Application of Machine Learning in Market-Neutral Pairs Trading in the Indian Stock Exchange

[1]Chetan Tayal, [2]Lalitha V.P.
[1]Student, [2]Associate Professor
[1,2]Department of Computer Science and Engineering
[1,2]RV College of Engineering, Bangalore, Karnataka

Abstract : Pairs Trading is employed widely as a market neutral trading strategy using the concepts of statistical arbitrage. The theory behind this technique is based on the idea of mean-reverting time series and is designed to illustrate the relationship between the two assets. In order to minimize risk when the market goes in only one direction, investors may benefit by profiting on one side of the bet. One of the hardest aspects of pairs trading is deciding which pairs have a link that can be called fundamental and which are merely coincidental. The proposed machine learning method in this research is unsupervised clustering for construction of the search space for pair selection, and an algorithm to evaluate the search space is presented. In an effort to recognize profitable trading pairs, not only do we realize we are picking out better pairs, these pairs have less exposure to the market and are less volatile, thus presenting less risk to an investor.

IndexTerms **- Pairs Trading, OPTICS, Unsupervised Clustering, Spread, Back test**

I. INTRODUCTION

A significant proportion of trading techniques that are created have substantial market exposure. This is proof that a large percentage of the strategies are driven by the market, rather than in-depth quantitative analysis, making them incapable of generating 'alpha'. These techniques generally perform well when the markets are doing well, but drawdowns in the markets may negatively affect the performance of many of the methods because of the substantial exposure to the market. So, one of the main issues with quantitative finance is creating strategies that are neutral to the market - methods that generate profits while also minimizing risk for investors. Pairs Trading is a widely utilized quantitative trading technique, especially for strategies that are market neutral and may perform well in both bullish and bearish market conditions.

Using a Pairs Trading allows us to make gains while hedging our losses, even if the market goes in a different way. In this study, we are looking at a common pairs trading algorithm, in which machine learning is being used. We'd want to utilize unsupervised clustering as a way to identify pairings of assets that are lucrative. First, we run PCA to accomplish data reduction, and then we use OPTICS to identify similar assets in our search area, consisting of assets that match a specific set of criteria. We use conventional methodology in which stocks from different sectors are considered in order to determine which equities are paired and form portfolios from them. All code will be made publicly accessible on Github, and all the code for the study has been developed in Python 3.6.

II. PAIRS TRADING

Pairs trading is a quantitative trading approach that employs statistical arbitrage and attempts to profit from market inefficiencies. Pairs trading is a type of investment strategy where an investor identifies two assets (also referred to as instruments) which possess some underlying economic or statistical link, and buys a long position in one asset while simultaneously selling a short position in the other asset.The concept of mean reversion lies at the heart of this method. We expect the spread between the two assets to revert to the mean over time when we use Pairs Trading. Since this is the case, we should bet on which asset is going to increase in value and which asset is going to decline in value, with the expectation of profiting from both wagers.

It is the fact that, even if the value of both assets goes up or down, we can still profit. This indicates that just one side of the wager loses money, and we can earn money on the other part. Because we may hedge our losses, this technique is market neutral. Typically, in a Pairs Trading technique, two stages are required. Each of the following sections will be examined more closely, and references to the relevant literature will be provided. Our process will also include applying unsupervised clustering as a component.

2.1 Pairs Selection

The first step in implementing the method is to identify a pair of assets that, though separate, are economically linked. The purpose of this stage is to search across the whole search space to uncover all viable candidates, and then to identify the best trading pair possibilities among them. In this approach, older literature tends to present two approaches of locating pairs. as advocated by [21] Krauss, Do, & Huck (2017) and [4] Caldeira & Moura (2013).

The second approach is to use equity groups to segment equities. Typically, you will do this search by grouping equities within the same sector and limiting the search space for combinations. Equity valuations were assigned based on how relevant the valuation method was to each equity. (The study of [9] Do & Faff (2010) and [11] Dunis et al. (2010) demonstrated this idea of meaningful clustering of equities.)

Both approaches suffer from drawbacks as well as present various advantages. This first method is capable of finding more interesting matches, but it is vulnerable to false correlations due to the larger size of the search space. When the other approach is used, we lessen the possibility of identifying erroneous associations, but we have to work with a limited number of comparisons.

