EXPLORING STATISTICAL ARBITRAGE IN THE NAIROBI SECURITIES EXCHANGE

Stanley Sayianka

2022-10-30

INTRODUCTION

Statistical arbitrage is a trading strategy born in the late 1980s as a proprietary strategy by the quant group under Nunzio Tartaglia in Morgan Stanley. The aim of the strategy is to take advantage of a mis-pricing between two or more assets which exhibit co-movement(commonly referred to as: *cointegration*) in their share prices, by assuming a LONG position in one (or several) assets which are thought to be under-valued and a SHORT position in the other(s) which are though to be over-valued, betting that the asset prices will converge to their "*equilibrium"* value thus locking in profit. This similar historical price movement is usually due to some fundamental reason such as: assets which have the same risk-factor exposures i.e. stocks belonging to the same industry, and thus they have the same dynamics in their share price movements.

The basic intuition governing statistical arbitrage is the fundamental principle of trading: *"Buy low, Sell high"*. However trivial this principle sounds, it would require that a trader is able to determine the intrinsic value of an asset in order to determine if the asset is over-valued or under-valued and thus trade accordingly. Pricing securities in this manner would require asset pricing models.

Pricing models in finance could be divided into relative pricing models and absolute pricing models. Absolute pricing values securities from their fundamentals such as Earnings, fundamental ratios and most commonly discounted value of the future dividend yields. This is common among *value* investors. Relative pricing on the other hand stipulates that two securities which could be substitutes for each other should trade for the same price(without explicitly stating what that said price is). This is commonly referred to as LOP: *The Law of One Price*. Ingersoll(1987) states the Law of One Price as the ***"proposition . . . that two investments with the same payoff in every state of nature must have the same current value."***

In the statistical arbitrage framework, an asset is priced relative to its peers such that: assets which exhibit similar risk factor exposures, are likely to have similar price dynamics, and thus any deviation of one, or a set of assets from this price dynamics would be interpreted as a relative mis-pricing which is likely to be corrected. This gives the reason why such arbitrage strategies are also referred to as *Relative-*

value arbitrage strategies.

The divergence in a set or pair of assets is quantified using a *spread* time series for the pair or set of assets. This spread time series quantifies the magnitude of divergence enabling an analyst to determine the extent of mis-pricing that has occurred and thus trade accordingly. Several methods exist in literature for computing the divergence or spread such as: using distance-based methods, ratio-based methods, the minimum variance portfolio method, the Ordinary Least Squares method, using the Total Least Squares method, Kalman filtering, stochastic control approach among others.

This study aims to explore the statistical arbitrage strategy on stocks listed in the Nairobi Securities Exchange. The framework for implementing the strategy will be as follows:

- Pairs search and selection: Identify possible pairs of assets which have co-moving price dynamics
- Spread modelling: Construct and model the spread using the several methods listed.
- Trading Rules: Come up with optimal trading rules such as maximum holding periods, optimal entry and exits, stop losses etc.

The questions we seek to answer in this study are:

1. Are there any cointegrated pairs of stocks listed in the Nairobi Securities Exchange for which the co-movement has been shown to be consistent historically, both in the price dynamics and in their risk factor exposure profiles ?
2. Among the spread modelling techniques used, which is the best performing, in terms of consistency, and portfolio returns ?
3. What are the optimal rules for trading under the statistical arbitrage framework ?

DATA & METHODOLOGY

Data

The data used in this study was fetched from WSJ(Wall Street Journal) who is a verified data vendor. The data was available for each listed stock from its inception into the exchange to date, and provides the OHLCV format common for financial datasets. Data on the factors used such as oil spot prices(BRENT OIL) and gold spot prices(XAUUSD) were fetched from yahoo finance, for the period under analysis. The data on the market prices(NSE All Share Index) were fetched from Investing.com for the period under investigation. Data on inflation rates, currency conversion rates(USDKES) and treasury bill rates were downloaded from the Central Bank of Kenya website.

Methodology

The following methodology was followed through this study in line with the framework suggested above for statistical arbitrage:

We design a universe of stocks from which to test for suitable pairs. This is especially suitable since there is a large number of stocks to analyze and the number of pairs to test grows exponentially. For this we choose only the constituents of:

- FTSE NSE 25 - A portfolio of market capitalization weighted index of top 25 companies listed on NSE. This also is useful, since we are assured that the liquidity the 25 companies trade at is acceptable as compared to other minority companies which may pose a liquidity risk in trading pairs.

By choosing from this basket we eliminate liquidity risk, since statistical arbitrage involves frequent trading, then being able to liquidate positions instantly when certain conditions are met is important, and thus selecting most largest traded assets in the Nairobi Securities Exchange helps in affirming that there is little-to-no liquidity risk. (Gatev, Goetzmann, & Rouwenhorst, 1998), (Krauss, Do, & Huck, 2017), (Stübinger, 2019) also use the same approach when selecting the universe of stocks to test for pairs cointegration.

We construct a cointegration test by choosing a suitable back test range over which to search for cointegrated pairs, which we chose to be 1200 days(approximately 4 years). The test used is the Engle & Granger two-step cointegration model. The data used for actual back-testing of the strategy will be different from the data used in pairs searching in order to avoid data snooping bias.

Once we identify several pairs to analyze, we choose the most suitable pair to trade based on the following factors:

- That the pair satisfies the cointegration tests, at the 5% level.
- That the pair comes from the same sector e.g. Banking, etc, this will enable the portfolio constructed from the spread to have some fundamental validity.
- For pairs rejected at the above point, we could re-consider them in scenarios where i.e. pairs for which there seems to be a fundamental reason for them to be cointegrated, e.g. a pair comprised of insurance and banking stock, is ignored in step 1, but it is still viable considering that banking and insurance both fall under "Financial sector".
- Data availability and consistency for the two stocks.
- Market cap of companies, and good fundamentals available, for this reason, the FTSE NSE 25 ensures a requirement of 1 billion KES market cap, thus this is automatically qualified by the universe selected.

