements for the Professional Masters Degree in Financial Mathematics*, .

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257. doi:https://doi.org/10.1016/0377-0427(87)90125-7.

Ruiz-Torrubiano, R., & Suarez, A. (2010). Hybrid approaches and dimensionality reduction for portfolio selection with cardinality constraints. *IEEE Computational Intelligence Magazine*, *5*, 92–107. doi:10.1109/MCI.2010.936308.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science*, *9*, 277–293. doi:10.1287/mnsc.9.2.277.

Shimizu, H., & Shiohama, T. (2020). Constructing inverse factor volatility portfolios: A risk-based asset allocation for factor investing. *International Review of Financial Analysis*, *68*, 101438. doi:10.1016/j.irfa.2019.101438.

Silva, A., Neves, R., & Horta, N. (2015). A hybrid approach to portfolio composition based on fundamental and technical indicators. *Expert Systems with Applications*, *42*, 2036–2048. doi:10.1016/j.eswa.2014.09.050.

Slime, M., B. adn Hammami (2016). Concentration risk: The comparison of the ad-hoc approach indexes. *Journal of Financial Risk Management.*, *5*, 43–56. doi:10.4236/jfrm.2016.51006.

S&PGlobal, A. D. (2022). Equity s&p 500. *A Division of S&P Global*, .

Tayali, H. A., & Tolun, S. (2018). Dimension reduction in mean-variance portfolio optimization. *Expert Systems with Applications*, *92*, 161–169. doi:10.1016/j.eswa.2017.09.009.

Tayali, S. (2020). A novel backtesting methodology for clustering in mean–variance portfolio optimization. *Knowledge-Based Systems*, *209*, 106454. doi:10.1016/j.knosys.2020.106454.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *The Review of Economic Studies*, *25*, 65–86. doi:10.2307/2296205.

Tola, V., Lillo, F., Gallegati, M., & N., M. R. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, *32*, 235–258. doi:10.1016/j.jedc.2007.01.034.

Tu, J., & Zhou, G. (2010). Incorporating economic objectives into bayesian priors: Portfolio choice under parameter uncertainty. *Journal of Financial and Quantitative Analysis*, *45*, 959–986. doi:10.1017/S0022109010000335.

Wang, W., Li, W., Zhang, N., & Liu, K. (2020). Portfolio formation with preselection using deep learning from long-term financial data. *Expert Systems with Applications*, *143*, 113042. doi:10.1016/j.eswa.2019.113042.

Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244. doi:10.1080/01621459.1963.10500845.