Now that we have built a list of potential trade pairs, we then define a mechanism for making the final selections. A common strategy to make a pair acceptable for trading is to employ three different selection criterias. They are: Distance, Correlation, and Cointegration.

[13] Gatev, Goetzmann, and Rouwenhorst (2006) proposed the distance approach. If the total of the squared distance between two asset's historical price data is less than a specific threshold, then they select a pair for trading. While there are disadvantages to this method, such as only two assets having a narrow spread of values, two assets with negligible differences in their values won't be able to present any trading chances.

We observe the usage of Pearson correlation as a technique of identifying ideal pairs, but instead of price data they used returns data in [6] Chen, Chen, Chen, and Li (2017). Using a correlation metric, the results of [13] showed a raw average return of 1.70%. A good point has been raised in that two securities which are positively associated in terms of their returns could have an inverse relationship and exhibit unforeseen diversions in the future.

In the final step, pairs of time series that are cointegrated are selected. In other words, X and Y are two time series and their linear combination Z is cointegrated. A stationary time series returns to its mean by way of its mean reversion property. So if a divergence in the spread is seen, a convergence is likely to follow. In the study of [30] Vidyamurthy (2004), it can be seen that trial-and-error approaches for systems of strategies that utilize cointegration are implemented. Meanwhile, the cointegration technique did better in the study conducted by [15] Huck and Afawubo (2015).

To come up with our candidates, we then run a number of computer simulations that look for patterns between the two assets based on their spread which is a correlation between the two assets. For determining the spread, it is generally acceptable to utilise the ratio of past prices over a specific period of time. When the spread between the two assets (or when the spread between the two assets increases or decreases) is wide, we expect the spread to narrow soon. When the spread is mean-reverting, the strategy of Pairs Trading is able to conduct transactions.

2.2 Pairs Trading Model

To determine the pair's spread, we select a pair of assets, and then calculate the spread. Based on that spread, we choose a trading position. The model most frequently used to implement this trading strategy is based on the one defined by [13]. It is composed of four distinct stages, which are identified as follows:

1. Use the mean and standard deviation of the spread over the formation period to calculate the spread between the two assets.

2. The upper, lower, and exit threshold levels for your spread must be defined. Long positions are triggered by the upper threshold, short positions by the lower, and exits by our exit threshold.

3. Adjusting the threshold when the spread changes to keep the process in check.

4. Whenever the upper threshold is crossed, we long the spread by longing Y and shorting X. Similarly if the lower threshold is crossed, we short the spread by shorting Y and longing X. Subsequently our positions are exited when the exit threshold is crossed.

This trading model is as simple as can be, and has been used by corporations for algorithmic trading since it incorporates notions of statistical arbitrage. Predictive models are often employed at this step of the financial trading process to forecast future price movements.

There is older research which incorporates studies that use Neural Networks, as outlined in [10] Dunis, et al, (2009) and [12] Dunis et al (2015), in which they examine how the distribution of movement changes when a Neural Network is used to simulate the action. The experimental system, developed by Thomaidis, Kondakis, and Dounias (2006), uses a Neural Network-based GARCH model to simulate market movements for relative prices between two stocks. Finally, [21] Krauss et al. (2017) performed a study to measure the efficiency of deep neural networks in applying statistical arbitrage.

In this paper, we examine the application of machine learning during the initial stage, which is to identify the candidates that one would want to pair. It is the aim of this project to identify an optimal strategy for identifying pairs that eliminates exhaustive search across all assets and while minimizing the exposure to frequent difficulties. A detailed description of this strategy was given earlier; one of the most typical ways to discover this type of pair is to search for pairings in the same industry sector, as they are more likely to be impacted by a shared underlying issue. Due to how widespread this strategy is in the industry, we will likely identify pairs that are being actively traded and that will leave us with very little profit margin, which is then consumed by fees.

In addition, a full search of the universe of assets should be done. As a general rule, this approach is avoided due to two primary concerns. The primary issue is the high computation demand required to do such a detailed search, and this in turn raises concerns about the phenomenon of multiple comparison bias, which can lead to misleading pairwise comparisons.

To put it another way, in this study we explore how we can utilize machine learning to search for pairs while shrinking the search space using clustering approaches. For trades to be considered, a pair must have price cointegration. Our hypothesis is that in unsupervised clustering approaches, certain assets will be able to be picked out that do not belong in the same sector, however without encountering the problems of multiple comparison bias. Once these groupings of pairs have been found, we test them against randomly chosen pairs generated using a standard technique.