Once identified as the suitable pairs for trade, we model the two stocks using the common trends model, which supposes that the price of an asset can be broken into a common factor part and a stationary part. In modelling the asset prices using the common trends model, we model the common factor part using either a single-factor model or a multi-factor pricing model composed of several selected variables. We proceed to work with the best model between the two, using the coefficient of determination R^2. The approach on using multi-factor models closely follows the work of (Vidyamurthy, 2004) which uses the APT framework.

For the single-factor model, we use the log market prices, such as the log prices of a market index such as NASI[1], while for the Multi-factor model, we consider the following factors to include in the model:

[1] NSE All Share Index

- USD/KES currency conversion rate.
- The inflation rate (daily).
- The rate of interest e.g. risk free rate (91-day treasury bill)
- NASI log prices - the market.
- BRENT OIL Price/Barrel.
- XAUUSD - Gold spot prices.

We proceed with the chosen pairs to use the methods listed above for computing the spread(divergence) between the two. For this task we explore several models for computing the hedging ratio. The hedge ratio the most optimal portfolio holding in the second leg of the pair, given that we LONG/SHORT one unit in the first leg of

the pair, in order for us to be hedged. The spread from a portfolio depends on the hedging ratio chosen. The models for computing hedging ratios which we consider are regression based methods. When minimizing variance of the resulting portfolio then the hedge ratio obtained is equivalent to the slope term in an OLS model, this approach is also called the Minimum variance approach. To see this relationhip, we consider a linear combination of the two assets(say A, B) making up the pair:

$$r_A - \lambda r_B$$

were: r_A is the returns from security A, and r_B is the returns from security B. The value of λ which results in the least portfolio returns variance is the best hedge ratio. To compute the variance, we take:

$$(r_A - \lambda r_B)^2 = r_A^2 + \lambda^2 r_B^2 - 2\lambda r_A r_B$$
$$Var(r_A - \lambda r_B) = Var(r_A) = \lambda^2 Var(r_B) - 2\lambda Cov(r_A, r_B)$$

To find the value of λ which minimizes the portfolio variance, we differentiate the resulting funtion while equating it to zero.

$$2\lambda Var(r_B) - 2Cov(r_A, r_B) = 0$$
$$\lambda = \frac{Cov(r_A, r_B)}{Var(r_B)}$$

which also happens to be the definition of the dlope obtained in simple linear regression using the same returns series.

For hedge ratios and the spread constructed using the OLS, we try out various models such as: - Static OLS - We construct the spread by using the hedging ratio obtained by running a regression using all available back-test data. Hence the hedging ratio is constant all through the backtest. - Expanding-window OLS- We construct the spread using the hedge ratios obtained by running an expanding window OLS, which could be refreshed(computed) daily, or after k days(referred to as : *refresh rate*). Hence at any point in time, the hedge ratio is obtained using all the data since inception of the back test. This ensures we have no data snooping bias i.e. *using data which we couldn't possibly have on that day.* Hence the hedging ratio is a dynamic one. - Rolling-window OLS - We construct the spread using hedge ratios obtained by using rolling-forward regression which only uses data of the last n days(referred to as :*lookback period*) in regression. Where the regression statistics could be re-computed after k days(The *refresh rate*). - TLS - We give a motivation for using TLS, since OLS only accounts for variability in one direction while TLS accounts for variability in both directions, and thus could yield more

realistic hedge ratios regardless of the manner in which we position the legs of the pair, as compared to the OLS. We compute both the static TLS, an expanding-window TLS, and a rolling-window TLS.

For every regression method above(static, expanding-window and Rolling-window) except for the TLS approach, we implement both a weighted regression scheme and an unweighted one. We explore this since financial data is usually highly dynamic, and thus weighted regression i.e. which gives more weight to most recent observations than past observations is sensible in the financial modelling context. The weighting technicque we use is linearly growing weights.

We now go into the tradability of the pairs, by analyzing the spread using the Ornstein-Uhlenbeck process which is a model for mean reverting continuous-time processes. We analyze the spread in order to determine:

- Perfect cointegration or simply strong cointegration by analyzing the common trends model to determine the equilibrium and cointegration coefficient as well as to determine the SNR ratio, which will indicate whether the spread might be a profitable trade.
- Consistency of mean reversion, which will be the important in determining whether the pair is practically tradeable.
- Mean of the spread and deviations, which will help us in better understanding the spread and the behavior we expect it to have in the future.
- Threshold levels and styles to use, since different spreads could warrant using different styles e.g. static thresholds, Adaptive thresholds, unequal static thresholds, etc.
- The half life of reversion - to help us determine holding periods and lookback periods when designing adaptive bands around the spread series.

We take transaction costs into account, by assuming a transaction cost of 1% per trade.

For testing the profitability of the pairs trading, we use various statistics to compare it to the Kenyan benchmark: NSE All Share Index by computing:

- Annualized returns.
- Sharpe ratio: A meaure of return per unit risk
- Alpha: A measure for the portion of the strategy's returns that are not attributable to "Beta", or the portion of performance attributable to a benchmark
- Beta: A measure of exposure to market risk of the strategy

ANALYSIS

As at August 9, 2020, the FTSE NSE 25 was comprised of the following securities with their respective industries:

