A MODEL-BASED APPROACH MACHINE LEARNING TO SCALABLE PORTFOLIO SELECTION

Ana Paula S. Gularte[1, 2] and Vitor V. Curtis[1, 2]

[1]Department of Aerospace Science and Technology, Aeronautics Institute of Technology (ITA), Marechal Eduardo Gomes Square, São José dos Campos, São Paulo, Brazil
[2]Department of Science and Technology, Federal University of São Paulo (UNIFESP), Cesare Mansueto Giulio Lattes Avenue, São José dos Campos, São Paulo, Brazil

ABSTRACT

This study proposes a scalable asset selection and allocation approach using machine learning that integrates clustering methods into portfolio optimization models. The methodology applies the Uniform Manifold Approximation and Projection method and ensemble clustering techniques to preselect assets from the Ibovespa and S&P 500 indices. The research compares three allocation models and finds that the Hierarchical Risk Parity model outperformed the others, with a Sharpe ratio of 1.11. Despite the pandemic's impact on the portfolios, with drawdowns close to 30%, they recovered in 111 to 149 trading days. The portfolios outperformed the indices in cumulative returns, with similar annual volatilities of 20%. Preprocessing with UMAP allowed for finding clusters with higher discriminatory power, evaluated through internal cluster validation metrics, helping to reduce the problem's size during optimal portfolio allocation. Overall, this study highlights the potential of machine learning in portfolio optimization, providing a useful framework for investment practitioners.

KEYWORDS

Portfolio Selection, Cluster Analysis, Hierarchical Risk Parity, Inverse-Variance Portfolio, Mean-Variance

1. INTRODUCTION

1.1. Motivation

Machine learning has received significant attention in modern financial research as a scalable approach to portfolio selection. One of the recurring challenges in this field is optimizing portfolios through the selection and appropriate allocation of stocks. An important consideration in portfolio selection is evaluating the intrinsic value of companies through fundamental analysis, in addition to analyzing technical indicators to strike a balance between these approaches and gain a comprehensive view of selected assets while minimizing losses during market downturns [1, 2, 3, 4]. This critical area of study in machine learning, data mining, and data science presents significant potential for advanced algorithmic techniques and data-driven models. In finance, people commonly employ classic models like the ones suggested by Markowitz [5]. Among these, the MV model stands out for achieving an optimal allocation of financial assets through diversification basedonthe expected risk of the assets for a given target return.

The Mean-Variance model was the genesis for numerous research studies that extendedMarkowitz's [5, 6] 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

David C. Wyld et al. (Eds): SCDD, CIOS, EDUR, CSEAI -2023
pp. 23-40, 2023. IJCI – 2023 DOI:10.5121/ijci.2023.120303

sensitivity in estimates of expected returns and covariances [7, 8, 9, 10, 11, 12, 13, 14, 15, 4, 3, 16, 17], among others. Such examples do not necessarily indicate that the risk-return optimization theory needs to be revised. Instead, the classical framework must modify itself 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 investment decision-making, financial portfolio allocation complex portfolio optimization methods without inputting high-quality assets a step before portfolio allocation; however, few researchers perform preliminary asset selection [16, 18]. [19] 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 management, and acardinality 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 [20, 10].

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 [21, 22, 19]. 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 [23, 21]. 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 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 [24, 25, 26, 27]. The numerical experiments present in [28, 29] highlight tthe feasibility and effectiveness of UMAP in processing data in complex systems such as the financial market.

Kalayci et al. [1] conducted a comprehensive review of the literature devoted to mean-variance portfolio optimization in recent articles from the last two decades and evidenced that machinelearning algorithms represent 12% of the solutions applied to the MV problem, with the K-meanstechnique 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 [30, 19]. Researchers currently use clustering methods in portfolio construction to group highly correlated stocks and then use these clusters to build the MV portfolio, as demonstrated by [23]. Meanwhile, Marvin [30] 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. [31] proposed an investment decision model that uses Support Vector Machines (SVMs) to classify assets and combines them with the 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 [19] 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 inputsubset 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. [16] 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.

1.2. Objective

The goal of this research is a scalable quantitative proposal via machine learning for asset selection and allocation through clustering.

Two stages divide the essence of this model: asset selection and portfolio optimization. The firststage involves integrating the UMAP method in the dimensional transformation of the time series into a new input for the clustering models. Then we apply ensemble algorithms to cluster the assets using the methods: i) K-means, ii) PAM, and iii) AHC 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 preselected in the previous step and perform the allocation with the HRP model, comparing the optimal portfolio results with two other models: i) MV and ii) IVP. 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 inputs to the optimization models. Also, in this sense, it offers investors more stable and diversified portfolios with better Sharpe ratios in the out-of-sample results [32, 33, 3].

1.3. Contributions

This article provides important contributions to the field of finance by addressing several gaps in the literature. Firstly, the study highlights the limited use of machine learning algorithms in mean-variance portfolio optimization solutions, with only 12% of studies utilizing these techniques in the last two decades. This finding underscores the need for further research to explore the potential of machine learning algorithms in portfolio optimization.