III. DATA

First, let's go over the process that we use to select our data. As of March 31st, 2021, the NSE includes over 2,000 securities with 11 key industries. We would have to perform a 1,842,240-comparison search throughout the universe of assets. It would be impossible to carry out such many comparisons since the processing resources needed would be beyond what would be available in this universe of assets. Not only that, the liquidity in this universe of assets is likely not enough to execute the trades that we want to carry out.

First, we'll focus on the top 500 assets by market capitalization listed on the National Stock Exchange because these are the most liquid assets in the market. In common parlance, this universe is known as the Nifty 500 index. So to now carry out a comprehensive search, we have reduced the number of comparisons to 124,750.

This next step removes equities that have not traded for five years or more. We are able to eliminate assets with smaller track records and less stability in price fluctuations. This manner, the prices are more likely to follow a stochastic motion rather than have false correlations with some other asset.
This eventually provides us with a universe of 396 equities, and with that, 78,210 comparisons are now possible.
A time series is established for each asset, which begins on January 1, 2016, and extends to the last day of January 4, 2021. Next, we will conduct two separate sets of experiments to evaluate unsupervised clustering for forming pairs of data. Now we have a dataset with historical price data for 396 equities.

Lastly, instead of modelling the raw price values, we model the returns generated by the assets. We were able to normalize our dataset without doing any extra preprocessing techniques by using this. Also, the work we do with lower floating-point digits at the end of the method helps speed up the results.

IV. DATA

In order to successfully form pairings, we use a two-step process. Our universe comprises of assets with price series. The first stage consists of translating the price series of these assets to a compact representation using PCA. This approach requires transforming the time series into a new representation using clustering methods, and then using unsupervised clustering methods to identify assets and find cointegrated pairings in these clusters.

We're doing two sets of experiments: The first set consists of pairs that originate from clusters of traditional sectors while the second set comprises pairs that come from clusters formed using unsupervised clustering. This is in addition to the fact that the two sets have been back-tested over two separate time periods: one that included periods of low volatility and another that included periods of higher volatility, mainly during the time of the COVID pandemic.

4.1 PCA on asset prices

In PCA, a set of observations made up of possibly correlated variables is transformed into a set of linearly uncorrelated variables by using an orthogonal transformation. Principal components are variables which have no significant correlation with other variables. This allows us to reduce a huge group of variables into a smaller subset that contains most of the information. Our dataset contains 1286 data points for each asset's time series data.

After studying section 2.1, we use the returns data to derive price information and use PCA on the information contained in the returns space. Converting pricing data into returns data helps to standardize the prices of varying ranges as well since prices of different equities can exist on different price ranges. The formula in Equation 1 below is used to transform price data into return data.

$$R[i][t] = (P[i][t] - P[i][t-1]) \div P[i] \qquad (1)$$

where *P[i][t]* represents the pricing data for an asset '*i*' at '*t*' time period and *P[i][t-1]* is the price series at time '*t-1*'. By getting the daily return on the equity and then taking the percentage of the daily return, we are able to derive the daily percentage change.

In addition to providing the standardisation of price range, the conversion of pricing data into returns data also affords the opportunity to normalise the data. We found that, after doing PCA, over 90% of the asset return information was found in the first three primary components.

The graph in Fig. 1 displays the overall impact of each individual component, up to the fifth component. From the graph, we can see that the first three components of time series data contain the majority of the information.