Company Name	Ticker	Industry
Absa Bank Kenya PLC	ABSA	Banking
Diamond Trust Bank Kenya Ltd Ord 4.00	DTK	Banking
Equity Group Holdings Plc Ord 0.50	EQTY	Banking
I&M Holdings Plc Ord 1.00	IMH	Banking
KCB Group Plc Ord 1.00	KCB	Banking
NIC Group Plc Ord 5.00	NCBA	Banking
Stanbic Holdings Plc ord.5.00	SBIC	Banking
Standard Chartered Bank Kenya Ltd Ord 5.00	SCBK	Banking
The Co-operative Bank of Kenya Ltd Ord 1.00	COOP	Banking
Nation Media Group Ltd Ord. 2.50	NMG	Commercial and Services
WPP Scangroup Plc Ord 1.00	SCAN	Commercial and Services
Bamburi Cement Ltd Ord 5.00	BAMB	Construction and Allied
KenGen Co. Plc Ord. 2.50	KEGN	Energy and Petroleum
Kenya Power & Lighting Co Ltd Ord 2.50	KPLC	Energy and Petroleum
Total Kenya Ltd Ord 5.00	TOTL	Energy and Petroleum
Britam Holdings Plc Ord 0.10	BRIT	Insurance
CIC Insurance Group Ltd ord.1.00	CIC	Insurance
Jubilee Holdings Ltd Ord 5.00	JUB	Insurance
Kenya Re Insurance Corporation Ltd Ord 2.50	KNRE	Insurance
Liberty Kenya Holdings Ltd Ord.1.00	LBTY	Insurance
Centum Investment Co Plc Ord 0.50	CTUM	Investment
Nairobi Securities Exchange Plc Ord 4.00	NSE	Investment Services
British American Tobacco Kenya Plc Ord 10.00	BAT	Manufacturing & Allied
East African Breweries Ltd Ord 2.00	EABL	Manufacturing & Allied

Company Name	Ticker	Industry
Safaricom Plc Ord 0.05	SCOM	Telecommunications

Pairs search

We thus formulate the pairs search in such a way that the formation period takes the first 1200 days(approximately 4 years) since inception of the two assets. This is necessary since, all listed assets have different Initial listing dates, thus when joining ptice series of two assets, the starting date will equal the date of the latest Initial listing date between the two. The latest Initial listing date for the stocks included in FTSE NSE 25 is 2014.

Testing spread series for stationarity

The two-step Engle & Granger procedure for cointegration searches for the parameters $\beta_0,\ \beta_1\ \rho$ which yield the best fit to the following:

$$Y_t = \beta_0 + \beta_1 * X_t + \epsilon_t$$
$$\epsilon_t = \rho * \epsilon_{t-1} + \nu_t$$

where

$X_t,\ Y_t$: The log-prices of the assets in our pair.[2]

β_0 : The intercept term in our model, which we will later refer to as the *equilibrium*.

β_1 : The slope term in our regression model, which i commonly referred to as the *cointegration coefficient* or the *hedge ratio*.

ϵ_t : The idiosyncratic random error component, which is assumed to distributed as $\epsilon_t \sim \mathsf{N}(0, \text{œ}^2)$. This error component is actually referred to as the *spread series*.

The test used in determining if the spread is stationary is the Augmented Dickey-Fuller test.

Note that if we conduct a regression analysis of the form:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

we would get different values for the regression coefficients $\beta_0,\ \beta_1$ as well as a different spread series ϵ_t since in an OLS setting the order i.e. the series we use as the independent variable(or dependent variable) affects the value of the regression coefficients, and reversing the order of the regression variables won't equal inverting the coefficients of the first model. However this holds true for Total Least Squares regression, which we shall encounter in the next section.

To counter this in our pairs search, we do OLS regression twice for a single pair, i.e. given the ticker JUB and KNRE, we conduct

[2] We choose to conduct the regression analysis using log prices, although studies conducted on which price series to use indicate that results could be near identical when conducting regression using the log prices, or the real price series themselves.

regression twice. First when JUB is the independent variable, and KNRE the dependent variable, and secondly, we conduct the regression analysis when JUB is the dependent variable, and KNRE the independent variable.

We then filter only pairs for which the ADF p value is less than or equal to a threshold value of 0.01, which would indicate a strationary spread and thus a strong co-integration relationship during the pairs search period. For this threshold value, we obtain 41 potential pairs to investigate.

The results of the ADF test for co-integration are shown below:

	pair	Alpha	Beta	P value
1	ABSA-COOP	1.727	0.442	0.0100000
4	IMH-SCBK	-7.568	1.903	0.0100000
8	SCBK-IMH	4.250	0.393	0.0100000
10	NMG-SCAN	3.380	0.379	0.0100000
13	SCAN-NMG	-4.365	1.651	0.0100000
26	DTK-IMH	2.655	0.772	0.0103041
28	IMH-ABSA	-3.053	2.006	0.0127137
31	IMH-DTK	-3.150	1.229	0.0139747
38	NCBA-IMH	1.549	0.687	0.0191635
40	COOP-ABSA	-2.468	1.718	0.0196202

For this analysis we proceed to work with the pair **DTK~IMH**, since the pair satisfies the requirement of cointegration, same sector, and that the two companies have the same business model.

The common trends model

(Vidyamurthy, 2004) shows that a more simpler approach to modelling the cointegrating relationship between two securities is through the common tredns model. Under this model, the security's price is exoressed as a simple sum of 2 component time series:

- A stationary component, the spread component.
- A random walk component, or the trend component.

This composition ensures that, if two series are cointegrated, then the trend component of the two series must be identical upto a scalar(the cointegrating coefficent). This however would only be true for perfectly cointegrated time series, however for partially cointegrated time series, a strong correlation coefficient between the two trend components is sufficient for a pair to be tradable.

Given two series: X_t, Y_t:

$$Y_t = \eta_{y_t} + \epsilon_{y_t}$$

$$X_t = \eta_{x_t} + \epsilon_{x_t}$$

$$\eta_{y_t} = \alpha\eta_{x_t},\ \alpha : Cointeration\ coefficient$$

Given that we now have suitable pairs to test and trade, we turn to asset pricing models with an aim of giving a fundamental reason as to why the pairs above have a co-integrated relationship. In this section, we investigate if the assets comprising the selected pair have similar risk factor exposure profiles.

