BUSINESS SCHOOL
AALBORG UNIVERSITY

Cointegrated Pairs Trading with Gold Tracking ETFs

Rodrick Ngeh 20212049

Finance Master Thesis |
05-02-2024

Abstract

This thesis questions the profitability of pairs trading using two US Gold Tracking ETFs, namely the SPDR Gold Shares (GLD-US) and iShares Gold Trust (IAU-US). The data used ranges from 1st January 2015 to 25th October 2023, by implementing a cointegration approach. The empirical results suggest that the two ETFs are co-integrated and stationary during the tested periods, with the statistical properties of the ADF and the Johansen tests used to test the data. The strategy's performance during the backtesting period is compared to that during the training period to evaluate if the strategy is profitable. The performance results in terms of returns show that the training period outperforms the backtesting period, despite the backtesting period still producing sufficient annualized returns. This leads to the efficient market hypothesis to be put into question as this should not be the case. The sensitivity analysis determined the strategy to be robust as the scenario results did not deviate so much from the backtesting period results. Despite the limitations of this study, the results conclude that pairs trading is still profitable, even though the strategy's efficiency drops when introduced to new data.

Key words: Cointegration, Pairs trading, Statistical Arbitrage, Stationarity, Efficient market, Sensitivity Analysis, Profitability.

Table of Contents

1. Introduction

Traders who seek to make profits from anomalies in financial markets use statistical arbitrage techniques such as pairs trading. The concept of pairs trading came to light in the mid-1980s at Morgan Stanley, as Nunzio Tartaglia's group of physicists, computer scientists and mathematicians on Wall Street generated millions of profits from their trading algorithm. After continuous losses, the group disbanded in 1989 (Huck, 2010). Pairs trading was introduced in academic literature by Gatev et al. in 1999 (Gatev et al. 1999). This set the basis for pairs trading before several authors began to test, verify and evolved the distance approach used by Gatev et al. (1999). A more statistic-based method to pairs trading using cointegration was introduced by Vidyamurthy (2004).

This thesis aims to test for cointegration between two gold tracking Exchange Traded Funds (ETFs) whose prices follow the same path. The assumptions of co-movement of the assets and mean reversion stated by Schizas et al. (2011) aim to exploit stock price divergence, as they take a short position on expensive stocks and long the cheaper stock, thereby making a profit when both prices converge/mean revert to their fundamental characteristics.

The increasing popularity of ETFs and the high frequency at which they are being traded gave rise to this thesis and the thesis problem statement; **if a cointegrated pairs trading strategy will be a good and efficient method that yields profit using the two chosen ETFs.** The thesis problem will be tackled by first, testing the two ETF prices for cointegration and persistent cointegration through the Johansen and Augmented Dickey Fuller methods, then regressing one of the price on another to calculate the constant and coefficient of the dependent variable, calculating the spread from the resulting regression equation afterwards, identifying trends using z-score of moving average, building a trading strategy to take advantage of mispricing, followed by returns computation. Finally, the strategy will be backtested using new data, in order to evaluate its performance. This thesis will therefore test the profitability of two Gold Tracking ETFs, since several studies: (Do and Faff (2012); Rad et al. (2015); Smith and Xu (2017)) have indicated a decline/disappearance in profitability. This entails taking advantage of mispricing by challenging the concept of Market Efficiency through the implementation of cointegrated trading strategies that will generate positive returns as shown by other studies: Ackaert and Tian (2008); Petajisto (2016). This thesis will by no means compare its results to those of other articles. Its main purpose is to derive profits or losses by implementing a co-integrated pairs trading strategy. The question of, if profits are actually declining as shown by other articles, which can be determined by comparing this thesis' results to those of other articles, is another topic which will not be discussed in this thesis.

The thesis will be presented and structured in an intuitive manner without much technicality. Thereafter, section 1 above contains the introduction, with the definition, history and concept of statistical arbitrage coming in section 2. Literature review will be in section 3, closely followed by data and methodology in section 4. The backtesting, performance measures and result analysis will be in sections 5, 6 and 7 respectively. Section 8 will provide the sensitivity analysis, section 9 with the assumptions, followed by the thesis limitations in section 10, and the conclusion will come in last in section 11.

2. Statistical Arbitrage Definitions

Despite the numerous reviews of statistical arbitrage from academics and practitioners over the years, it is hard to pin a precise definition to the term. Avellaneda, M. and Lee, J. H. (2008) defined the term in a manner that includes several strategies with systematic trading signals, market trades and statistical methods. Montana, G. (2009) on the other hand looked at statistical arbitrage as an investment strategy that takes advantage of patterns observed in financial data streams. Do, B., Faff, R. and Hamza, K. (2006) stated statistical arbitrage as an equity trading strategy that finds mispricing between stocks through time series methods.

Engelberg et al. (2009), Caldeira and Moura (2013) and Vidyamurthy (2004) describe pairs trading as statistical arbitrage, which implies taking advantage of the mean reversion characteristics of two identical securities through an established trading algorithm. Pairs trading is further defined as a risk arbitrage strategy by Do and Faff (2012), elaborating that pairs trading does not profit from market trends but rather from two opposing trading positions from two securities that are a pair. The opposing positions are both long and short and are seen as market neutral (Vidyamurthy, 2004; Gatev et al., 2006; Huck and Afawubo, 2015; Schizas et al., 2011).

2.1. Statistical Arbitrage History

Nunzio Tartaglia developed the idea of pairs trading together with a group of mathematicians, computer scientists and physicists in the 1980s at Morgan Stanley. The initial idea was to trade two securities by identifying those pairs of securities that move together over time, but usually diverge from equilibrium for a brief period before converging back (Vidyamurthy, 2004; Gatev et al., 2006; Thorp E.O., 2003). Tartaglia's team made a profit of $50million in trades in 1987 (Gatev et al. 2006). The team dissolved in 1989 after making losses in the years that followed the profit year. From here on, pairs trading became a very common and popular investment strategy for hedge funds and investors (Vidyamurthy, 2004; Gatev et al., 2006; Engelberg et al., 2009). Statistical arbitrage has since evolved, and it is being used on equities, commodities and cryptocurrencies.

2.2. Fundamentals of Pairs Trading

Pairs trading is a trading technique in which the price of two assets are analyzed and compared at the same time. Two categories of analysis can be identified when it comes to how prices are paired - fundamental and technical analysis. Fundamental analysis looks at company properties, company industry, its current situation, and the entire economy. Technical analysis on the other hand focuses on historical prices, which is the basis of this thesis (Ehrman, 2006). Appropriate pairs when using technical analysis must be cointegrated. Once a pair is found to be cointegrated, it is essential to evaluate if the current pricing still follows past historical pricing methodology relative to each other (Vidyamurthy, 2004).

Pairs trading is fundamentally a trading strategy that exploits market mispricing through the use of statistical arbitrage (Gatev et al., 2006; Huck and Afawubo,2015). In this regard, the fundamental hypothesis that the market is fully efficient is put to the test. Pedersen (2015) defines an efficient market as a market where prices reflect all the relevant available information, as well as the fundamental value of the security. This therefore means that two securities that are close substitutes or that yield similar returns should have the same price (Gatev et al., 2006). If the market is efficient, there is no need for active investors trying to beat the market as market returns reflect the best risk returns. With this claim, Pedersen (2015) asked the question of whether it is the market or investors that are inefficient or perhaps both.

A pair trader’s role is to test for co-integration between a pair of assets. If cointegration exists, the trader then evaluates if their prices have been moving together historically (Vidyamurthy, 2004). If not, then their prices are judged to have momentarily deviated from each other. The element of arbitrage here is the ability of the trader to recognize which asset is overpriced or underpriced asset relative to the other. The trader can then sell (short) the overpriced asset while buying or taking a long position on the underpriced asset. But because both asset prices are cointegrated, it is believed that they will eventually go back to how they were historically (Ehrman, 2006). The process of both prices going back to their historical levels is described as mean reverting.

In summary, pairs trading is a strategy that exploits divergence in asset prices with the hope that they converge back to equilibrium.

3. Literature Review

The primary focus of this thesis is on the cointegration approach presented by Vidyamurthy (2004) and Caldeira and Moura (2011), who made the most of the cointegration relationship between two assets in order to evaluate the performance of pairs trading strategies. Studies like Denis et al. (2010) based their results on the Engle-Granger 2-step method also used by Caldeira and Moura (2011) to test for cointegration. Other studies like Dunis and Ho (2005) and Afawubo (2014) used the Johansen's approach. Both methods, though different, are used to test long-term cointegrated relationships between assets.

Gatev et al. (2006), Caldeira and Moura (2013) and Smith and Xu (2011), define cointegration approach to pairs trading as an approach that involves selecting pairs with the same cointegration order, testing the pairs through Augmented Dickey-Fuller (ADF) test, with the objective of finding pairs with mean reverting spread.

Smith and Xu (2017) investigated the different methodologies and parameters involved in the cointegration of stocks using a large data sample and concluded that it is not profitable. The large sample size does not really apply to this thesis as the thesis aims at using just 2 Gold Tracking ETFs. Smith and Xu tested a 9 and 12 months formation period and showed that the 12 months period had more returns than the 9 months and suggested that a longer formation period could be more profitable.