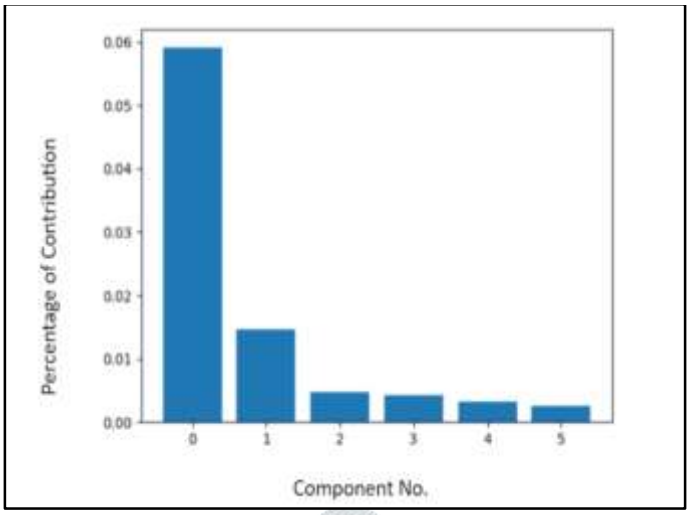

Figure 1. Histogram Plot of PCA

4.2 Unsupervised Clustering

In order to represent a specific asset, we now construct a compact image of it. Next, we establish clusters within our overall set of assets. The simple step of using information about an equity to cluster it into an existing sector is our first cluster formation. Table 1 illustrates the number of stocks in the various sectors, according to the sectors on the table. 11 clusters appear in the market, with each cluster representing an existent sector. Once the statistical tests have been completed, only the assets in their own cluster will be compared to each other. By executing an exhaustive search across a cluster, we avoid the effort required to conduct an exhaustive search across the universe.

This time we use unsupervised clustering in the second cluster set. We believe that unsupervised learning can discover relationships buried in a narrow search space, even if they aren't connected to the same sector's assets. For increased meaningfulness, we implement several ground rules before constructing the clusters.

1. While traditional clustering methods like K-Means require specifying the number of clusters in advance, we don't need to do that here.
2. There is no requirement that securities be grouped to a specific cluster. This allows us to have ungrouped assets.
3. Equity assignments should be tightly limited to the cluster to which they are given.
4. There are no assumptions about the form of clusters as we develop our theories, hence no need to assume the distribution of the data of the equity search space.

Table 1. Sector Wise breakdown of NSE 500

S.No	Sector	No. of Assets
1	Industrials	62
2	Basic Materials	75
3	Healthcare	40
4	Energy	11
5	Utilities	15
6	Consumer Cyclical	64
7	Consumer Defensive	31
8	Communication Services	12
9	Technology	24
10	Real Estate	9
11	Financial Services	11

When outliers are no longer required, we no longer have to classify every asset in the universe. We maintain the assignments confined to one cluster for one equity otherwise comparisons will occur out-of-cluster because now assignments are calculated between assets that belong in the same cluster.

We arrive at this conclusion with density-based grouping rules in mind. Instead of assuming that the data has a known distribution, we can let the data form clusters with any forms (like a gaussian one for example). To meet rule 1 and 2, we are able to use a density-based approach and we will be able to handle outliers since it does not require us to specify the number of clusters.

The algorithm of choice is DBSCAN, which is one of the most often used density-based clustering algorithms. The DBSCAN cluster identification algorithm uses the density of the data points to find clusters in the search space. The two parameters that are used in DBSCAN are e (the minimum distance between two cluster neighbors) and minPts (the number of points that is the smallest within an actual cluster). If a data point lies outside of a cluster with a minimum number of points (between minPts and e) and an outer radius of, that data point is designated as an outlier.

DBSCAN, however, still has one flaw. Clusters are assumed to be equally dense, thus each cluster has the same number of entities. In some cluster locations, e with a constant value might not be appropriate. In Figure 2, we see that cluster A, B, and C have the same value of e, but cluster A1 and A2 are separate entities because they display a distinct density value.

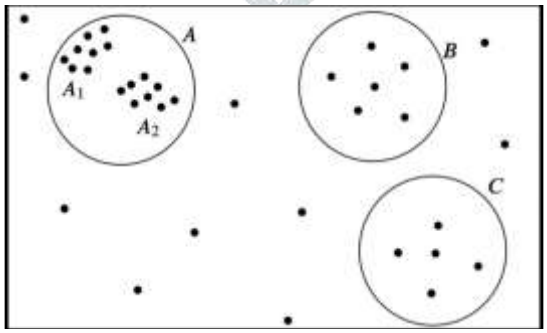

Figure 2. Clusters of Varying Density

We propose using the OPTICS clustering approach to get around this difficulty. OPTICS follows DBSCAN, but varies the e parameter. By using this algorithm, we will be able to locate a group of equities in the generated equity space even if they have different densities, which will allow us to group equities according to their portfolio composition regardless of the overall amount of assets.