Secondly, the article identifies the lack of maturity in automated asset clustering methods for diversification, leaving room for future developments. Specifically, the K-means technique has been the most commonly used in the subcategory of research utilizing machine learning algorithms.

Furthermore, as noted in previous research, the study addresses the limited explicit knowledge of the MV model concerning performance measurement. By addressing these gaps, this article provides a useful framework for future research in portfolio optimization.

In terms of methodology, this article proposes a dynamic walk-forward method that updates cluster parameters for asset selection in each investment horizon window, improving the model's ability to generalize to new data and avoid overfitting. This approach significantly improves over existing methods and accurately represents the underlying relationships between assets.

Additionally, the article introduces a novel objective function for the asset selection process that considers various financial performance and risk metrics for each company, enhancing the risk-return tradeoff for selected assets.

Another contribution of this study is the availability of the application algorithm on Github, making the approach scalable and enabling other researchers and financial market professionals to reproduce the results obtained in this study.

Finally, the article compares and validates the results from three allocation models using out-of-sample data, comprehensively evaluating the proposed approach's effectiveness.

Overall, this article provides innovative methodologies and techniques that can improve the accuracy and reliability of portfolio optimization models. These contributions have significant implications for financial practitioners, providing them with better tools for managing risk and maximizing returns.

2. MODERN PORTFOLIO THEORY

Markowitz's [5] introduction of the concept of risk in finance as the variance or deviation from an average marked a turning point in the development of Modern Portfolio Theory (MPT) [5, 6]. The theory proposed the classic MV mathematical model [2], which quantifies the degree to which returns deviate from the mean return by weighing the squared deviations of individual asset returns from their expected return by their respective probabilities or weights and dividing this sum by the total weight or probability of the portfolio. This statistical measure provides insight into the level of risk or uncertainty associated with an asset or portfolio, making it a valuable tool for investors seeking to optimize their portfolios. As a result, this theory has revolutionized the field of finance and is widely adopted in investment management, known as the Modern Portfolio Theory.

However, as forecasts do not obtain sufficient accuracy [33], quadratic optimization models such as MV open possibilities for the emergence of new portfolio allocation methodologies, especially those that have, in the covariance matrix, its main modeling object, among them, stands out the Risk Parity [14]. 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 from contributing equally to the portfolio's overall volatility and being less sensitive to errors in parameter estimates, such as those related to the covariance matrix [34, 35]. The Inverse-Variance Portfolio (IVP) is a classic framework comparable to Risk Parity. However, it does not avoid instability problems in the face of many conditions [36].

In addition to the traditional approaches cited above [3, 33] introduced Hierarchical Risk Parity (HRP), which applies complex models, including graph theory and unsupervised machine learning, involved in this study. The methodology uses the IVP in a quasi-diagonal matrix by grouping the weights into bisections of a subset. The Hierarchical Risk Parity (HRP) method calculates the weights of groups of similar assets iteratively using inverse variance, constructing a quasi-diagonal matrix by dividing the assets into subsets. The approach involves three main steps: Tree Clustering, Quasi-Diagonalization, and Recursive Bisection, which are explained in more detail in [3, 33].

Problems of instability in the portfolio performance of quadratic optimizers are present in recent studies by [37, 3, 33] and [38] 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 [33], 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 [39]. Hierarchical clustering generates robust, diversified, and better risk-adjusted out-of-sample performance portfolios than traditional optimization techniques [40, 41]. Although its features are attractive, the empirical literature on out-of-sample portfolio performance compared to the HRP approach is still very scarce [41].

3. DIMENSIONALITY REDUCTION AND CLUSTERING METHODS

This section briefly reviews 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: AHC, which has a concept of chaining the data. K-means and PAM fall into the category of compactness with different approaches; the first (K-means) 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 (PAM) seeks to reduce the sum of dissimilarities, which guarantees the compactness property of the objects and is less sensitive to outliers and noisy data. And as a preprocessing technique, we use UMAP, a graph learning algorithm for dimensional data reduction.

3.1. Uniform Manifold Approximation and Projection

Authors [24] 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 [24, 25] for details on the mathematical foundations.

3.2. K-Means

James McQueen first proposed k-means in 1967 [42], and it has become one of the most widely used partitional clustering models due to its relative simplicity and ability to scale to large datasets. The mathematical basis of k-means involves selecting k random data points as the initial centroids in the first step of the algorithm. Next, the algorithm iteratively moves the data points between clusters to improve the clustering criterion and calculate the minimum distance between each data point and the k centroids to associate each data point with its nearest centroid. The squared error (ESS) for a cluster containing k clusters is the sum of the cluster variation, where i denotes the sample i, and j denotes the centroid j. The Euclidean distance is the most common metric to calculate the distance between two points [43].

3.3. Partition Around Medoids

Another clustering algorithm we used was PAM, proposed in 1987 by Kaufman and Rousseeuw[44]. 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.

3.4. Agglomerative Hierarchical Clustering

Finally, the third algorithm we use is 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 [45, 46]. This strategy facilitates the 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.