Schizas et al (2011) indicated that profitability stayed robust irrespective of the number of best selected pairs used during the trading period. Their study equally found a decrease in the portfolio standard deviation and sharp ratio due to larger number of pairs.

4. Data and Methodology

The thesis is aimed at finding if the two chosen ETFs can generate profit through the set trading strategy. To come to this conclusion, quantitative analysis will be conducted with statistical models applied to financial data, which are time series ETF closing prices. Econometrics is viewed as the application of statistical models to analyze economic data (Stock & Watson, 2015). For the processing and interpretation of data, R and R-Studio was used for coding and computations.

Firstly, data will be downloaded and converted to the natural logarithmic form, after which it will be tested for cointegration using the Johansen method. After that, the data will be split into training and test data set. An ordinary least square regression will be applied to the training data set, with the constant and coefficient of the dependent variable both extracted and used to construct the spread. The spread will be tested for co-integration again through ADF method in order to confirm the cointegration relationship between both ETF log price training data. The Z-score is calculated afterwards, and the trading and stop-loss signals are generated based on the observed z-score.

4.1. Data

The two Gold Tracking ETFs are the SPDR Gold Shares (GLD-US) and iShares Gold Trust (IAU-US). The data used in this thesis are daily adjusted closing prices, downloaded from FactSet into an excel sheet. FactSet is a financial data and software company that provides flexible, open data and software solutions to professional investors throughout the world. They provide investors with access to financial data and analytics to facilitate their decision-making processes (FactSet, 2023).

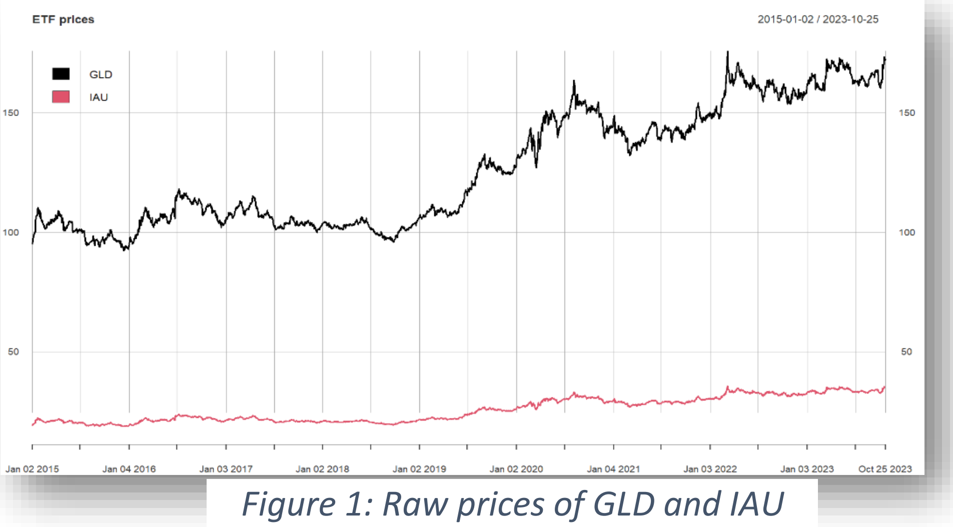


Figure 1: Raw prices of GLD and IAU

The data from FactSet ranges from 1st January 2015 to 25th October 2023, giving a total of 2219 observations.

The daily adjusted closed prices of (GLD-US) and (IAU-US) will be converted to log form as raw ETF prices are rarely stationary, before being tested for co-integration using the Johansen approach and Engle Granger Augmented Dickey Fuller Test (EG-ADF). Using both the ADF test and the Johansen test to test for co-integration gives more solid evidence of co-integration. And if both prices are co-integrated using both methods, it should eliminate any doubts of false positive results and prove that there is persistence in co-integration. The first 70% (1553 observations) of the data will then be set for the training period and the remaining 30% (666 observations) for the testing period. The EG-ADF test will be done on the spread of the training and testing data set to ensure co-integration exists both during the training and testing period.

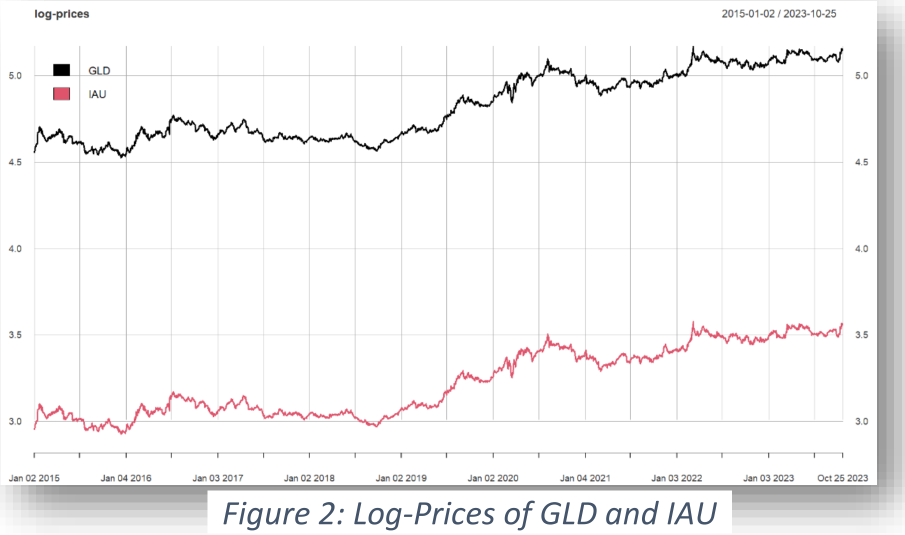


Figure 2: Log-Prices of GLD and IAU

Figure 2 gives a much clearer picture of how the log prices of both ETFs move in a similar direction, giving some cointegration indication.

4.2. Methodology

The methodology is based on theories that deal with testing for cointegration and stationarity using Johansen cointegration method and the ADF method, computing the spread, z-score, the trading strategy or rules and returns calculation.

4.2.1. Co-integration and Stationarity

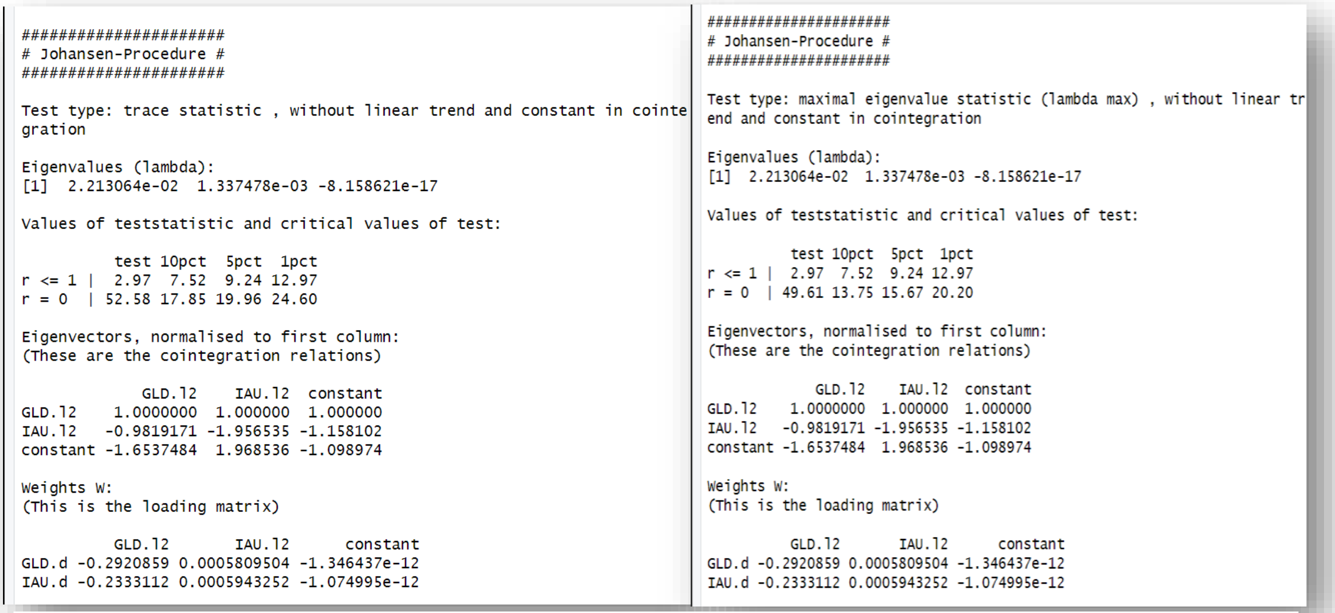
Finding potential pairs of assets whose prices move together is not that simple. If an asset price is said to be stationary, it means it will revert to the mean and therefore be suitable to a mean reverting trading strategy, even though most raw asset price series are not stationary. Never-the-less, it is possible to make a non-stationary time series stationary through co-integration, which refers to co-movement in asset prices. Two methods are useful for this thesis when it comes to cointegration: The Johansen approach and the Augmented Dickey Fuller (ADF) approach.