4.3 OPTICS Clustering

For the results of running OPTICS clustering on our dataset, we observed the following. We see that 277 equities are allocated to a single outlier cluster as OPTICS managed to identify 119 out of the total 396 into some form of cluster that we have in the search area. To see the 29 clusters that were generated amongst the 119 assets, we employ the use of TSNE and present the figure in Fig. 3.

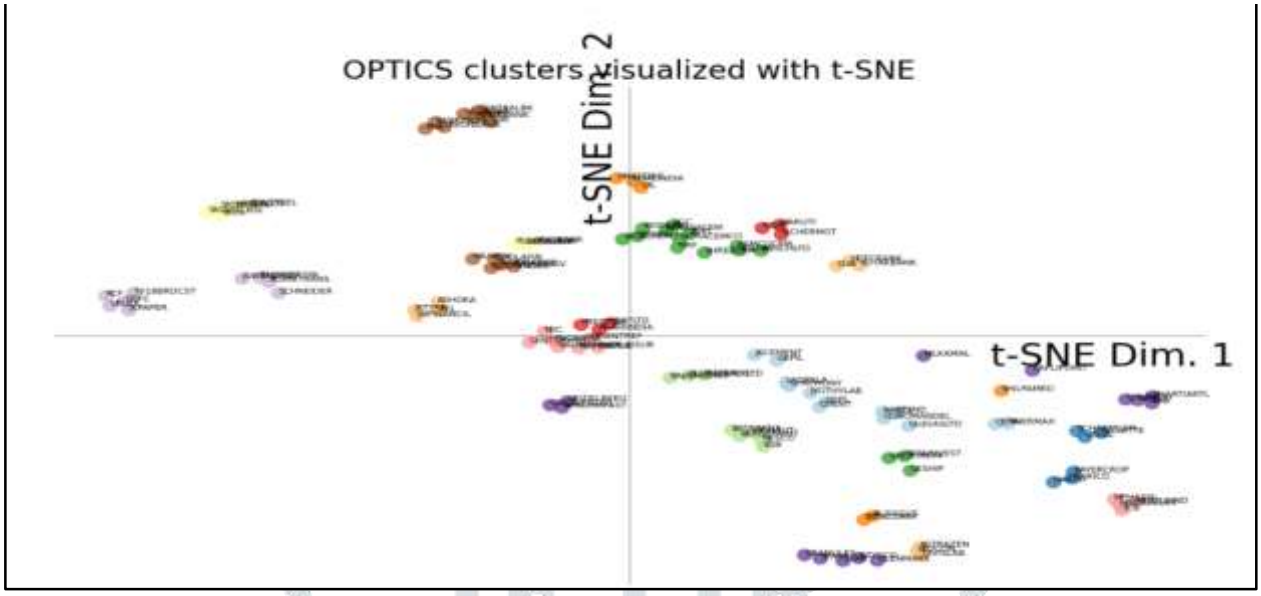

Figure 3. Plot of OPTICS clustering on our returns data. The 29 clusters shown here do not contain equities that OPTIC failed to classify into a cluster.

An in depth look at the generated clusters showed that although many clusters were made from assets that were part of the same sector, they also comprised assets belonging to completely different sectors. 20 clusters out of a total of the 29 clusters originally formed demonstrated this property.

Our original premise was confirmed, demonstrating that unsupervised clustering generates clusters made up of shares from different sectors and is able to discover linkages that might not be obvious if all equities from the same sector were grouped. The number of comparisons that is required right now is 207. Once we have generated these clusters, we will now perform a series of statistical tests to choose the pairs of cointegrated variables that remain.

4.4 Pair Formation Rules

After we produce a collection of assets, we identify a set of selection rules to conduct the search for pairs. Clusters generated using sector-wise information and optics information are identified by naming them "Sectorwise Cluster" and "Optics Cluster" respectively.

1. If the assets are well-integrated, we'll choose a pair. Cointegration of a set of time series is defined as when all series are I(1) (known as “integrated 1”), and the weighted sum of them is I(0) (known as “integrated 0”). A common feature of all I(0) series is that they are all mean-reverting.