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 chose to use Ward's method [47] among the various types of linking strategies described in the paper, as it minimizes the total variation within the cluster after merging two clusters into one. This method calculates the distance between clusters by considering the sum of squared differences between their observations. At each step, the method considers the union of all possible cluster pairs and merges the two clusters whose fusion results in a minimum increase in information loss [48]. This approach is particularly effective for datasets with continuous variables, as it produces clusters of comparable sizes and shapes that can be easily understood and visualized. Additionally, Ward's method performs well for quantitative variables and has been the only clustering strategy in a previous study [19] that resulted in a positive outcome for investors [48]. The minimum loss in terms of the sum of squares of errors defines Ward's method.

3.5. Clustering Validation

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. The Silhouette Coefficient [49] aims to evaluate a partition present in the structure of the data set. The silhouette measure compares the disparity within each cluster and between clusters obtained by implementing a clustering algorithm. The Davies-Boudin index [50] 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.

4. METHOD

The data consists of twelve fundamental 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, it was composed of 93 stocks from 90 companies [51]. The 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 [52]. The time chosen was from January 1, 2016, to December 31, 2021 (i.e., six years of data resulting in 1504 scenarios of daily changes in the indices). The source code of our framework, database, and experiments are publicly available at: https://github.com/ComputerFinance/SCDD

The variables are Debt ratio (Debt-to-Equity (DTEQ), Debt-to-EBITDA (DTEB)), Liquidity ratio (Current ratio (CR), Quick ratio (QR)), Profitability ratio (Net Profit Margin (NPM), Return on Assets (ROA), Return on Equity (ROE), Log-return Trimestral (R), Internal Rate of Return (IRR), Sharpe Ratio (SR)), Volatility ratio (Bollinger Bands Quarterly (BB), Volatility (V), Value at Risk (VaR), Maximum Drawdown (MDD), Beta (B)) and Market ratio (Price-to-Earnings (PTE), Price-to-Book (PTB), Enterprise Value (EV), Enterprise Value-to-EBITDA (EVTE), Price-to-Free Cash Flow (PFCF)).

For evaluating the performance of the optimized portfolios, we use the closing asset price data from Yahoo Finance for the same period as the indicators dataset. We compare the three optimization methods (HRP, MV, and IVP) to validate the portfolio's financial performance using in-sample and out-of-sample data and use the most popular metrics in portfolio management for long-term horizons, which are the Herfindahl-Hirschman index, average portfolio return, annualized volatility, Sharpe ratio, maximum drawdown, and beta index.

The proposed framework, depicted in Figure 1, consists of two stages. The first stage is responsible for asset preselecting. The second stage is responsible for optimal portfolio allocation.

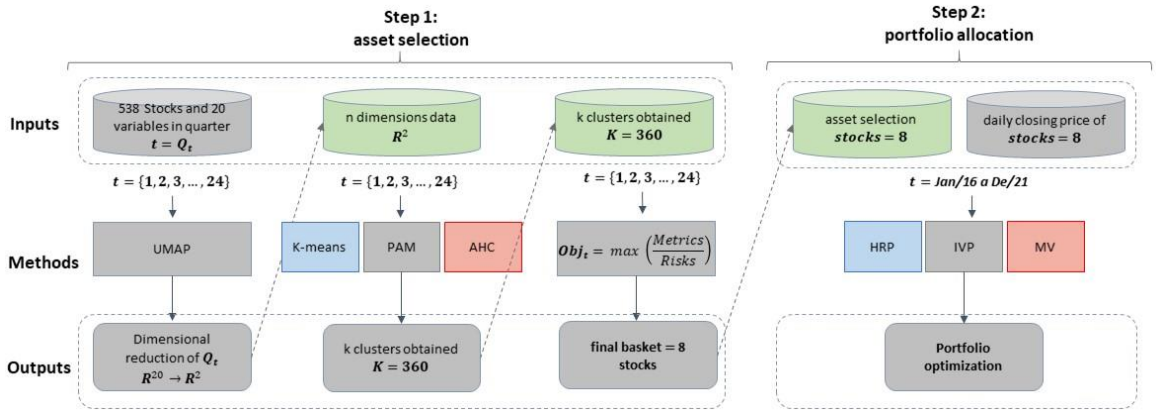

Figure 1. Flowchart for pre-selection of stocks using the clustering approach in step 1 and portfolio allocation in step 2.

4.1. First Stage Analysis: Asset Selection

To clarify the method, the first step involved extracting time series data for 538 stocks in the Ibovespa and S&P 500 indexes, using 20 fundamental and technical indicators from January 1, 2016, to December 31, 2021. This six-year data resulted in 1504 scenarios of daily changes in the indices. The dataset was then divided into quarters, labeled as Q_t, where $t = [1, 2, 3, \dots, 24]$. For each quarter t, we included each Stock's indicators and the quarterly average stock price as an additional feature. Subsequently, we processed the data individually, incorporating the pre-selection steps. Figure 1 provides a visual representation of this first step in the process.