The Johansen approach to cointegration identifies long-run equilibrium between two prices. If a long-run equilibrium is found, any deviations are short-lived and are corrected by an adjustment in one or both prices (Vidyamurthy, 2004). The Johansen cointegration test involves estimating the rank of the cointegrating matrix (Π).

For two integrated time series of order one $(I(1))$ to be cointegrated, there must be a linear relationship between the two that is integrated of order zero $(I(0))$ (Caldeira and Moura, 2013). To put this in context, two ETFs are cointegrated when their prices follow a similar trend, which therefore means they should have a long-run constant relationship. The Johansen test provides statistics of two kinds; Trace and Eigen Statistic. Both produce very similar results when used.

These tests are likelihood-ratio tests, which assess a model's goodness of fit overall. The tests statistic indicates whether the rank (Π) = 0, with the null hypothesis stated as rank (Π) = 0, and the alternative test is written as 0 ≤ rank (Π) ≤ r, with r being the maximum possible cointegrating vectors and Π is the cointegrating matrix as mentioned above. The Johansen test measures the degree of cointegration between variables and makes it possible to rank the most cointegrated pairs (Asteriou and Hall, 2011).

The Johansen cointegration test results will be presented at this level because it was used on the entire data set. The results for both the Trace and Eigen statistic are presented on the figure 3:



```
######################
# Johansen-Procedure #
######################

Test type: trace statistic , without linear trend and constant in cointe
gration

Eigenvalues (lambda):
[1]  2.213064e-02  1.337478e-03 -8.158621e-17

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 1 |  2.97  7.52  9.24 12.97
r = 0  | 52.58 17.85 19.96 24.60

Eigenvectors, normalised to first column:
(These are the cointegration relations)

             GLD.l2    IAU.l2  constant
GLD.l2    1.0000000  1.000000  1.000000
IAU.l2   -0.9819171 -1.956535 -1.158102
constant -1.6537484  1.968536 -1.098974

Weights W:
(This is the loading matrix)

          GLD.l2       IAU.l2      constant
GLD.d -0.2920859 0.0005809504 -1.346437e-12
IAU.d -0.2333112 0.0005943252 -1.074995e-12
```

```
######################
# Johansen-Procedure #
######################

Test type: maximal eigenvalue statistic (lambda max) , without linear tr
end and constant in cointegration

Eigenvalues (lambda):
[1]  2.213064e-02  1.337478e-03 -8.158621e-17

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 1 |  2.97  7.52  9.24 12.97
r = 0  | 49.61 13.75 15.67 20.20

Eigenvectors, normalised to first column:
(These are the cointegration relations)

             GLD.l2    IAU.l2  constant
GLD.l2    1.0000000  1.000000  1.000000
IAU.l2   -0.9819171 -1.956535 -1.158102
constant -1.6537484  1.968536 -1.098974

Weights W:
(This is the loading matrix)

          GLD.l2       IAU.l2      constant
GLD.d -0.2920859 0.0005809504 -1.346437e-12
IAU.d -0.2333112 0.0005943252 -1.074995e-12
```

Figure 3: Johansen Cointegration Results with Trace (Left) and Eigen Test Statistic (Right)

The Johansen test for cointegration in figure 3 was performed on the entire data set. At rank r=0, the null hypothesis of no cointegration was rejected using both the trace and maximum eigen statistic. The test statistics were greater than the critical values at 10%, 5% and 1% significance level. There was no cointegration at r≤1, as the trace and Eigenvalue statistics value of 2.97 was less than the critical value at all the significance level.

Both tests conclude that there is at least 1 cointegrating relationship between GLD and IAU.

For the **ADF**, in the context of this thesis, let y represent logGLD and x be log IAU. If y_t and x_t are both log-prices of two assets, at any given moment in y where $y_t - yx_t$ is stationary, y and x can be said to be cointegrated. That is, y and x have a long-run equilibrium relationship that when disturbed, tend to adjust or return to that equilibrium (Engle and Granger, 1987). This relationship is built to be stable in the long run to avoid investment systems from incurring substantial losses (Do & Faff, 2010). Vidyamurthy (2004) used the co-integration technique to construct relationships between assets based on the co-integration concept laid down by Engle and Granger (1987).

The ordinary least square (OLS) cointegration equation is as follows:

$$y_t = \alpha + \beta x_t + \varepsilon_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where:

- α is a constant or intercept and β is the coefficient of x at time t, (represented by mu and gamma respectively in my R code)
- ε_t is a stationary zero-mean variable (not necessarily white noise), then y_t (logGLD) and x_t(logIAU) are co-integrated,
- $\alpha + \beta x_t$ represent the long-term equilibrium,
- $\varepsilon_t = y_t - \alpha - \beta x_t$ represent deviations from the equilibrium.

Thus, co-integration is tested through the Engle-Granger method using the Augmented Dickey Fuller test (ADF test). Selecting an appropriate number of lags when running the ADF test is important, and these lags are selected based on the Information Criterion. AIC yields optimal amount of information based on the minimum information criterion (Stock and Watson, 2015). The lag will be 2 for reasons stated in section 10. The ADF tests the null hypothesis of non-stationarity against the alternative hypothesis of stationarity. With a 5% (p value =0.05) statistical significance level, the null can be rejected (Stock and Watson, 2015).

Therefore, the first step of the Engle-Granger ADF test is to estimate the cointegration coefficient, β and the intercept α through ordinary least squares (OLS) method stated in equation (1) (Stock and Watson, 2015). The results can be interpreted as the premium for holding y_t over x_t (Vidyamurthy, 2004). Co-integration is confirmed with the ADF test for unit root conducted on the residuals extracted from equation (1) with the alternative hypothesis that $y_t \ and \ x_t$ are cointegrated at the 5% level of statistical significance (Stock and Watson, 2015), with the AIC used to select the appropriate lag length. The corresponding test statistic is obtained by estimating an autoregression of ΔY_t on its own lags and Y_{t-1} using OLS on the following ADF equation:

$$\Delta Y_t = \alpha + \beta t + (\emptyset - 1)Y_{t-1} + \cdots + \delta_{p-1}\Delta Y_{t-p-1} + \varepsilon_t$$

Where ΔY_t is the differenced time series, α is a constant, β the coefficient on a time trend, p the lag order of the autoregressive process and δ_{p-1} are the coefficients of lagged differences.

The above ADF equation is written as such so that a linear regression can be applied on ΔY_t against t and Y_{t-1}, making it possible to test if $(\emptyset - 1)$ is significantly different from zero. If $(\emptyset = 1)$, then it's a random walk process. If that is not the case, and $\emptyset \in (-1,1)$, then it is a stationary process.

4.2.2. The Spread

The Spread of a pair of assets indicates how the current relationship between both assets are different from its historical. Divergence from equilibrium is said to be significant depending on the distance to the mean, and the spread of a cointegrated asset pair is considered stationary and mean reverting (Vidyamurthy, 2004). The spread, ε_t of logGLD and logIAU is defined below;

$\varepsilon_t = \log(GLD_t) - \beta\log\,(IAU_t)$……………………………………………………………………………(2)

It should be noted that the spread relies on a regression process and does not include an intercept (α_0). Creating room for an intercept and regressing the log-price of GLD on the log-price of IAU yields the following equation,

$\log(GLD_t) = \alpha + \beta\log(IAU_t) + \varepsilon_t$…………………………………………………………………………(3)

The above equation can be rewritten as follows;

$\log(GLD_t) - \alpha - \beta\log(IAU_t) = \varepsilon_t$…………………………………………………………………………(4)

Where the ε_t, is the spread at point in time, t.

4.2.3. The Z-Score

The z-score measures the distance to the long-term mean in units of long-term standard deviation and helps in generating signals that will long and short the ETF pairs. When the Z-score deviates significantly from its historical mean, it suggests that the spread between the two assets is unusually high or low. This flaw can be exploited with the expectation that the spread will revert to its historical mean. (Caldeira and Moura,2013). The strategy in this thesis uses the z-score to generate signals that detect abnormal price deviations between the two ETF price time series. From a statistical point of view, the z-score can either be negative or positive values that are relative to the mean. An arbitrary Simple Moving Average (SMA) window size of 20 was chosen to help carve out clear market trends. The window size cannot be too small (e.g., 10) because it would react more quickly to short-term fluctuations but might be more sensitive to noise. A larger window size on the other hand (e.g., 50 or 200), would result in a smoother moving average but may lag behind changes in the underlying data (Murphy, 1999). Thus, a window size of 20 seems appropriate. The Z-score was computed using the difference between log prices, the moving average and standard deviation:

$$Z - score = \frac{\varepsilon_t - SMA_{20}}{SD_{20}}$$

ε_t is the value of the price spread at time t,
SMA_{20} is the moving average of window size 20,
SD_{20} is the standard deviation of window size 20.

In summary, the Z-score and moving averages are tools used in identifying deviations from the historical mean and make mean-reverting opportunities in pairs trading much more visible. They both provide the quantitative framework upon which the entry, exit, and stop-loss are built within the context of a pairs trading strategy.