In this analysis, we use the log prices instead of the real prices themselves, since it is common practice to model asset prices using the log-normal model in finance.[3]

[3] When modelling prices using a log-normal model, i.e. we assume that the prices are log normally distributed, this enables the random walk model to be applied to the security prices, and the normal distribution to be applied to the log returns

Single-factor models

Many common single-factor asset pricing models such as the CAPM state that the risk-free adjusted returns from a security are fully explained by the risk-free adjusted returns of the market, which gives rise to a linear model of the form:

$$E(R_i) = R_f + [E(R_m) - R_f]\beta_i + \epsilon_t$$

where:

$E(R_i)$: The expected returns from security i

R_f : The risk-free rate of return

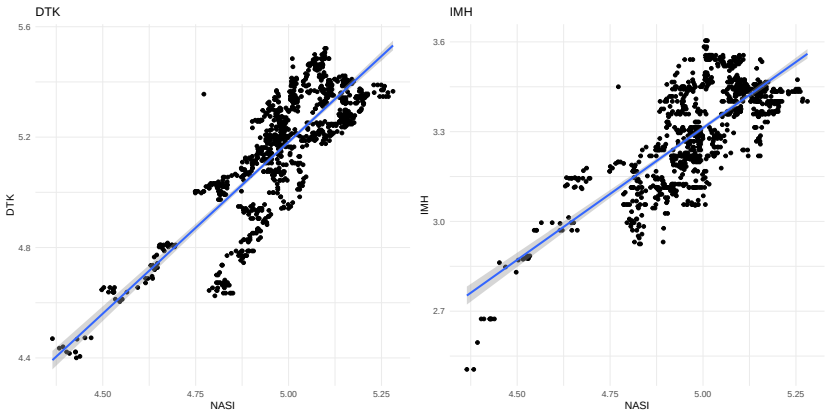
$E(R_m)$: The expected returns from a market portfolio such as a market index.

ϵ_t : The idiosyncratic random component, $\epsilon_t \sim N(0, \sigma^2)$

Thus for this section our interest is to model the log prices and not returns. We fit a regression model for the two assets(Which act as the response variable in our case) using the log prices of the NASI(which acts as a proxy for market) as the only explanatory variable and investigate if the model is a good fit for the data.

Our hypothesis is that: There exists a strong positive linear relationship between the market movement and the movement of the log prices of both assets.

A plot of the relationship between the two is illustrated:



It is evident that the relationship is strongly linear as expected, and the coefficient of determination in the models .68 for DTK, and .57 for IMH. This implies that: 68% of total variability in DTK log prices are explained by log prices of NASI, while only 57% of total variability in the log prices of IMH is explained by the log prices of NASI.

Table 3: Model fitted on DTK log prices

term	estimate	std.error	statistic	p.value
(Intercept)	-1.038101	0.1337518	-7.76140	0
log(NASI)	1.244103	0.0268292	46.37117	0

Table 4: Model fitted on IMH log prices

term	estimate	std.error	statistic	p.value
(Intercept)	-1.0956844	0.1201688	-9.117878	0
log(NASI)	0.8816134	0.0241046	36.574452	0

Multi-factor models

We consider multi-factor models, where instead of a single explanatory variable such as in single-factor models, we utilize multiple variables in an attempt to price a security. The model is of the form:

$$E(R) = R_f + \beta_1 R_1 + \beta_2 R_2 + ... + \beta_p R_p + \epsilon_t$$

where:

R_i : The return contribution from each of the p factors.

β_i : The factor exposures for the respective factors chosen.

R_f : The risk free rate of interest.

ϵ_t : The idiosyncratic random error component

For our analysis we chose the following variables:

- USDKES: The currency conversion rates for Kenyan shillings and US Dollars.
- NASI: The proxy for market returns in Kenya
- Inflation Rates: Kenya's annual inflation rates
- Interest rates: The 91-day treasury bill rates.
- BRENT: Oil price per barrel, as listed in NYSE
- XAUUSD: The spot prices for Gold as listed in the NYSE

We seek to investigate if the two assets making our pair are affected by the above factors in the same manner, since this would be a good justification for their cointegrated relationship. We utilize multiple linear regression to model the log prices of the assets making our pair using the above listed factors as the explanatory variables.

We create a feature vector for the BRENT and XAUUSD by adjusting them using the currency conversion rate, so that we have the spot prices for oil/barrel and gold in Kenyan currency. We also create a feature called RROI[4], which is a rate of interest adjusted for inflation as follows:

We expect the following hypotheses to stay true to the relationship between the log prices and the factors chosen:

[4] Real Rate of Return

Relationship with market movement

- We expect a positive relationship between log prices of assets and the market prices (NASI).[5]

[5] This was illustrated earlier in the section on Single-factor asset pricing models

Relationship with inflation and interest rates

- We expect a negative relationship between the log prices of the assets and inflation, as well as log prices and interest rates, since during period of inflation, rational decision makers prefer consumption as opposed to savings and investments.

When interest rates are higher, then short term securities which are risk free such as 91-day treasury bills are attractive, this drives investors to quit investing in risky assets such as stocks and invest in risk-free government assets. The chart below shows the relationship between the Real Rate of interest(which is simply the rate of inteerst adjusted for inflation) and the log asset prices.

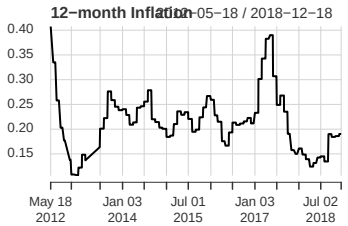


Figure 1: Historical Inflation rates

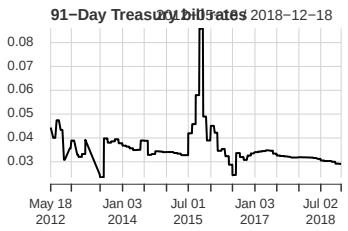
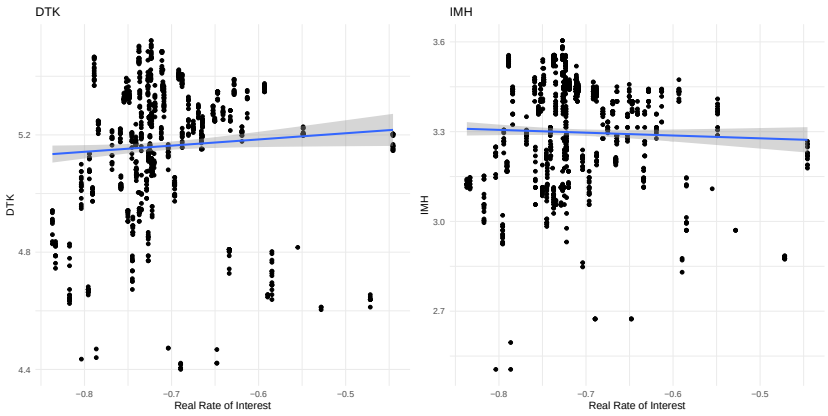


Figure 2: Historical 91-day treasury bill rates



Relationship with commodities

- We expect a negative relationship between log prices of the assets and oil prices or gold prices, since commodities such as gold and oil are usually considered by investors to be **safe haven** investments,

such that in periods of high inflation, uncertainty and increased volatility in stock markets, investors prefer having their investments in commodities such as gold and oil.