After dividing the dataset into quarters, we applied the method UMAP dimensionality reduction model to the Q_t sets in the next stage of our analysis. UMAP is a model that can reduce a high-dimensional dataset to a lower-dimensional one, generally two or three dimensions, by adjusting its main parameters to focus on a global or local aspect of the data. We utilized this model as a preprocessing step for clustering models to improve the distance relationships between data points and enhance computational performance. These dimensionality reduction models reduce the input data's dimensionality for clustering methods and can also unfold complex high-

dimensional structures, allowing for better results. Furthermore, the data showed different distributions of elements between clusters, even without using a dimensionality reduction model as a data preprocessing step. This approach allowed us to evaluate the performance of each clustering method and select the best one for our portfolio optimization model.

After preprocessing the data with UMAP, we applied the K-means, PAM, and AHC clustering models to generate clusters from the Q_t set. The literature suggests various indexes for specifying and evaluating the appropriate number of clusters present in the data structure. While there is no consensus on which measures to use, [53] summarize the main aspects ofusing an evaluation index, with internal validation measures being among the most relevant. These measures assess the degree to which an obtained partition is justified, considering the data's compactness and separability. Therefore, we opted to use the Silhouette Coefficient and the Davies-Bouldin Index.

In the next stage, we present our approach to selecting stocks for each cluster and model based on the objective function defined in Equation 3. To break down the objective function into financial metrics (Equation 1) and risk factors (Equation 2), we aim to evaluate portfolio performance and exposure to particular sources of uncertainty. We express the function as follows:

The sum of the financial metrics that measure each company's financial health and performance denoted as Metrics, is obtained by Equation 1:

$$Equation\ 1: Metrics = CR + QR + NPM + ROA + ROE + IRR + SR, \quad (1)$$

Where CR is the current ratio, QR is the quick ratio, NPM is the net profit margin, ROA is the return on assets, ROE is the return on equity, IRR is the internal rate of return, and SR is the Sharpe ratio.

Similarly, the sum of the risk factors that measure each firm's exposure to financial and market risks, denoted as Risks, is obtained by Equation 2:

$$Equation\ 2: Risks = DTEB + EVTE + DTEQ + PTE + PTB + PFCF + V + VaR + MDD + B, \quad (2)$$

Where $DTEB$ is the debt-to-EBITDA, $EVTE$ is the enterprise value-to-EBITDA, $DTEQ$ is the debt-to-equity, PTE is the price-to-earnings, PTB is the price-to-book, $PFCF$ is the price-to-free cash flow, V is the volatility, VaR is the value at risk, MDD is the maximum drawdown, B is the beta.

The objective function of the stock selection model for each cluster is defined in Equation 3. The objective is to maximize the portfolio score, calculated by dividing the financial metrics by the risk factors. A single score represents the portfolio's overall quality for the time t. Maximizing the portfolio score leads to a more precise selection of stocks in each cluster and, consequently, a better overall portfolio performance.

$$Equation\ 3: Obj_t = max\left(\frac{Metrics}{Risks}\right) \quad (3)$$

The objective function, Equation 3, calculates two pre-selection parameters: $(i)\ Obj_{stock_t}$, which is the stock objective function in Q_t, and $(ii)\ Obj_{\text{preselect}_t}$, which is the average objective function of $P_{\text{preselect}_t}$ in Q_t. As the asset selection is an iterative quarterly process, when $t = 1$, the set referring to Q_{t+1} is returned. Then, the intersection stocks between the consecutive

preselected portfolios $P_{\text{preselect}_{t-1}}$ and $P_{\text{preselect}_t}$ are identified and kept in $P_{\text{preselect}_t}$. The stocks that do not belong to the intersection must meet the condition $Obj_{stock_t} \geq Obj_{\text{preselect}_{t-1}}$, ensuring that the preselected stocks progressively improve the average objective function of the portfolio over the quarters.

In each quarter of the dataset, we apply the method iteratively to select the optimal number of clusters defined by K using ensemble of three algorithms: K-means, PAM, and AHC. We calculate the optimal K in each continuous window using the clustering validation. The algorithms generate a score by maximizing the Silhouette Coefficient and minimizing the Davies-Bouldin Index to obtain the cluster number. When applying the clustering method, we average the values to determine the optimal cluster number with a value of $K = \{2,\ldots,20\}$.

In the final step, the algorithm selects a subset of n stocks. It divides the financial time series dataset into a cluster structure of 5 and selects 2 representative assets from each cluster for the three clustering methods with the highest objective function values in Equation 3. The resulting subset reduces computational complexity and serves as the input data for the portfolio optimization methods, HRP, MV, and IVP, based on the target number of the entire universe of assets.

4.2. Second Stage Analysis: Portfolio Allocation

In the second step, we obtain the resulting $P_{\text{preselect}_t}$ by taking the union of the two previous conditions and repeating the process until $t = 24$, when we have the final $P_{\text{preselect}_t}$ for HRP, MV, and IVP portfolio optimization methods.

The resulting subset serves as input for the asset allocation task, where assets are reordered to similar groups, ones together and dissimilar ones further apart. This reordering helps investors make more informed asset allocation decisions and build more diversified portfolios, as demonstrated in previous studies [33, 3].