4.2.4. Trading Strategy including transaction Cost

The trading strategy will be similar to that proposed by Caldeira and Moura (2011). Their strategy specified trading rules which indicated when exactly to engage and cut long and short positions. The trading here aims to take advantage of the mean-reverting behavior in the spread Z-score, entering positions when the spread deviates significantly from its historical mean and closing positions based on specified stop-loss conditions. It's a simple example of a mean-reversion trading strategy with risk management through stop-loss levels.

The trading strategy is based on the already calculated z-score and is detailed as
follows:

Thresholds:

- A long (buy) signal is generated when the Z-score is less than or equal to the long threshold (-1).
- A short (sell) signal is generated when the Z-score is greater than or equal to the short threshold (+1).

Stop-loss Levels:

- If in a long position and the Z-score turns positive or falls below the stop-loss level (-2.5), close the long position.
- If in a short position and the Z-score turns negative or rises above the stop-loss level (+2.5), close the short position.

In summary,

Thresholds:

Long (buy) signal: z-score < threshold long (-1)

Short (sell) signal: z-score > threshold long (1)

Stop-loss Levels:

Close long position: z-score ≥ 0 or z-score < stop_loss_long (-2.5)

Close short position: z-score ≤ 0 or z-score > stop_loss_short (2.5)

Opening or closing a short position means buying and selling both pairs simultaneously. If the spread is given as in equation 4, the trading rule will be to open a position when the z-score is 1 standard deviation thresholds from above or from bellow. If the z-score hits the -1 standard deviation threshold, it means that the pairs are below the long-run equilibrium value. This means buying GLD and selling IAU. If the z-score hits the 1 standard deviation threshold from above, the pair is overvalued and there should be a short-sell, implying selling GLD and buying IAU. The position is then closed as the z-score approaches zero or when it reaches ±2.5.

The trading volume initiated, as well as closed, depends enormously on the chosen threshold values. The lower the threshold value, the higher the number of open positions there would be, leading to higher trading costs. On the other hand, the higher the threshold value, the fewer the number of open positions for trade. This means that threshold values will be chosen and tested empirically rather than theoretically as proposed by Avellaneda and Lee (2008).

Transaction Cost: Given how difficult it is to have a precise and generally acceptable level or percentage for transaction costs, as it varies across different market conditions and different trading platforms, a 5% transaction cost was used. These costs will be standard costs attributable to trading such as slippage costs, brokerage fees, rental costs, and other expenses that come with using a trading platform. The transaction cost is subtracted from the current position based on the sign of the change in position. This means that transaction costs are incurred whenever there is a change in the trading position, reflecting a per-trade transaction cost. In other words, transaction costs are subtracted from the current position whenever a change in position occurs.

4.2.5. Returns Computation

The cumulative returns will be used to assess the trading performance of overtime. It will be calculated as the aggregate return gained or lost on the trade over the trading period. If the simple one period return is given by (Tsay, 2010):

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Where R_t is the return at time t, P_t is the price at time t, and P_{t-1} is the price at time t-1.

Then holding an asset for k periods between dates (t-k) and t gives a k-period simple gross return given by the formula:

$$1 + R_t[k] = \frac{P_t}{P_{t-k}} = \frac{P_t}{P_{t-1}} \times \frac{P_{t-1}}{P_{t-2}} \times \cdots \times \frac{P_{t-k+1}}{P_{t-l}}$$

$$= (1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1})$$

$$= \prod_{j=0}^{k-1} (1 + R_{t-j})$$

5. Backtesting

The testing period is used for back-testing the strategy, where the parameters computed in the training period are used to run the back-test. The testing period runs from 05/03/2021 to 25/10/2023, constituting the remaining 30% (666 observations) from the total data set.

Pair's spread

The pair's spread, computed as log difference of the price series will be generated as input for the model. A cointegration test using the ADF method as described in section 4.2.1 will be applied to the spread to test for persistence in cointegration during the testing period.

Returns

The returns will equally be computed as in section 4.2.5

6. Performance Measures

Performance measures will evaluate the strategy performance during the training and testing period. This will be measures like the annualized return, annualized volatility, Sharpe ratio and the maximum drawdown.

6.1. Annualized Returns

The compounding effect of returns over time will be calculated. The total returns of the strategy will be divided by the number of trading days and multiplied by 100 to have the results in percentage.

6.2. Annualized Volatility

Volatility was used to evaluate the riskiness of strategy by measuring the dispersion of the returns around the average (Mateus, 2022). Volatility is given by the formula below:

$$\sigma = \sqrt{Var(R)}$$

Where, σ is volatility or standard deviation, Var is variance and is given by $Var(R) = \sum_{i=1}^{n} pi(R_i - E[R])^2$ and R represents the returns.

6.3. Sharpe Ratio

The sharpe ratio was introduced in 1966 and has been widely used in trading to estimate the reward-to-variability ratio for certain portfolios ever since (Sharpe, 1994). The formula below is used for the sharpe ratio:

$$Sharpe\ ratio = \frac{(R_p - R_f)}{\sigma_p}$$

Where R_p is annualized returns, R_f is risk-free rate, and σ_p is annualized standard deviation of the returns.

As stated in the assumptions, risk free rate is zero, giving the sharpe ratio as:

$$Sharpe\ ratio = \frac{(R_p)}{\sigma_p}$$

6.4. Maximum Drawdown

The Max Drawdown measures how much an investment or trade has declined from its peak during a given period (Caldeira and Moura, 2013). In other words, it shows the maximum percentage loss from a strategy's highest point to its lowest point over a specific period of time, giving indication of potential risk losses. It is given by the formula below:

$$Max\ Drawdown = \frac{Trough_t - Peak_t}{Peak_t}$$

7. Results and Analysis

The results from the methodology will be presented in two parts for simplicity: the training period and testing or backtesting period.

7.1. Training Period

As mentioned before, the training period consists of the first 70% (1553 observations) of the total 2219 gold tracking ETFs observations, GLD and IAU. The focus here is to generate parameters that will be used to backtest the trading strategy and evaluate its performance with new data.

The intercept and the coefficient of IAU were found to be 0.985 and 1.64 respectively after running the OLS regression. The regression equation following equation 1 was as follows:

$$Log(GLD)_t = 0.985 + 1.64Log(IAU)_t + \varepsilon_t$$

The two values were then used to construct the spread as seen below:

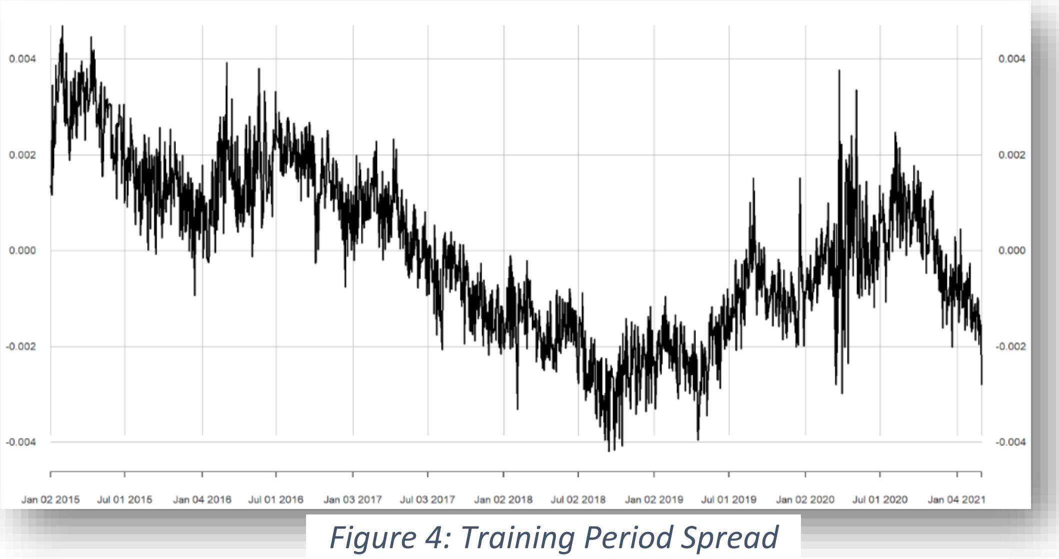
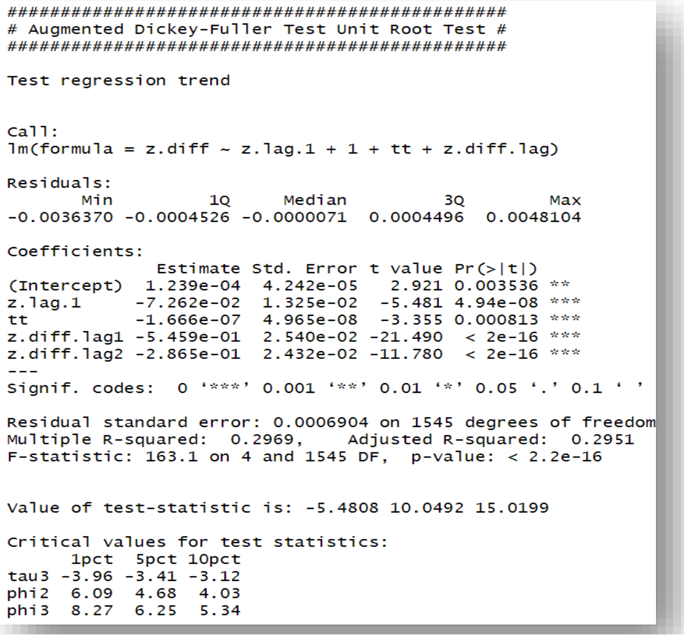