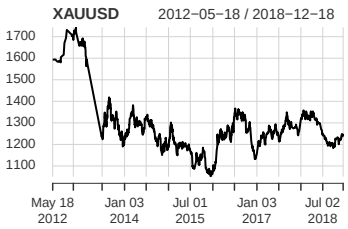
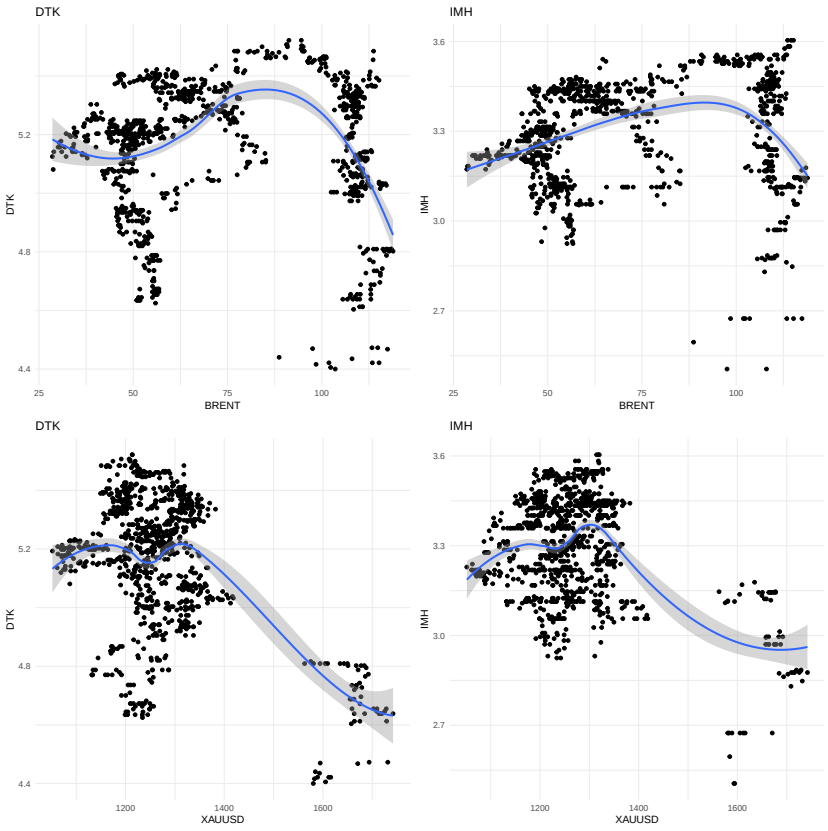


Figure 3: Historical XAUUSD(Gold) spot prices

Figure 4: Historical BRENT(Oil) prices



We fit a no-intercept multiple linear regression model to the dataset, where the explanatory variables are: BRENT prices in KES, Gold spot prices in KES, The real rate of return(RROI), and the market prices(NASI), while the response variables are the log prices for the two assets. The choice for the no-intercept model arises during step wise model selection, whee we found out that the no-intercept model is significantly superior than models with the intercept term. Another possible reason for omitting the intercept in our case could also be the fact that, the intercept in most asset pricing models is usually interpreted as the risk free rate of return, however in our case we already included the risk-free rate of return[6], hence there is no need to conclude an intercept in the model.

[6] We include the risk-free rate of return through the variable RROI, which as defined earlier is the inflation-adjusted risk-free rate of return

From the fitted models, the coefficient of determination(R-squared) in the models are: .9978 for DTK, and .9979 for IMH. This implies that: 99.78% of total variability in DTK log prices are explained by factors chosen, while only 99.79% of total variability in the log prices of IMH is explained by the factors chosen. This level of improvement over the single-factor model is noteworthy, and thus we prefer to use the multi-factor models for the common trends.

Table 5: Model fitted on DTK log prices

term	estimate	std.error	statistic	p.value
NASI	0.0153702	0.0003272	46.97153	0
RROI	-1.7274345	0.0860365	-20.07794	0
`BRENT(KES)`	0.0000703	0.0000036	19.36050	0
`XAUUSD(KES)`	0.0000097	0.0000005	20.24915	0

Table 6: Model fitted on IMH log prices

term	estimate	std.error	statistic	p.value
NASI	0.0102082	0.0002045	49.90617	0
RROI	-1.1634575	0.0537814	-21.63307	0
`BRENT(KES)`	0.0000545	0.0000023	24.00506	0
`XAUUSD(KES)`	0.0000048	0.0000003	16.15332	0

It is evident that the two asset log-prices are affected in the same manner and upto the same magnitude with the macro-economic factors chosen, i.e. they both have a positive relationship with: NASI, BRENT(KES), XAUUSD(KES) and have a negative relationship with: RROI as expected. This helps explain their co-integrated relationships, since they have the same risk-factor exposure profiles.

From the fitted models above, the residual series from the two model would serve as the stationary component in the common trends model, we therefore proceed to test for their stationarity using the ADF test. The reisudals from the IMH model give a p value of 0.01, and DTK residuals give a p value of 0.05 confirming that they are indeed stationary at the 5% level[7]

We proceed to test for perfect/partial cointegration, which naturally arises from using the common trends model[8]. We specifically compute the pearson's coefficient correlation between the trend component fo the two series, such that for perfectly cointegrated series, the correlation coefficient will take a value either +1/-1, while for partial cointegration, the correlation coefficient will strongly be near +1/-1. For this analysis the correlation coefficient was found to be 0.9926, which implies near-perfect cointegration for the trend component for the two price series. This implies that this pair is a good candidate for statistical arbitrage.