The three proposed optimization methods, which include the traditional Markowitz mean-variance (MV) approach, the hierarchical risk parity (HRP) method, and the inverse-variance portfolio (IVP) technique, have demonstrated promising results in-sample. However, it is important to note that the portfolio with the minimum variance in the sample may not necessarily have the minimum variance out-of-sample. This raises concerns about the reliability of these methods and the need for further evaluation.

To address this issue, we applied the Monte Carlo method to evaluate the performance of different portfolio allocation methods out-of-sample. Monte Carlo simulations generate random data by sampling from a probability distribution, useful when actual data are unavailable or to evaluate a method's performance on hypothetical data. This study used the multivariate normal distribution to generate hypothetical stock returns based on historical data.

Although the minimum variance (MV) portfolio had a lower risk than the in-sample portfolio of the traditional risk parity's (HRP) allocation, it is important to note that the portfolio with the minimum variance in-sample may not necessarily have the minimum variance out-of-sample. To avoid the potential overfitting and selection bias issues.

Specifically, we simulated 10,000 portfolios using the same data for all methods and assessed each allocation approach's in-sample and out-of-sample performance. To evaluate the performance of each method on the datasets, the variances of the portfolios were distributed into

histograms based on 260 observations, equivalent to a one-year frequency. We calculated the average of the in-sample and out-of-sample variance distributions for this purpose.
Monte Carlo simulations are a tool for identifying methodological features that confer a preferential status over others, irrespective of arbitrary exceptions. These simulations enable us to assess and compare the efficacy and reliability of different approaches, thereby enhancing the accuracy and robustness of our model.

5. RESULTS

Our analysis examined the K-means clustering model with and without using UMAP in the preprocessing step. We applied the Silhouette Coefficient and Davies-Bouldin Index in both scenarios to determine the optimal number of clusters. As a result, two and five classes were formed without and with UMAP preprocessing, respectively. We have presented the findings of our analysis in Table 1.

Table 1. Impact of UMAP preprocessing on K-means.

Cluster Id	**Stocks per cluster**
K-means with UMAP preprocessing	
0	22,3%
1	26,7%
2	21,9%
3	14,1%
4	15,0%
K-means without UMAP preprocessing	
0	98,9%
1	1,1%

The results presented in Table 1 indicate that when applied without preprocessing, K-means failed to accurately identify the clusters, with only 1% of the data belonging to the second group. However, preprocessing with UMAP enabled us to identify the clusters well-distributedly. UMAP's nonlinear approach allows faith capture of the underlying geometric structure of the data, and the overall topology of the dataset is more accurately preserved even at low dimensions, both local and global structures.

Iteratively, we apply the method at each quarter in the dataset to select the optimal number of clusters defined by K for each of the algorithms, namely K-means, PAM, and AHC. We calculate the optimal K in each continuous window using the internal validation measures. The algorithm generates a score to obtain the cluster number that maximizes the Silhouette Coefficient and minimizes the Davies-Bouldin Index. We apply the clustering method with a value of $K = \{2,\dots,20\}$ and then average between the values to determine the optimal cluster number.

Figure 2 shows the result of the three clustering methods for the last quarter of 2021 and the number of clusters for $K = 5$.