2. The second rule seeks to quantify the likelihood of a time series either being mean-reverting or trending. When the time series is mean-reverting, values lie somewhere between 0 and 0.5. Even though the time series reveals that it will not revert to the mean in the future, if the value obtained is more than 0.5, then the time series is not expected to follow its mean reversion pattern in the future. The second rule reinforces the mean-reversion property of the spread by using the Hurst component that tries to quantify the likelihood of a time series being either mean-reverting or trend following. A value lying between 0 and 0.5 is a sign that the time series is likely to be mean-reverting

3. This third criterion assumes that the time series will always return to its mean when it has completed half of its half-life. Basically, we want to filter out series that either go back to their mean within a day (which is less than a week) or a year (which is greater than a year). This would ensure that all distributions with misleading behaviour and which aren't likely to return to their mean are eliminated.

With these principles now put out, we've created a framework for pairing. First, we do an analysis that utilises two independent ways to define clusters (based on sectorwise information or optics based). Following this, we look at equities within the same clusters to check if they conform to the previously mentioned rules. Any combination that fulfils all three requirements will be designated as experiment candidates and be mentioned in section 5.1. We will then perform a backtesting experiment by running a simulation in Python using Backtrader for the pairs in the experiment cluster.

V. BACKTESTING

Back testing is a strategy evaluation procedure, similar to simulation, in which historical data is examined to forecast how well the strategy would have performed under a specific set of conditions in the past. To assess the accuracy of our backtests, we examine the resulting profits and use them to estimate future profits. Backtesting can be defined as running a trading strategy on real market data and evaluating our strategy by using the results from backtesting. An analysis of the strategy also includes determining the level of risk and the financial return to be gained before we commit actual capital to the approach.

The first step to testing and validating a strategy is to run a backtest that yields positive results. A negative or positive result in real time might depend on the investor's perspective, but it is much more likely that a negative backtest will fail. A positive result can mean either mean higher profits at the expense of higher risk or vice versa. Next we will look over the experimental setup, the metrics we us for evaluation of the strategies and finally take a look at the results obtained.

5.1 Experimental Setup

Figure 4. Normalized price series plot over 5 years for a cluster generated from OPTICS.

We will be conducting a total of four experiments in our experimental setup. A backtesting experiment for clusters formed using OPTICS and clusters generated using Sector information is conducted on two different time frames in order to evaluate the efficiency of clustering as a superior way of choosing pairs. In Fig. 4, you can see an example plot for one of the clusters that was created.

One time frame corresponds to a period when the market was reasonably steady (2017-2019), while the other corresponds to a period when the market suffered significant volatility (2017-2019). The pairs are then back tested in a relative future time frame to determine their effectiveness under previously unforeseen situations. In this way, we will be able to ensure that our plan actually has predictive potential.

Backtesting was carried out with the help of Backtrader, which is a free and open-source backtesting library built in Python. Backtrader allows us to concentrate on designing reusable trading methods rather than having to worry about developing the infrastructure to support those techniques. In each single experiment, we have a different cluster based on the strategy used to construct the cluster as well as two different time frames, for a total of four possible clusters in each experiment. After that, we produce the pairs for each cluster using the technique outlined in section 4.4.

N number of pairs are now contained within each set of clusters, and we form n-separate portfolios, each given a starting investment of Rs. 1,00,000 to carry out trading. For the purpose of calculating the spread between the returns of the two equities, we first do an OLS regression between their respective time series returns. The normal equation is used to conduct regression in this case, assuming that X1 and X2 correspond to the returns of two different stocks. The procedure is as follows:

$$\phi = (X^t * X) * (X^t * y) \qquad (2)$$

If slope is *m* and the intercept is *b* , which have been obtained via regression, then the spread *S* is calculated as follows:

$$S = X1 - (m * X2 + b) \qquad (3)$$

Finally the normalized spread can be calculated using :

$$S_{normalized} = (S - S_{mean}) \div (S_{std}) \qquad (4)$$

This normalized spread is used to generate our trading signals. The upper and lower thresholds are defined on this spread to identify points of performing a BUY or a SELL, as discussed earlier in 2.2.

Lastly we use an industry standard 1% commission on each trade to keep the simulation environment as close as possible to real life trading situations.

Finally for each pair in the cluster, backtesting is carried out, and the average of the evaluation metrics outlined in the following section is used to calculate the final metrics in each experiment. In subsection 5.3, we conclude by comparing the outcomes of all of the studies.

5.2 Evaluation Metrics

To gauge the effectiveness of our trading techniques, we utilised the following criteria. It is important to understand that these values are relative to each experiment.