[7] The p-value for the DTK Residuals are not strongly significant, although it could still be regarded near-stationary

[8] Since if two series are cointegrated, then their trend components must be identical upto a scalar.

STRATEGY FORMATION

Since the pair chosen, is a suitable candidate for the pairs trade, in this section we proceed further with the pair and compute their cointegration coefficient[9]. We use the methods listed in the methodology section to compute the hedge ratios.

[9] commonly referred to as the hedge ratio

Static OLS Hedge ratio

In this section we compute the hedge ratio using data since inception up to the start of the trading period to compute the hedge ratio. We then utilize this hedge ratio for the rest of the strategy back-test period. This hedge ratio stays constant, hence the name static. We implement both a weighted version of the static OLS, where the weights are arithmetically increasing, and a static version without weighting.[10]

[10] For this analysis, the static OLS is fitted using data from 2009 to 2011, and the hedge ratios estimated are used for the period 2012 to 2018.

Table 7: Static OLS Model fitted (Un-weighted)

term	estimate	std.error	statistic	p.value
(Intercept)	1.309413	0.0578426	22.63751	0
IMH	1.168153	0.0175172	66.68594	0

Table 8: Static OLS Model fitted (Weighted)

term	estimate	std.error	statistic	p.value
(Intercept)	3.611470	0.1040730	34.701304	0
IMH	0.389896	0.0472909	8.244632	0

Expanding Window OLS

An expanding window OLS enables us to compute dynamic hedge ratios, where the regression is conducted in an expanding-window walk-forward fashion so that the last hedge ratio is computed using data since inception.

A motivation for using this model is: Financial time series data such as stock prices are usually subject to changes due to changing market regimes and varying volatility, therefore having a model that is at least adaptive to the changing data does justice to the analysis, instead of having a single static hedge ratio, which would likely be making a strong assumption that things are stationay across time, which isnt the case for financial data. For the expanding window

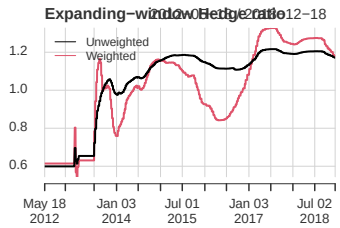


Figure 5: Expanding-window OLS Hedge ratios

OLS, we also compute a weighted version of the same, where the weights increase linearly with time.

Rolling Window OLS

The rolling window OLS enables us compute the hedge ratio in a walk-forward fashion where we set a lookback period equal to 30 days, and a refresh rate of 5 days, such that the regression is only conducted using the *30* previous data points and recomputed after *5* days.

This enables the hedge ratios from the model to be even more adaptive to the changing market conditions, with the downside that the hedge ratios would be too noisy. We implement a weighted version for the same.

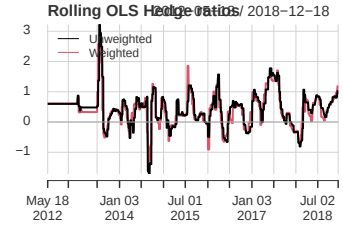


Figure 6: Rolling-window OLS Hedge ratios

Total Least Squares

A major downside to using the OLS as a model for hedge ratio computation, is that the hedge ratios computed are dependent on which asset we choose as our independent/dependent variable. This is because in the OLS framework, the explanatory variable is assumed to be constant, and known. This assumption might be too strict since in our analysis where we use one price series to explain the other, both price series are subject to random fluctuation as demonstrated in their High-Low range. [For this analysis, the static TLS using data from 2009 to 2011 gives an intercept of 0.7849, with a beta(hedge ratio)of 1.3272]

Total Least Squares on the other hand, takes care of this scenario, since in the TLS framework, variability/errors from both series is taken into account in the regression model. TLS does this by minimizing sum of perpendicular squared distance between the data points and the regression line.For TLS regression, the hedge ratio computed when asset A is the independent variable equals the inverse of the hedge ratio computed when asset A is the dependent variable, which would be intuitive for a pairs trade. It is important however to note that the TLS is unstable when few data points are used, thus when setting look-back period, we set a larger period than when using the rolling OLS.[11]

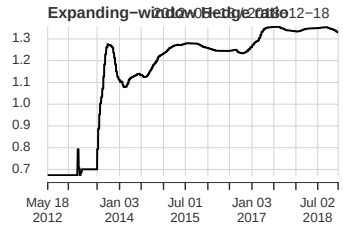


Figure 7: Expanding-window TLS Hedge ratios

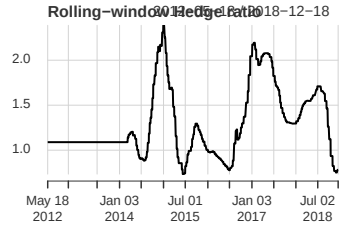


Figure 8: Rolling-window TLS Hedge ratios

[11] We use an arbitrarily chosen 200-day rolling OLS

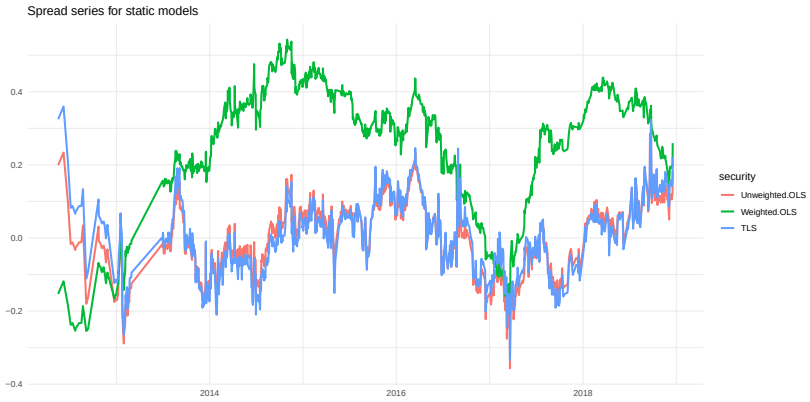
Spread Modelling

Once the hedge ratios are obtained we proceed to construct the spread, which would be the residuals of the linear regression model fitted, a plot of the spreads is shown. It is clearly seen that the spreads are mean-reverting around a mean of 0[12].