Figure 4: Training Period Spread

The spread was tested for cointegration using the ADF method described in section 4.2.1 and the results were as follows:



```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0036370 -0.0004526 -0.0000071  0.0004496  0.0048104 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.239e-04  4.242e-05   2.921 0.003536 ** 
z.lag.1     -7.262e-02  1.325e-02  -5.481 4.94e-08 ***
tt          -1.666e-07  4.965e-08  -3.355 0.000813 ***
z.diff.lag1 -5.459e-01  2.540e-02 -21.490  < 2e-16 ***
z.diff.lag2 -2.865e-01  2.432e-02 -11.780  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.0006904 on 1545 degrees of freedom
Multiple R-squared:  0.2969,    Adjusted R-squared:  0.2951 
F-statistic: 163.1 on 4 and 1545 DF,  p-value: < 2.2e-16


Value of test-statistic is: -5.4808 10.0492 15.0199 

Critical values for test statistics: 
      1pct  5pct 10pct
tau3 -3.96 -3.41 -3.12
phi2  6.09  4.68  4.03
phi3  8.27  6.25  5.34
```

Figure 5: ADF test results on the training data spread

With the spread constructed as in equation (4), all the test statistics rejected the null of a unit root at the 5% level of significance as seen in figure 5, suggesting GLD and IAU were cointegrated during the training period. The p-value of less than 0.05 (2.2e-16) also suggested co-integration.

The Z-score was computed using the spread, a window size of 20 for the moving average and standard deviation. This resulted to a more mean-reverting z-score spread as seen below:

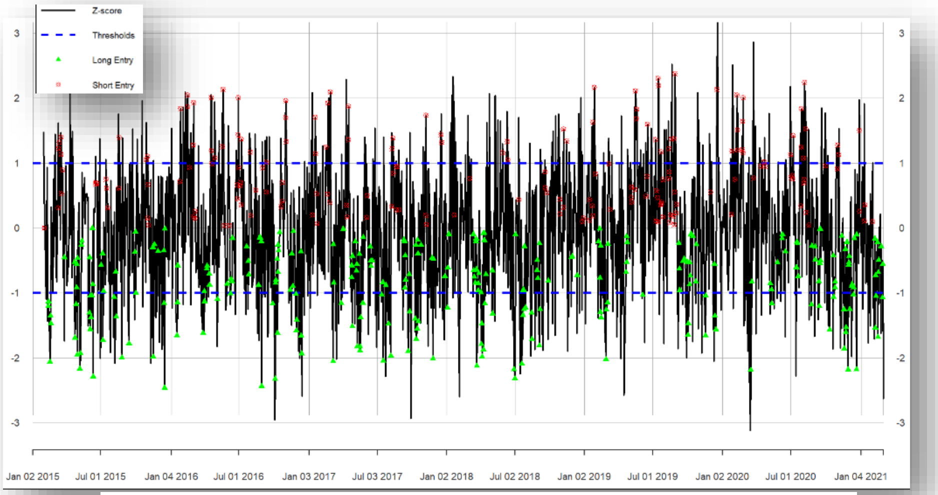


Figure 6: Training Period Z-Score with Trading Signals

Figure 6 equally included the long and short thresholds, as well as the long and short entry levels. The trading signal was initiated at 0, which is a no position. This means that the generated code makes the signal to check conditions for entering long positions at -1 thresholds or short positions at +1 thresholds. Thus, If the previous position was a long position (green points on figure 6), the signal checks conditions for exiting the long position which was specified to be either when z-score is non-positive or reaches the stop-loss level of -2.5. Conversely, If the previous position was a short position (red dots on figure 6), the signal checks conditions for exiting the short position stated as either z-score is non-negative or reaches the stop-loss level of +2.5.
A transaction cost of 5% was deducted from the signal value when a trade was made.

The traded returns were calculated after executing the trading rules. The traded returns integrated the trading decisions into the returns, simulating the performance of the strategy. If the strategy was in a long position, the return was multiplied by 1; if it was in a short position, the return was multiplied by -1; if there was no position, the return was multiplied by 0.

The cumulative profits and losses were then calculated as the cumulative sum of the returns at each time point.

Table 1 below summarizes the annual statistical performance of the pairs trading strategy. The table includes the annual returns and volatility, sharpe ratio, maximum drawdown as well as the number of trades.

Performance Measure	Percentage/Ratio
Annualized Return	5.11%
Annualized Volatility	0.88%
Sharpe Ratio	5.82
Max Drawdown	0%
Number of trades	907

Table 1: Strategy Performance

The strategy yielded annual positive returns, with a 0% max drawdown. The sharpe ratio was high and stood at 5.82, which is generally considered to be good and suggested the strategy can attract risk adjusted returns. A graphical presentation of how the returns evolved throughout the training period is seen in Figures 8 and 9.

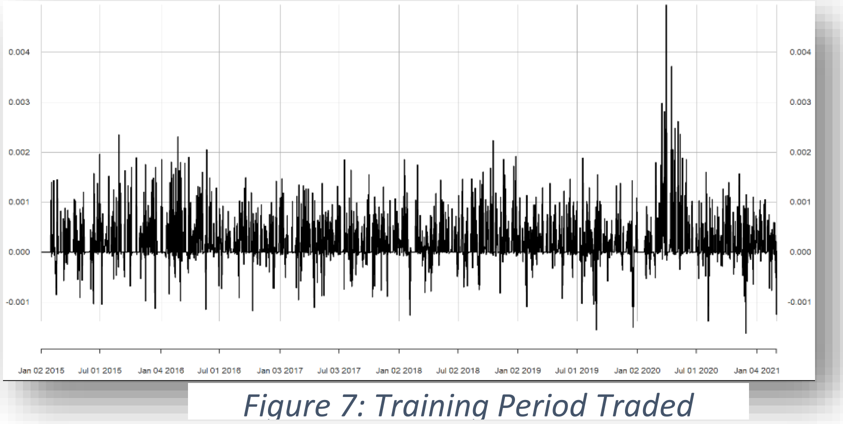


Figure 7: Training Period Traded

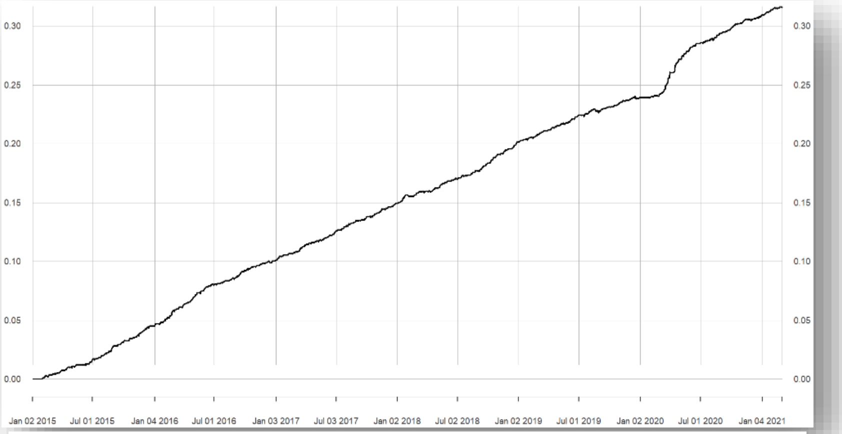
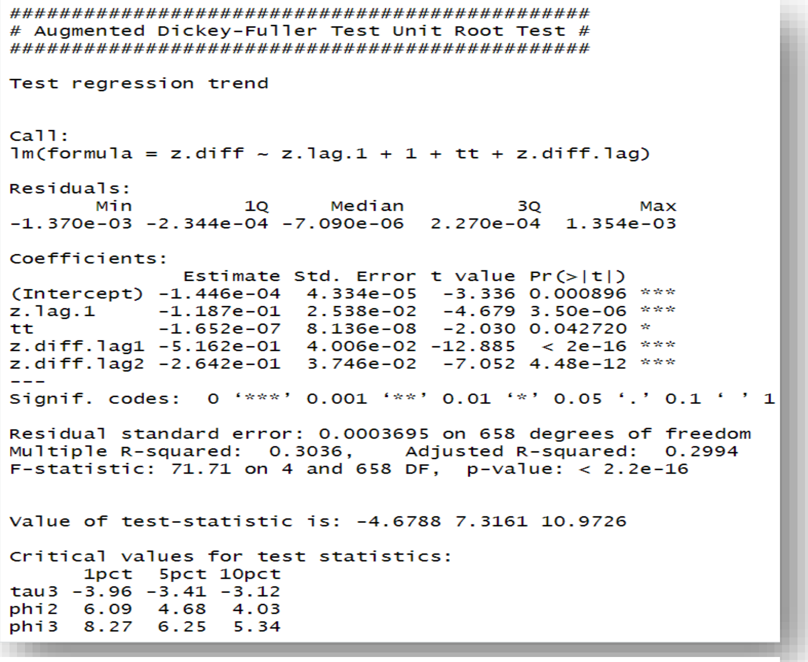


Figure 8: Cumulative Profits and Losses of the Training

The trading strategy had its highest returns in the early part of 2020 as indicated by the high spikes in figure 7, which results in the bulge in cumulative returns during the same period on figure 8.