1. Final Portfolio Value Average : In the experiments, n portfolios are created for n pairings. For each pair, each portfolio starts out with the identical initial capital. After running through the experiment, we calculate the average portfolio value at the end of the backtesting period.

2. Final Drawdown Average : The Drawdown of a portfolio measures the financial exposure to the market. We conduct an experiment in which we look at the drawdown experienced by each portfolio and then compute the average drawdown across all portfolios.

3. Sharpe Ratio : It is the profitability of an investment when the risk-free rate is above the market interest rate, taking into account the standard deviation. The results of this calculation are known as the return on investment (ROI), which includes both the risk-free return and the variance of the investment.

4. Risk Ratio: The variance of the returns is squared and compared to the mean daily return to determine risk ratio. The distance of the return from the mean is inversely proportional to volatility and hence is an indication of higher risk.

5.3 Experimental Results

At this point, we have identified our evaluation measures, and below is our research findings. Back testing results have been summarised in Table 2.

Table 2. Results from our 4 experiments

Type	Formation Period	Backtest Period	Average Value of Portfolio	Average Drawdown	Sharpe Ratio	Risk Ratio
Optics	01/01/17-01/01/19	01/01/1-01/01/21	102078.95	13.56%	1.43	0.99
Sector wise	01/01/17-01/01/19	01/01/1-01/01/21	91015.05	24.22%	-0.68	5.5
Optics	01/01/19-01/01/21	01/01/2-01/04/21	100148.41	0.88%	0.018	0.219
Sector wise	01/01/19-01/01/21	01/01/2-01/04/21	99924.40	0.99%	-0.013	0.215

For our two different sets of backtests, the results reveal that pairing of stocks chosen from clusters generated using OPTICSCAN performed better than pairs of stocks chosen from sector-wise clusters. In summary these were the observations :

- On average, the portfolio value obtained for OPTICS portfolios was higher.
- The resulting OPTICS average drawdown implies that portfolios are less exposed to the market.
- OPTICS portfolio's better Sharpe Ratio shows it has higher risk-adjusted returns.
- Looking at the Risk Ratios, it indicates that the OPTICS portfolios have lower overall risk than sector wise portfolios.

Backtrader generated the chart in Figure 5, which displays an example trade from one of our pairs. On the upper graph, you can see GILLETTE's price series. On the lower graph, you can see SCHAEFFLER's price series. On the graph located in between the two, you can see a representation of the spread between the two stocks.

Figure 5. Trade Execution plot generated using Backtrader. Red arrows show a SELL signal and Green arrows show a BUY signal.

VI. CONCLUSION

In light of our experimental findings, here is what we can say:

- Based on our experimentation, unsupervised clustering is a better apriori for portfolio construction, because it generates portfolios with more stable returns while reducing the exposure to the market.
- Because pairings created via unsupervised clustering have a lower average risk ratio, we may assume that pairs generated this way have less risk for investors, allowing us to make use of leverage for higher profit gains.

REFERENCES

[1] Amenc, Noël & Malaise, Philippe & Martellini, Lionel & Sfeir, Daphne. (2003). "Tactical Style Allocation—A New Form of Market Neutral Strategy". The Journal of Alternative Investments. 6. 8-22. 10.3905/jai.2003.319079.

[2] Baek, Seungho & Glambosky, Mina & Oh, Seokhee & Lee, Jeong. (2020). "Machine Learning and Algorithmic Pair Trading in Futures Markets". Sustainability. 12. 6791. 10.3390/su12176791.

[3] Batchu, Satish and Radha, K.V., "Cointegration and Causal Relationship between Exchange Rates and Stock Returns: A Study on Indian Context" (December 11, 2015). Available at SSRN: https://ssrn.com/abstract=2702218.

[4] Caldeira, J., & Moura, G. V. (2013). Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy. Available at SSRN 2196391.

[5] Chaudhuri, Kausik & Wu, Yangru. (2003). "Mean reversion in stock prices: Evidence from emerging markets". Managerial Finance.

[6] Chen, H., Chen, S., Chen, Z., & Li, F. (2017). Empirical investigation of an equity pairs trading strategy. Management Science.

[7] Christian L Dunis & Jason Laws & Jozef Rudy, 2011. "Profitable mean reversion after large price drops: A story of day and night in the S&P 500, 400 MidCap and 600 SmallCap Indices," Journal of Asset Management, Palgrave Macmillan, vol. 12(3), pages 185-202, August.