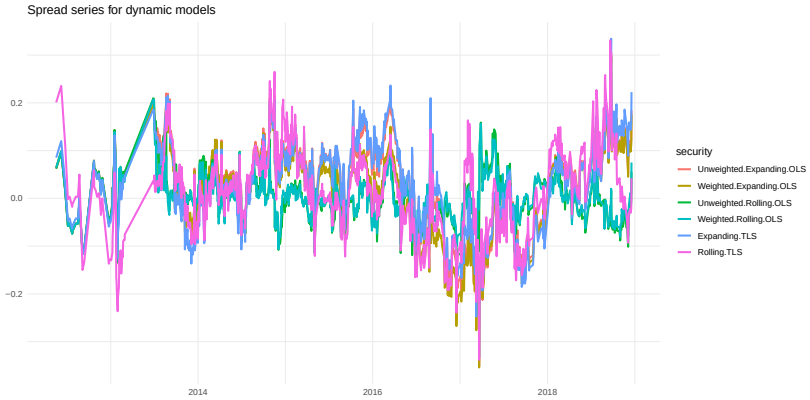
[12] The mean is not strictly 0, due to the equilibrium part(the intercept) of the regression models formulated

The mean reverting component is essential, since it enables us utilize models for mean reversion such as the Ornstein-Uhlenbeck process, to model the spread.

The spreads for the static OLS and TLS as shown below are not seen to be strongly mean reverting, this could be due to using static hedge ratios and equilibrium values which have not fully adapted to changing condition in the dynamics of the price series.



The spread for the expanding and rolling models as shown below are shown to be strongly mean-reverting about 0, due to their ability to quickly capture the changing dynamics of the price series.



To test for stationarity of the spread which is a requirement for the pairs trade to be effective, using a formal method, we use the ADF test. For the spreads generated using the static hedge ratios, the p vaues for the ADF test are all above 0.05(5%) except for the static TLS, indicating that the static TLS spread is stationary at the 5% level.

Spread	P value
Unweighted OLS	0.1400746
Weighted OLS	0.2555275
TLS	0.0462977

The spreads from the dynamic models, only spreads which exhibit stationarity at the 5% level for the ADF test are the rolling OLS, and TLS, which we will proceed to use for actual trade generation.

Spread	P value
Unweighted Expanding OLS	0.2159827
Weighted Expanding OLS	0.3003952
Unweighted Rolling OLS	0.0100000
Weighted Rolling OLS	0.0100000
Expanding TLS	0.1862421
Rolling TLS	0.0114211

From the above table, we therefore proceed with the spreads which exhibit stationarity (at the 1% ADF level), which are the spreads resulting from: Rolling OLS(both weighted and unweighted), and Rolling TLS.[13]

[13] We will also trade the static TLS spread, and show its trade statistics.

The Ornstein-Uhlenbeck process

The OU process is a continuous time stochastic process commonly used in finance to model a mean-reverting process whereby a process is said to follow the OU process if its stochastic differential equation is of the form

$$dX_t = \kappa(\theta - X_t)dt + \sigma dB_t, for\ dX_t^2 = \sigma^2 dt$$

where

dX_t : The process increment between time t and dt

θ : The expected value of the process in the long run, and is assumed constant. It is commonly referred to as the drift component.

κ : The speed of reversion of the process towards its long term expected value and is assumed constant.

dt : An infintesimal increase in time t

$\sigma, (\sigma > 0)$: Instantaneous diffusion term of the process, which is used to measure volatility and is assumed constant.

dB_t : Increment in interval $(t, t + dt)$ of a standard Brownian motion, under a probability measure P and is distributed as a $N(0, t)$ random variable.

Solving the SDE[14], we obtain that the mean reverting process X_t :

[14] Stochastic Differential Equation

$$X_t \sim \mathsf{Normal}(\mathsf{X_0 e^{-kt}} + \grave{}\,(1 - \mathsf{e^{-kt}}), \frac{\text{œ}^2}{2\check{}}(1 - \mathsf{e^{-2\check{}t}}))$$

such that:

$$E(X_t) = X_0 e^{-kt} + \theta(1 - e^{-kt})$$

$$Var(X_t) = \frac{\sigma^2}{2\kappa}(1 - e^{-2\kappa t})$$

Since our spreads are mean-reverting process, we utilize the OU-Process to model the resulting spreads from each model above, and obtain the following:

- The long-term mean
- The long-term variance
- The half life of mean reversion
- The speed of mean-reversion

The four components would be useful in signal generation and trading rules. To obtain the above statistics, we use the fact that the OU-process can be thought of as a continuous time AR(1) process, where the discretized version of the OU-process is shown below.

The limiting distribution of the OU-Process X_t has a mean and variance as shown below:

$$\lim_{t\to\infty} E(X_t) = \lim_{t\to\infty}[X_0 e^{-kt} + \theta(1 - e^{-kt})] = \theta$$

$$\lim_{t\to\infty} Var(X_t) = \lim_{t\to\infty}[\frac{\sigma^2}{2\kappa}(1 - e^{-2\kappa t})] = \frac{\sigma^2}{2\kappa}$$