7.2. Testing/Backtesting Period

The backtesting period evaluates how well the strategy performs when introduced to new data. The constant (0.985) and the IAU coefficient (1.64) already calculated during the training period are used as the input parameters. Based on this, the strategy was back-tested on the remaining 30% (666 observations) of the total GLD and IAU 2219 observations. The spread was built as in equation 4 using only the testing data and tested for cointegration using the ADF method. The ADF results were:



```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
       Min         1Q     Median         3Q        Max
-1.370e-03 -2.344e-04 -7.090e-06  2.270e-04  1.354e-03

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.446e-04  4.334e-05  -3.336 0.000896 ***
z.lag.1     -1.187e-01  2.538e-02  -4.679 3.50e-06 ***
tt          -1.652e-07  8.136e-08  -2.030 0.042720 *
z.diff.lag1 -5.162e-01  4.006e-02 -12.885  < 2e-16 ***
z.diff.lag2 -2.642e-01  3.746e-02  -7.052 4.48e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0003695 on 658 degrees of freedom
Multiple R-squared:  0.3036,    Adjusted R-squared:  0.2994
F-statistic: 71.71 on 4 and 658 DF,  p-value: < 2.2e-16


Value of test-statistic is: -4.6788 7.3161 10.9726

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.96 -3.41 -3.12
phi2  6.09  4.68  4.03
phi3  8.27  6.25  5.34
```

Figure 9: ADF test on the Test data Spread

The p-value of 2.2e-16 was less the 0.05 and the test statistics ones again rejected the null hypothesis for unit at 5% levels of significance as seen in figure 9, suggesting GLD and IAU were co-integrated during the testing period as well. This confirmed that there was persistence in co-integration.

The Z-score was then computed using the spread of the test data with the same methodology as in equation 4 and with a window size of 20 for the moving average and standard deviation. The mean-reverting z-score was obtained and plotted as seen below:

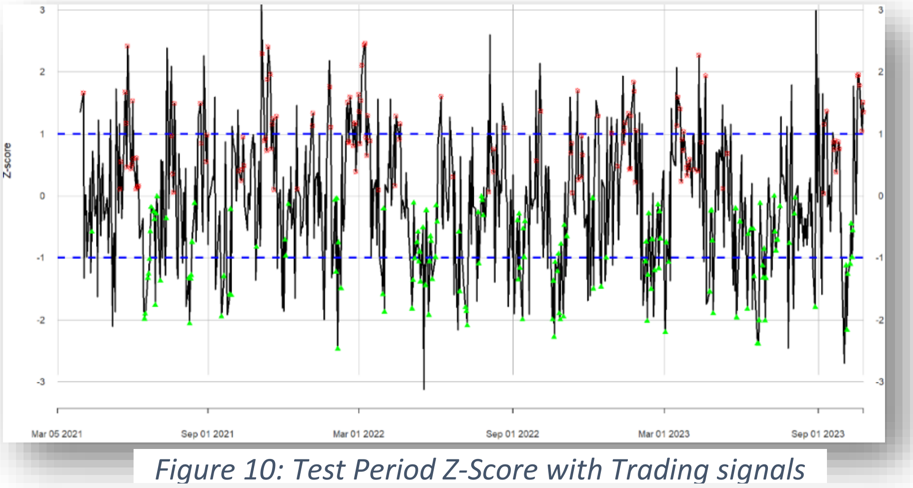


Figure 10: Test Period Z-Score with Trading signals

As already mentioned, the backtesting period evaluates the strategy performance on new data. This means that Figure 10 includes the same parameters from the training period. That is, the same trading rules, trading signals and transaction costs percentage stated in section 4.2.4.

Table 2 summarizes the annual statistical performance of the pairs trading strategy.

Performance Measure	Percentage/Ratio
Annualized Return	2.03%
Annualized Volatility	0.52%
Sharpe Ratio	3.9
Max Drawdown	0.03%
Number of trades	318

Table 2: Strategy Performance on New data

The 2.03% annualized return, though positive, was 3.08% short compared to that of the training period. The 0.03% max drawdown indicated the strategy had potential risk as it had dropped from its peak to its trough compared to the 0% obtained for the training period. The sharpe ratio at 3.9 was also smaller than that of the training period, indicating the testing period had less annual excess returns in comparison to the risk taken. The 0.52% volatility level was lower than the 0.88% for the training period, signifying the returns fluctuated less on average during the testing period and were more stable.

The returns and cumulative profits and losses can be visualized below:

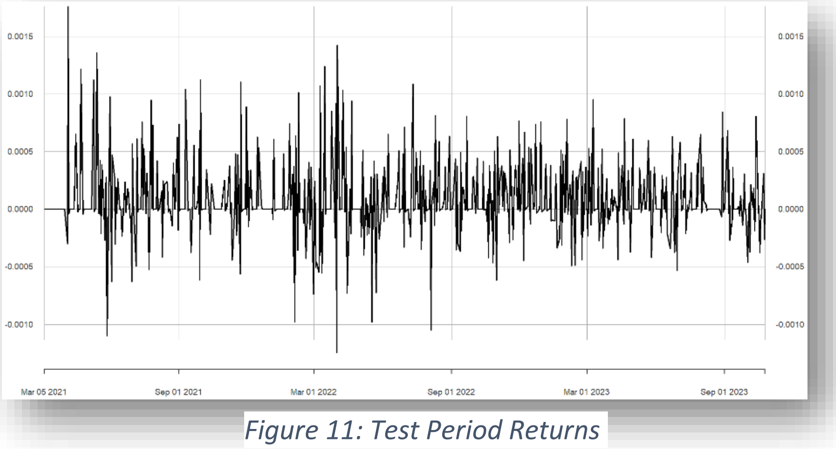


Figure 11: Test Period Returns

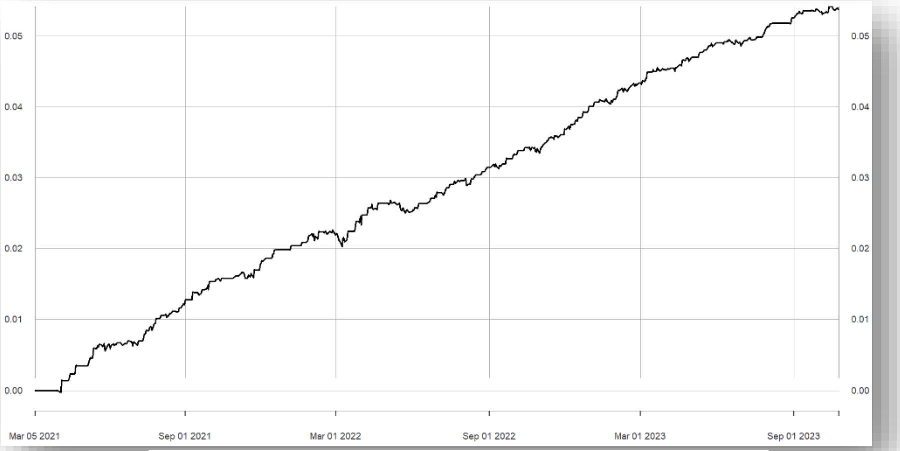


Figure 12: Test Period Cumulative Profits and Losses

8. Sensitivity Analysis

Sensitivity analysis will help assess how changing certain important parameters will impact the results of the strategy. This will determine the robustness of the backtesting period results presented in table 2 and should provide an understanding of how sensitive they are to variations in key factors. The analysis will be applied on parameters such as the window size for the moving average, the entry and exit thresholds. Only the testing period results will be compared to those of the sensitivity analysis as it is the most important.

In the first and second scenarios, only the moving average will be decreased to a window size of 10 and equally increased to a window size of 30 respectively. This means that while the MAs are being varied, the threshold as well as other parameters will not change.

The third scenario involves decreasing the short threshold value to 0.7 and long threshold to -0.7, while the fourth scenario will have an increased short threshold value to 1.2 and the long threshold to -1.2. In these scenarios, only the thresholds are varied and tested while the other factors are the same as in the original strategy. The results are as tabled below:

	Scenario 1 (10 MA)	Scenario 2 (30 MA)	Scenario 3 (±0.7 threshold)	Scenario 4 (±1.2 threshold)
PERFOMANCE MEASURES				
Annualized Return	2.49%	1.88%	2.25%	1.96%
Annualized Volatility	0.48%	0.49%	0.54%	0.51%
Sharpe Ratio	5.17	3.81	4.15	3.86
Max Drawdown	0.04%	0%	0.03%	0.03%
Number of trades	390	280	352	292

Table 3: Sensitivity Analysis Results

The overall results suggest the pairs trading strategy is rather robust as there were no large deviations in the scenario performances compared to those of the backtesting period. Scenario 2 and 4 underperforms when their annualized returns, sharpe ratio and number of trades are compared to those of the testing period. But the gaps are by small margins. Returns are less volatile for both scenarios in relation to the testing period but again the margins are small percentages.