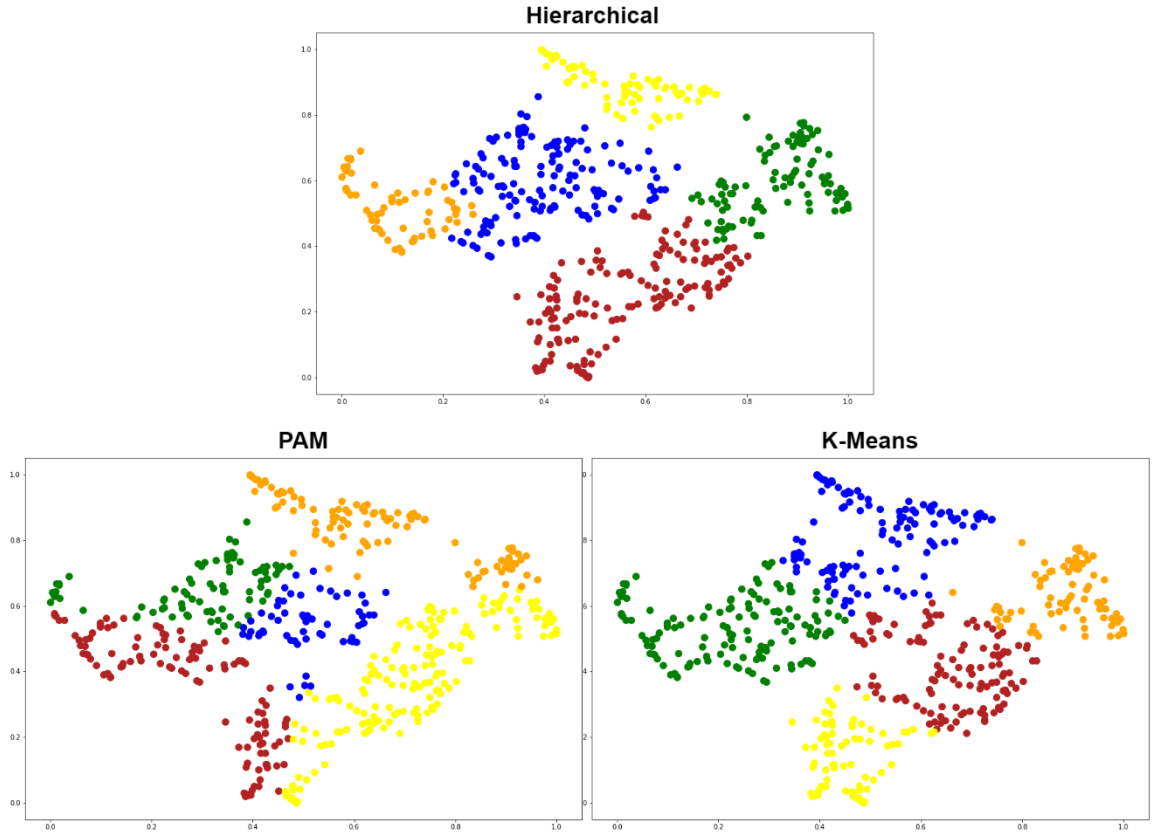

Figure 2. Clustering outputs for $K = 5$.

As a result, the method selected a set of 7 stocks from the S&P500 index and one from the 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 many different assets in the portfolio is unrealistic for individual investors. Many focus on ten or fewer assets, as evidenced in the works of [54] and [10] with five assets and [16] with ten assets, given that a very high number of assets are difficult to manage and can incur high transaction costs. In recent research, [31] 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. 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 Figure 3, 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.

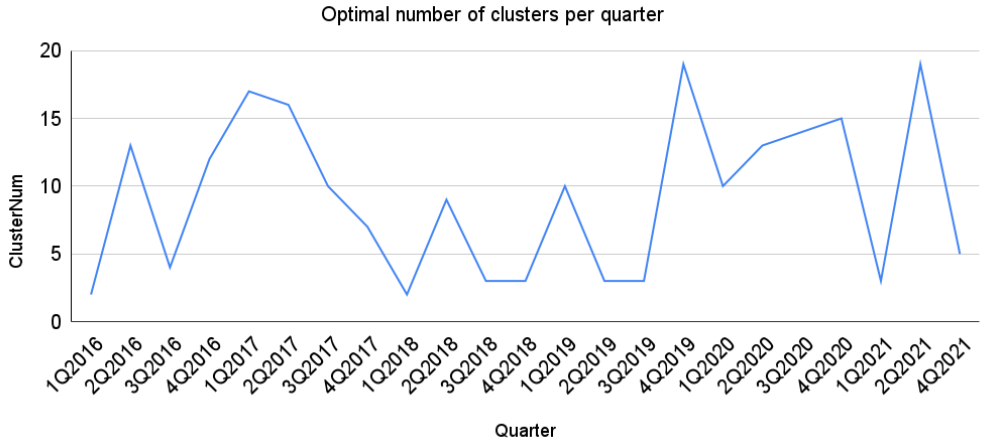

Figure 3. The optimal number of clusters per quarter.

Figure 4, through a Minimum Spanning Tree (MST) 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. In other words, it forms two clusters: one with the closest nodes, namely VRSN and NVDA, and the other with PFE and VRTX. The size of the nodes is proportional to the size of the annualized returns, and the green color signifies a positive performance in the period, which allows the calculation of the Degree Centrality (DC). The asset with the highest DC is VRSN (1.0), followed by ANET and NVDA (0.71). In parallel, BRAP4 and NLOK have the fewest links in the network and, therefore, the lowest DC (0.14 and 0.28, respectively).

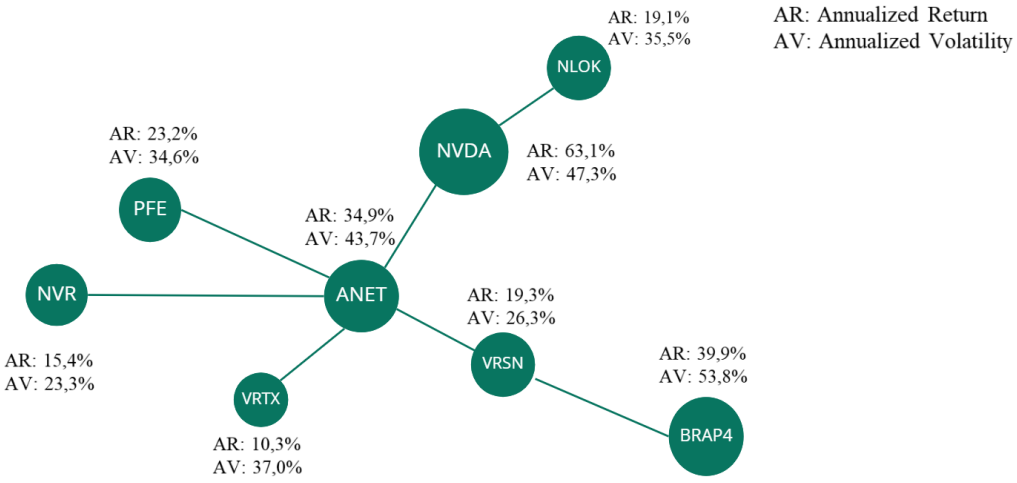

Figure 4. MST of the correlations among the selected assets.