[8] David Bowen, Mark C. Hutchinson and Niall O'Sullivan, "High Frequency Equity Pairs Trading: Transaction Costs, Speed of *Execution and Patterns in Returns*", SSRN, 2019.

[9] Do, B., & Faff, R. (2010). Does simple pairs trading still work? Financial Analysts Journal, 66(4), 83–95.

[10] Dunis, C. L., Laws, J., & Evans, B. (2009). Modelling and trading the soybean-oil crush spread with recurrent and higher order networks: A comparative analysis. In Artificial Higher Order Neural Networks for Economics and Business (pp. 348–366). IGI Global.

[11] Dunis, C. L., Giorgioni, G., Laws, J., & Rudy, J. (2010). Statistical arbitrage and high frequency data with an application to eurostoxx 50 equities Working paper. Liverpool: Business School.

[12] Dunis, C. L., Laws, J., Middleton, P. W., & Karathanasopoulos, A. (2015). Trading and hedging the corn/ethanol crush spread using time-varying leverage and nonlinear models. The European Journal of Finance, 21(4), 352–375.

[13] Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. The Review of Financial Studies, 19(3), 797–827.

[14] Guirguis, Michel, "Return Based Style Analysis of Equity Market Neutral Hedge Funds" (May 20, 2020). Available at SSRN: https://ssrn.com/abstract=3606480.

[15] Huck, N., & Afawubo, K. (2015). Pairs trading and selection methods: Is cointegration superior? Applied Economics, 47(6), 599–613.

[16] J Van Greunen et al., "The Prominence of Stationarity in Time Series Forecasting" , Journal of Studies in Economics and Econometric, 2014.

[17] Johannes Stubinger, Jens Bredthauer, "*Statistical Arbitrage Pairs Trading with High-Frequency Data*", International Journal of Economics and Financial Issues, 2017.

[18] Jose Fernando Ospino Lopez, Luis Fernandez Payeras and Oscar Carchano Alcina, "*Improving Pairs Trading Using Neural Network Techniques and Fundamental Ratios*", SSRN, 2020.

[19] José Pedro Ramos-Requena, Juan Evangelista Trinidad-Segovia, and Miguel Ángel Sánchez-Granero, "Some Notes on the Formation of a Pair in Pairs Trading", MDPI, 2020.

[20] Kamenshchikov, Sergey & Drozdov, Ilia. (2016). "Fractal Optimization of Market Neutral Portfolio".

[21] Krauss, C. (2017). Statistical arbitrage pairs trading strategies: Review and outlook. Journal of Economic Surveys, 31(2), 513–545.

[22] Leung, Tim & Li, Xin. (2015). "Optimal mean reversion trading with transaction costs and stop-loss exit". International Journal of Theoretical and Applied Finance. 18. 1550020.10.1142/S021902491550020X.

[23] Li, Zhixi & Tam, Vincent. (2018). "A Machine Learning View on Momentum and Reversal Trading". Algorithms. 11. 170. 10.3390/a11110170.

[24] Mushtaq, Rizwan, "Augmented Dickey Fuller Test" (August 17, 2011). Available at SSRN: https://ssrn.com/abstract=1911068.

[25] Ovadia, Y., "*Can You Trust Your Model's* Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift",NeurIPS, Vol 32, 2019.

[26] R.V.D Have, "*Pairs Trading Using Machine Learning : An Empirical Study*", Semantic Scholar, 2018.

[27] Rasekhschaffe, Keywan and Jones, Robert, "Machine Learning for Stock Selection" (February 8, 2019). Financial Analysts Journal, vol. 75, no. 3 (Third Quarter 2019), SSRN: https://ssrn.com/abstract=3330946.

[28] Rizwan Raheem Ahmed, Jolita Vveinhardt et al, "Mean reversion in international markets: evidence from G.A.R.C.H. and half-life volatility models", Taylor and Francis - Economic Research, 2018, pp. 1198-1217.

[29] Thomaidis, N. S., Kondakis, N., & Dounias, G. (2006). An intelligent statistical arbitrage trading system. SETN.

[30] Vidyamurthy, G. (2004). Pairs Trading: quantitative methods and analysis, (vol. 217). John Wiley & Sons.