The first and third scenarios perform rather well with higher annualized returns, sharpe ratios and a higher number of trades. But again, the differences are by small percentages and numbers when compared to testing period results. The returns might be less volatile in scenario 1, but the max drawdown is larger than in the testing period. Scenario 3 is quite opposite, with returns being more volatile but the maximum drawdown comes to be the same as that in the testing period.

It is thus fair to conclude looking at the scenario results that the differences in the scenarios' performance results and those of the backtesting period are not by large margins. In fact, the margins are relatively small, and very small in some of the cases. This therefore Implies the strategy can be assessed to be robust and stable.

9. Assumptions

Firstly, the thesis assumes a lag order of 2 throughout as the lag selection code kept repeating the maximum specified lag in all the information criteria. Thus, 2 was selected because it is the smallest lag that can be chosen to test for co-integration. I equally tested lag of 3 and 4 and had cointegration but decided to stick with 2 as it is widely used to test for cointegration in other pair trading articles that do not specify lag selection criteria.

secondly, the analysis would assume no leverage but will assume a transaction costs of 5%, for reasons being that the exact levels of leverage and transaction costs are challenging to estimate and vary across different market conditions and different trading platforms.

Also, cumulative P&L is assumed to start from zero as no initial capital was injected into the trading.

Furthermore, the sharpe ratio is calculated without a risk-free rate, implying the risk-free rate R_f is zero (0) when calculating the sharpe ratio.

10. Limitations

For the limitations, the pairs trading strategy proposed in this thesis produces good performance for both the training and backtesting periods. However, the framework still has limitations and room for improvements.

The first limitation is on the data frequency. This study is based on daily data, and it is known in an efficient market that prices react quickly to sudden changes in information. Trading on daily data prices may not capture market information effectively. The same research framework and backtesting can be done at higher frequencies, such as hourly data frequency.

The second limitation is the use of z-score and simple moving average to filter the spread and the direction of price movement. More rubost methods can explore other filters like the Kalman filter to produce better spreads, which could results to better performances.

Also, there is the possibility of overfitting or underfitting as the window sizes and range of certain parameters were predefined, which could be considered biased. This means the optimization results may not be a true reflection of the actual optimized parameters, as it is not certain that the optimized parameters of the training period will be thesame as those of the testing period.

11. Conclusion

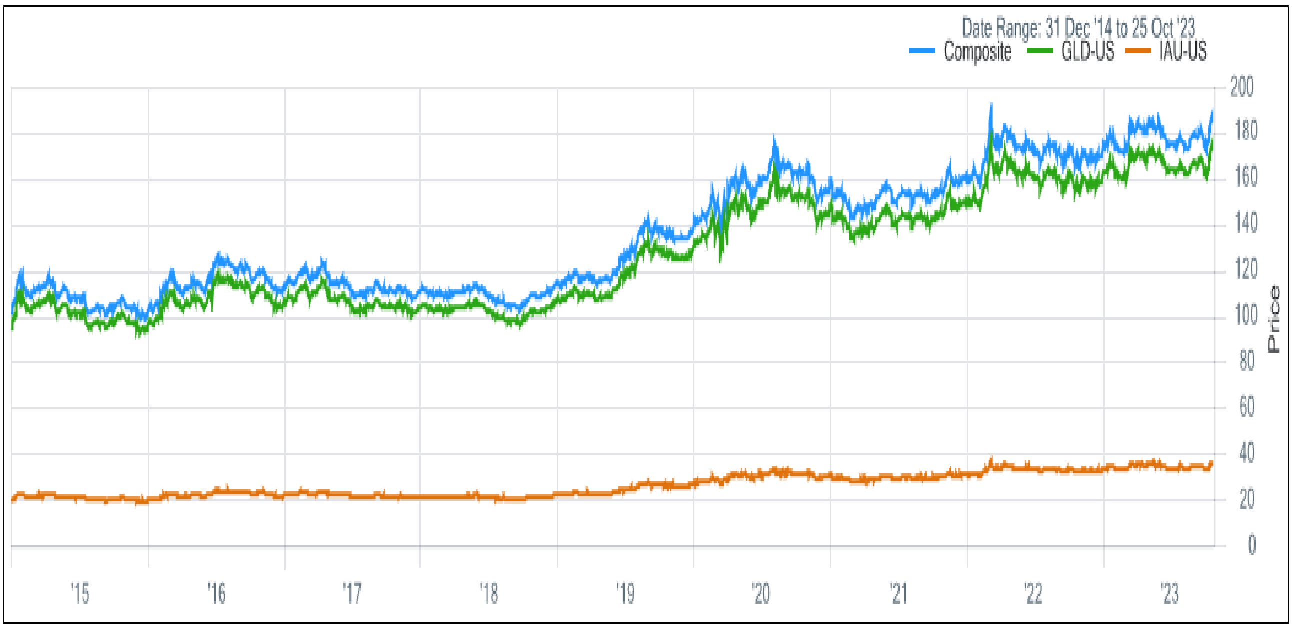
A statistical arbitrage pairs trading strategy using 2 gold tracking ETFs was proposed in this thesis, through the implementation of a cointegration approach. The data ranged from January 2015 to 25th October 2023, from which 70% was used for the training period and the remaining 30% for the testing period. Cointegration tests were applied at several levels and periods to ensure there was persistence in cointegration between the pair, which was found to be the case. The spread was calculated and normalized using the z-score and a simple moving average which helped to identify trends and mean reversion during the training and testing periods.

The annualized profits were 5.11% and 2.03% for the training and testing periods respectively, with very low risk levels as shown by the low volatilities, and a low maximum drawdown level as well. The sharpe ratios were equally high for both periods, which was good. One can however question the rather attractive results or perhaps the efficient market hypothesis. If the available information in the market were incorporated into the prices, such positive results should not be possible. Credit should not be taken away from the strategy which made sure losses were kept very low with the implementation of 2 stop-loss levels, thus the high profits.

As already mentioned, it would be intriguing to see what the results would look like if more sophisticated and clinical methods like hurst exponentials and Kalman filters were used. The hurst exponent has been documented to produce better trading pairs compared to the much older and traditional distance and correlation methods of selecting pairs for trading purposes. The hurst exponent uses mean reversion and correlation and the pairs with low Hurst exponent are selected with the idea that a low hurst exponent means co-movement of pairs and more mean reversion (Ramos-Requena, Trinidad-Segovia and Sánchez-Granero, 2017). The Kalman filter is known to produce smoother mean reversion as it filters noise and enhances the estimation of the hedge ratio for asset price series (Yang, Huang and Chen, 2023). Nevertheless, the attractive result establishes cointegration to still be an important tool in pairs trading that yields profits.

12. Appendices

12.1. Appendix A - SPDR Gold Shares (GLD-US) and iShares Gold Trust (IAU-US) price from FactSet



12.2. Appendix B – Backtesting period Spread.

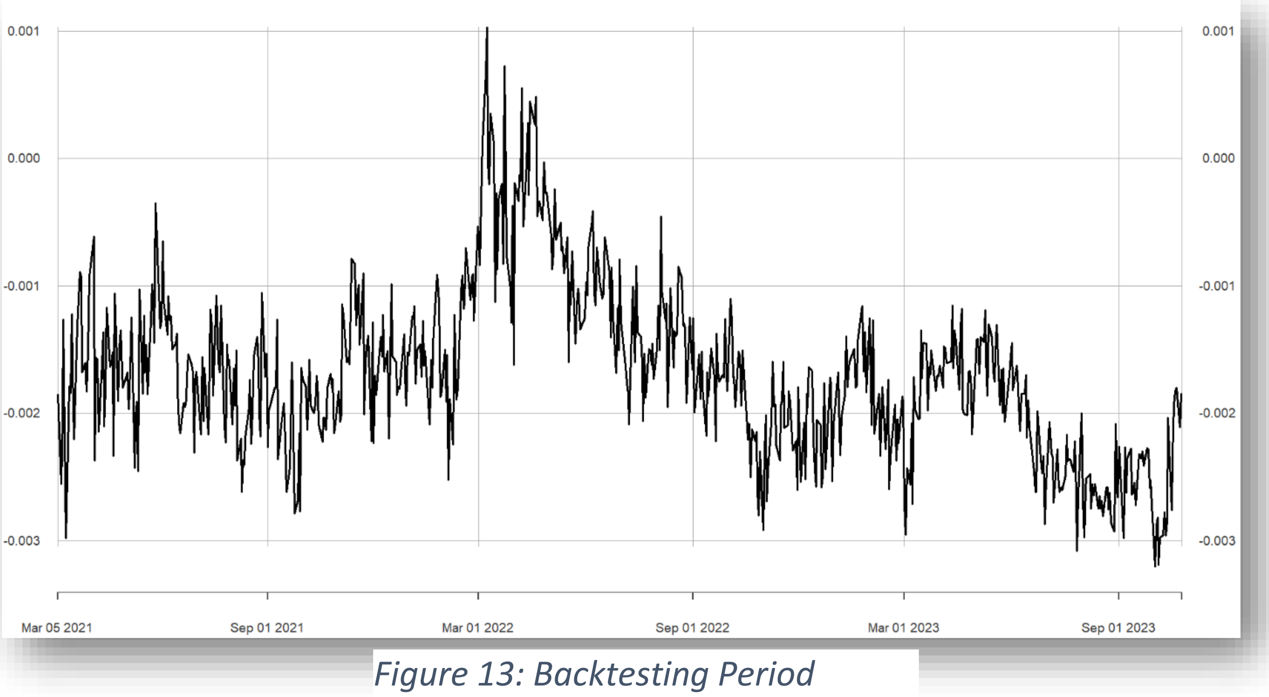


Figure 13: Backtesting Period

12.3. Appendix C - Benchmark

The London Bullion Market Association, LBMA Gold PM (GOLD-FDS); can be used as an appropriate benchmark. It is a global benchmark for gold and silver delivered in London and is regulated by ICE Benchmark Administration Limited (IBA) (LBMA., 2024). Data will be collected from FactSet, with the same start and end period as the backtesting period, which is from the 5th of March 2021 to 25th October 2023. Its performance will be compared to that of the backtesting period.