The construction of optimized portfolios applying the classic MV model, the IVP, and HRP generate weighted portfolios presented in Table 2.

Table 2. Portfolio Optimized.

Stock	MV	IVP	HRP	Economic Sector	Index
PFE	44,80%	26,50%	30,00%	Manufacturing Industry	S&P500
VRSN	20,40%	20,80%	17,10%	Professional, Scientific and Technical Services	S&P500
NLOK	14,10%	11,40%	12,40%	Information	S&P500
VRTX	2,90%	10,40%	11,40%	Manufacturing Industry	S&P500
NVR	13,80%	12,00%	11,20%	Construction	S&P500
ANET	1,10%	7,50%	7,00%	Manufacturing Industry	S&P500
BRAP4	2,30%	5,00%	5,60%	Management of Companies and Enterprises	iIbovespa
NVDA	0,60%	6,40%	5,30%	Manufacturing Industry	S&P500

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 sectors with 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.

In Table 3, we summarize these analyzed indicators. Among the characteristics obtained from the three methods studied, we have that MV concentrated the portfolio weights in four assets at 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. For HRP, the assigned weights 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.

Table 3. Portfolio Performances.

	MV	IVP	HRP
HHI	2828	1639	1710
Cumulative Returns	164,80%	228,80%	216,90%
Annual Volatility	18,70%	22,80%	22,10%
Sharpe Ratio	1,01	1,14	1,11
Max Drawdown	-29,90%	-29,60%	-29,90%
Beta S&P500	0,79	0,91	0,89
Beta Ibovespa	0,34	0,4	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 at 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 (0.99), followed by HRP (0.96) and MV (0.84), considering a 3% risk-free. In Figure 5, 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.

Figure 5. Portfolio Performance and Benchmarks.

To test the stability of the in-sample and out-of-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 10,000 simulated portfolios whose variances we distributed into histograms obtained based on 260 observations (equivalent to one-year frequency), both in-sample and out-sample. Table 4 presents the average of the in-sample and out-of-sample variance distributions.

Table 4. Average variances of the in-sample and out-of-sample distributions.

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

The simulations show 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.

In contrast, Figure 6 illustrates that the distributions were more shifted among themselves in MV, despite the smaller averages of 3.39% and 3.52% for in-sample and out-sample variances, respectively, compared to HRP. We can attribute this greater shift to the larger difference between these averages. Therefore, this experiment reinforces the results obtained by [33, 3] 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.

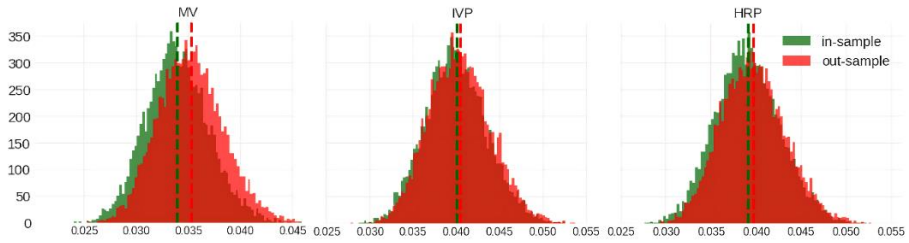

Figure 6. Monte Carlo simulations of in-sample and out-sample variances distributions.

6. CONCLUSIONS

In this paper, we introduce an innovative methodology for portfolio optimization, a recurrent issue in modern financial research for investment decisions. Our approach combines machine learning strategies with modern portfolio theory, providing additional insights to explore a dimensional reduction in high-dimensional portfolio optimization. It also uses the HRP model in a backtesting framework in each continuous window of the investment horizon via Monte Carlo simulation. The results show significant computational complexity reduction with preprocessing using UMAP, which allowed finding the clusters with the highest discriminatory power resulting in a final portfolio with eight assets. The input data for the optimization models favored the results of the computational reduction, and the HRP model stands out with better performance and higher Sharpe ratio (0.96) compared to IVP and MV. The portfolios also outperformed S&P 500 and Ibovespa in cumulative returns, with similar annual volatilities of 20%. Despite the impact of the pandemic, the portfolios recovered in 111 to 149 trading days after drawdowns of close to 30%.

Moreover, there is significant scope for future research on applying other predictive machine learning methods to composite out-sample data, implementing risk-smoothing mechanisms such as recurrence plots, refining the model for dynamic allocation, and reducing correlation with the market. However, these investigations will require more refined historical stock market data unavailable in the current dataset.

Our methodology presents an alternative financial application for portfolio managers, contributing to the literature on modern portfolio theory and machine learning theory. However, we acknowledge that portfolio selection still lacks a perfectly optimal solution. In this paper, we discuss the strengths and weaknesses of several existing machine learning methods, hoping that future researchers can develop more efficient or hybrid approaches based on the results of our processes.

ACKNOWLEDGMENTS

The authors are thankful to Economatica, one of the leading data providers for Latin American markets, for granting us access to their data platform and providing support and training in interpreting and utilizing the information. Their assistance saved us considerable time handling large volumes of data and contributed to more comprehensive analyses.

REFERENCES

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

[2] E. J. Elton, M. J. Gruber, S. J. Brown, W. N. Goetzmann, Modern Portfolio Theory and Investment Analysis - Ninth edition, 2013.

[3] M. P. Prado, Advances in Financial Machine Learning, John Wiley & Sons, Inc., Hoboken, New Jersey, 2018.

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

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

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

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

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

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

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

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

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

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

[14] E. Jurczenko, Risk-Based and Factor Investing, ISTE Press, Elsevier, London, 2015.

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

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

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

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

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

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

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

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

[23] Z. Ren, Portfolio construction using clustering methods, file:///https://core.ac.uk/download/pdf/213002338.pdf, 2005.

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

[25] L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, Statistics, Machine Learning 3 (2020). doi:arXiv:1802.03426.
[26] E. Becht, L. McInnes, J. Healy, C. Dutertre, E. W. Kwok, L. G. Ng, F. Ginhoux, E. W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nature Biotechnology 37 (2019). doi:10.1038/nbt.4314.
[27] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, C. Trapnell, Dimensionality reduction by UMAP to visualize physical and genetic interactions, Nature Communications 11 (2020).doi:10.1038/s41467-020-15351-4.
[28] A. M. Lopes, J. A. T. Machado, Dynamical analysis of the dow jones index using dimensionality reduction and visualization, Entropy 23 (2021) 600. doi:10.3390/e23050600.
[29] C. Pealat, G. Bouleux, V. Cheutet, Improved time series clustering based on new geometric frameworks, Pattern Recognition 124 (2022) 108423. doi:0.1016/j.patcog.2021.108423.
[30] K. Marvin, Creating diversified portfolios using cluster analysis, file:///https://www.cs.princeton.edu/sites/default/files/uploads/karina_marvin.pdf, 2015.
[31] F. Paiva, R. Cardoso, G. Hanaoka, W. Duarte, Decision-making for financial trading: A fusion approach of machine learning and portfolio selection, Expert Systems with Applications 115 (2019) 635–655. doi:10.1016/j.eswa.2018.08.003.
[32] D. León, A. Aragón, J. Sandoval, G. Hernández, A. Arévalo, J. Niño, Clustering algorithms for risk-adjusted portfolio construction., Procedia Computer Science 108 (2017) 1334–1343.doi:10.1016/j.procs.2017.05.185.
[33] M. L. Prado, Building diversified portfolios that outperform out of sample., Journal of Portfolio Management 42 (2016) 59–69. doi:10.3905/jpm.2016.42.4.059.
[34] 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.
[35] Qian, E. Y. (2005). Risk parity portfolios: Efficient portfolios through true diversification. PanAgora Asset Management, Inc., 1, 1–6.
[36] 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.
[37] Chen, J., & Yuan, M. (2016). Efficient portfolio selection in a large market. Journal of Financial Econometrics, 14, 496–524. doi:10.1093/jjfinec/nbw003.
[38] 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.
[39] 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.
[40] Raffinot, T. (2017). Hierarchical clustering-based asset allocation. The Journal of Portfolio Management., 44, 89–99. doi:10.3905/jpm.2018.44.2.089.
[41] 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.
[42] MacQueen, J. (1967). Classification and analysis of multivariate observations. In 5th Berkeley Symp. Math. Statist. Probability (pp. 281–297).
[43] Manning, S. H., C. D. (1999). Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts, london, England: MIT Press.
[44] Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. Statistical Data AnalysisBased 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.
[45] K.Sasirekha, P. (2018). Agglomerative hierarchical clustering algorithm- a review. International Journal of Scientific and Research Publications, 3, 2–4.
[46] Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern Classification, 2nd Edition. Menlo Park, California: A Wiley Interscience Publication.
[47] 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.
[48] Nielsen, F. (2016). Introduction to HPC with MPI for Data Science. Springer Cham. doi:10.1007/978-3-319-21903-5.

[49] 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.

[50] 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.1979.4766909.7.

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

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

[53] JAIN, A. K.; DUBES, R. C. Algorithms for Clustering Data. Englewood Cliffs, New Jersey: Prentice Hall Advanced Reference Series, 1988.

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

AUTHORS

Ana Paula, dos Santos Gularte,is anM.Sc.student in operational research from the Brazilian Aeronautics Institute of Technology (ITA) concentration area in data science. Studies and conducts scientific research on machine learning, artificial intelligence, big data, and numerical optimization methods. His research interests include the application of artificial intelligence and data science in business, finance, and economics.

VitorVenceslau Curtis received the M.Sc. and Ph.D. degrees in electronic engineering and computer science from the Brazilian Aeronautics Institute of Technology (ITA) in 2013 and 2018, respectively. In 2017, he was a Visiting Scholar with the Viterbi School of Engineering, University of Southern California (USC), under the supervision of Prof. V. K. Prasanna. Since 2018, he has been an Assistant Professor with the Computer Science Division, Department of Computer Systems, ITA. His research interests include high-performance computing, mathematical optimization, and applied mathematics.