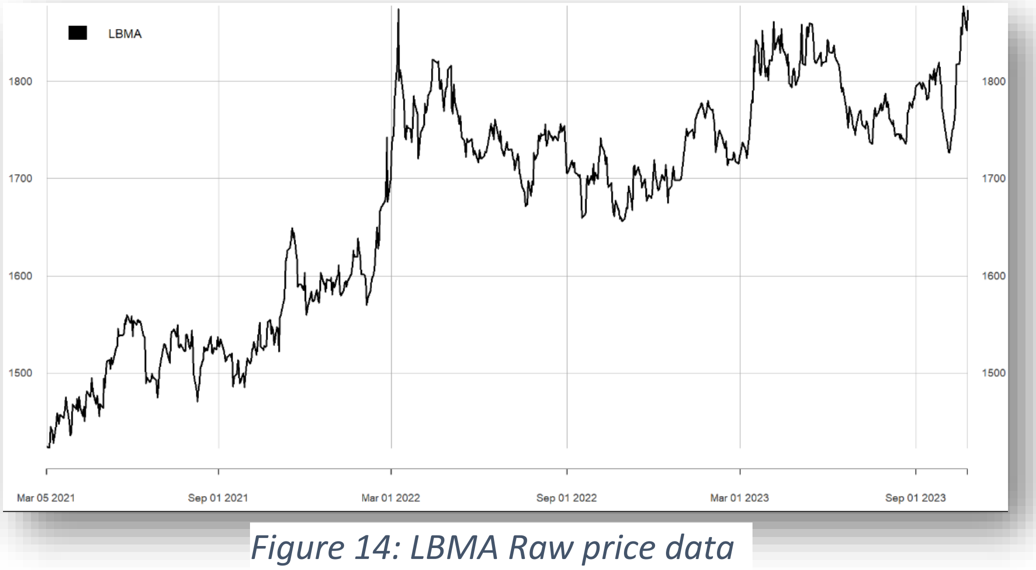


Figure 14: LBMA Raw price data

	Bactesting Period	Benchmark
PERFOMANCE MEASURES		
Annualized Return	2.03%	1.41%
Annualized Volatility	0.52%	1.18%
Sharpe Ratio	3.9	0.78
Max Drawdown	0.03%	1.64%

Table 4: Backtesting Period Performance Vs Benchmark Performance

It is quite clear from the above table that the strategy proposed in this thesis produces superio performance when compared to those of the LBMA benchmark. The strategy produces superior annualized returns which are less volatile than than the benchmark's. It equally has a promising sharpe ratio with a very small maximum drawdown percentage, indicating it produces more risk adjusted returns and has a smaller potential risk as its returns do not drop too deep down from its peak point. The benchmark on the other hand has a smaller sharpe ratio, meaning less attractive risk adjusted returns and the high maximum drawdown shows LBMA stands a higher risk as its returns drop highest from its peak to its trough.

13. Reference list

ACKAERT Lucy F., TIAN Yisong S. (2008). *Arbitrage, liquidity, and the valuation of Exchange-Traded Funds*, New York University Salomon Center, Financial Markets, Institutions & Instruments, V. 17, No. 5, December 2008. Published by Wiley Periodicals, Inc.

Asteriou, D. and Hall, S.G. (2011). *Applied Econometrics*. Palgrave Macmillan.

Avellaneda, Marco and Lee, Jeong-Hyun (2008). *Statistical Arbitrage in the U.S. Equities Market*. Available at SSRN: https://ssrn.com/abstract=1153505 or http://dx.doi.org/10.2139/ssrn.1153505.

Caldeira, João and Moura, Guilherme Valle, (2013). *Selection of a Portfolio of Pairs Based on Cointegration: A Statistical Arbitrage Strategy*. Available at SSRN: https://ssrn.com/abstract=2196391 or http://dx.doi.org/10.2139/ssrn.2196391

DO Binh, FAFF Robert (2012*). Are pairs trading robust to trading costs*, The journal of financial research, Vol. XXXV, No. 2, pp 261-287. https://doi.org/10.1111/j.1475-6803.2012.01317.x.

Do, Binh & Faff, Robert & Hamza, Kais. (2006). *A New Approach to Modeling and Estimation for Pairs Trading.*

Ehrman, Douglas S. (2006). *The Handbook of Pairs Trading.* John Wiley & Sons.

Engelberg, J., Gao, P., & Jagannathan, R. (2009). *An Anatomy of Pairs Trading: The Role of Idiosyncratic News, Common Information and Liquidity*. SSRN: https://ssrn.com/abstract=1330689 or https://doi.org/10.2139/ssrn.1330689

FactSet. (2023). *Our company*. Available at: https://www.factset.com/our-company.

Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). *Pairs Trading: Performance of a Relative-Value Arbitrage Rule*. Review of Financial Studies, 19(3), 797–827. https://doi.org/10.1093/rfs/hhj020

Gatev, E., Goetzmann, W., & Rouwenhorst, K. G. (1999). *Pairs Trading: Performance of a Relative Value Arbitrage Rule*. NBER Working Paper No. 7032, 797–827.https://doi.org/10.3386/w7032

Giovanni Montana, Kostas Triantafyllopoulos, Theodoros Tsagaris (2009). *Flexible least squares for temporal data mining and statistical arbitrage, Expert Systems with Applications.* ISSN 0957-4174: https://doi.org/10.1016/j.eswa.2008.01.062.

Huck, N., & Afawubo, K. (2015). *Pairs trading and selection methods: is cointegration superior?* Applied Economics, 47(6), 599–613. https://doi.org/10.1080/00036846.2014.975417

LBMA. (2024). *The Independent Precious Metals Authority*. Available at: https://www.lbma.org.uk/.

Mateus, C. (2022). Session 1 presentation: *Return and Risk*, [Topics in Asset Management.] Aalborg University.

Murphy, J.J. (1999). *Technical analysis of the financial markets: a comprehensive guide to trading methods and applications*. New York: New York Institute of Finance.

Pedersen, L. H. (2015). *Efficiently inefficient: how smart money invests and market prices are determined.* Princeton, United Kingdom: Princeton University Press.

Petajisto, Antti, (2016). *Inefficiencies in the Pricing of Exchange-Traded Funds.* Available at SSRN: https://ssrn.com/abstract=2000336 or http://dx.doi.org/10.2139/ssrn.1572907

Rad, H., Low, R.K.Y. and Faff, R.W. (2015). *The profitability of pairs trading strategies: distance, cointegration and copula methods,* Quantitative Finance. https://doi.org/10.1080/14697688.2016.1164337.

Ramos-Requena, J.P., Trinidad-Segovia, J.E. and Sánchez-Granero, M.A. (2017). *Introducing Hurst exponent in pair trading*. doi:https://doi.org/10.1016/j.physa.2017.06.032.

Schizas, Panagiotis and Thomakos, Dimitrios D. and Wang Tao (2011). *Pairs Trading on International ETFs.* Available at SSRN: https://ssrn.com/abstract=1958546 or http://dx.doi.org/10.2139/ssrn.1958546

Sharpe, W.F. (1994). *The Sharpe Ratio.* The Journal of Portfolio Management. doi: 10.390501/jpm.1994.409501.

SMITH Todd R., XU Xun (2017). *A good pair - Alternative pairs-trading strategies*, Swiss society for Financial Market Research. https://doi.org/10.1007/s11408-016-0280-x

Thorp, E.O. (2003). *A Perspective on Quantitative Finance: Models for Beating the Market.*

Tsay, Ruey S. (2010). *Analysis of Financial Time Series*. New Jersey: John Wiley & Sons.

Vidyamurthy, Ganapathy. (2004). *Pairs Trading - Quantitative Methods and Analysis.* John Wiley & Sons.

Yang, Shuo and Huang, Ke and Chen, Yao-wen, (2023). *Research on Hierarchical Pair Trading Strategy Based on Machine Learning and Kalman Filter*. Available at SSRN: https://ssrn.com/abstract=4590815 or http://dx.doi.org/10.2139/ssrn.4590815.