The half life of mean reversion which is the distance half way between the mean of the long-term mean of the process and the current value of the spread series, we compute it as shown below:

$$Recall\ dX_t = \kappa(\theta - X_t)dt + \sigma dB_t, for\ dX_t^2 = \sigma^2 dt$$

where $E(X_t) = X_0 e^{-kt} + \theta(1 - e^{-kt})$, the half way point is therefore:

$$X_0 + \frac{\theta - X_0}{2}$$

We therefore need to find h so that:

$$E(X_h) = X_0 + \frac{\theta - X_0}{2}$$

thus:

$$X_0 e^{-\kappa t} + \theta(1 - e^{-\kappa t}) = X_0 + \frac{\theta - X_0}{2}$$

$$e^{-\kappa h}(X_0 - \theta) + \theta = \frac{\theta - X_0}{2}$$

$$e^{-\kappa h}(X_0 - \theta) = \frac{-2\theta + 2X_0 + \theta - X_0}{2}$$

$$e^{-2\kappa h} = \frac{1}{2},\ taking\ log$$

$$h = \frac{ln(2)}{\kappa}$$

Therefore the half life of mean reversion only depends on κ the speed of mean reversion. It is important to note that, the higher the speed of mean reversion, the shorter the period needed for the mean to reach the midway point between the current value of the process and its long term mean, thus for a pairs trade, we would want the half-life of mean reversion to be shorter, since longer time periods could result to losses.

Given all the information upto time $t-1$ F_{t-1}, we can discretize the exact analytical solution of the OU Process, to be as shown below:

$$X_t = X_{t-1}e^{-\kappa\Delta t} + \theta(1-e^{-\kappa\Delta t}) + \sigma\sqrt{\frac{1}{2\kappa}(1-e^{-\kappa\Delta t})}\epsilon_t,\ \epsilon_t \sim N(0,1)$$

We further simplify it to:

$$X_t = X_{t-1}e^{-\kappa\Delta t} + \theta(1-e^{-\kappa\Delta t}) + \epsilon_t,\ \epsilon_t \sim N(0, \frac{\sigma^2}{2\kappa}(1-e^{-\kappa\Delta t}))$$

Based on the above equation, we consider the OU Process as a continuous-time version of the discrete time AR(1) Process of the form:

$$X_t = \alpha + \beta X_{t-1} + \epsilon_t$$

where:
$\alpha = \theta(1-e^{-\kappa\Delta t})$
$\beta = e^{-\kappa\Delta t}$
$SE(Standard\ error) = \sigma\sqrt{\frac{1}{2\kappa}(1-e^{-\kappa\Delta t})}$

We therefore regress the process current spread value against its previous lag value, to obtain the regression parameters for the AR(1) model: β_0, β_1, and map them back to the parameters of the OU-Process. For an AR(1) model of the form:

$$X_t = \beta_0 + \beta_1 X_{t-1} + \epsilon_t$$

we obtain the parameter for the OU Process as:

$$\hat{\theta} = \frac{\beta_0}{1-\beta_1}$$
$$\hat{\kappa} = \frac{1}{\Delta t} log_e(\frac{1}{\beta_1})$$
$$\hat{\sigma} = SE\sqrt{\frac{1-\beta_1^2}{2\hat{\kappa}}}$$

For the spread series chosem above the OU-process statistics computed are as follows:

	Mean(theta)	Reversion speed(kappa)	Half life	variance
Static TLS	-0.0016624	0.0749272	9.250941	0.0014379
Unweighted Rolling OLS	-0.0000961	0.1683943	4.116215	0.0008636
Weighted Rolling OLS	-0.0000202	0.2239602	3.094956	0.0007839
Rolling TLS	-0.0017320	0.0981610	7.061333	0.0015461

TRADE EXECUTION

In this section, we proceed to formulate rules for trading signals based on the spread, its long term mean, variance and half life. We consider the approach of an expanding horizon equi-distant 2-standard deviation bands as used by several researchers such as in (Gatev et al., 1993). We also consider the approach of using adaptive bands, which utilize the concept of moving averages instead of the constant average and standard deviation used in equi-distant bands. This concept of moving bands is illustrated in (Vidyamurthy, 2004), where he considers that in prescence of perfect cointegration, then mean and variance are constant over time and thus using equi-distant bands is sufficient, however in the prescence of a **mean drift**[15], since the variance might grow linearly with time, we use moving bands, which will be more adaptive to the movements in the spread as opposed to the equidistant bands.[16]

The mean drift component is illustrated as shown:

Considering two securities A and B, the returns from the portfolio constructed under the common trends model is given below:

$$r_A = r_A^{cf} + r_A^{spec}$$
$$r_B = r_B^{cf} + r_B^{spec}$$

Constructing a portfolio where we LONG 1 unit of A and SHORT β units of B, we obtain the portfolio: $r_A - \beta r_B$.

This is equivalently written as:

$$r_A - \beta r_B = (r_A^{cf} - \beta r_B^{cf}) + (r_A^{spec} - \beta r_B^{spec})$$

Thus when the common factor component are not identical upto the scalar β as in the context of perfect cointegration, then the portfolio spread becomes:

$$spread_{portfolio} = spread_{cf} + spread_{spec}$$

So that if the $spread_{cf}$ is non-stationary, then the overall portfolio spread equals the specific spread(which i stationary) with a stochastic drift to its mean value(due to spread from the common factor component, this would violate the cointegration condition on spread stationarity.

In the abscence of perfect cointegration, (Vidyamurthy, 2004) computes a statistic called the SNR[17] ratio, which would be a measure of whether a pair is still tradable, when the cointegration conditions are violated.

[15] When using the common trends model to formlate the pairs trading strategy,since the common factor/trend component of the two assets are identical upto a scalar and will cancel out in the case of perfect cointegration, then the overall portfolio spread will equal the spread from the linear combination of the stationary components only, which will be stationary as well, however when the cointegration relationship is not perfect, then the portfolio spread will also have a spread from the trend components of the two series, whose variance grows linearly with time, this component is what is termed as the mean dfrift

[16] We do not consider the approach on using rolling-adaptive bands in the study where most researches suggest setting the lookback roling period equal to the half life of mean reversion, since by setting the half life equal to the lookback period of the adaptive bands, we get noisy bands, due to the short half life periods such as 5 days. For more on rolling-adaptive bands, commonly called bollinger bands, seeErnest P. Chan's book

[17] Signal-To-Noise

$$SNR = \frac{\sigma^2_{st}}{\sigma^2_{non-st}}$$

where:

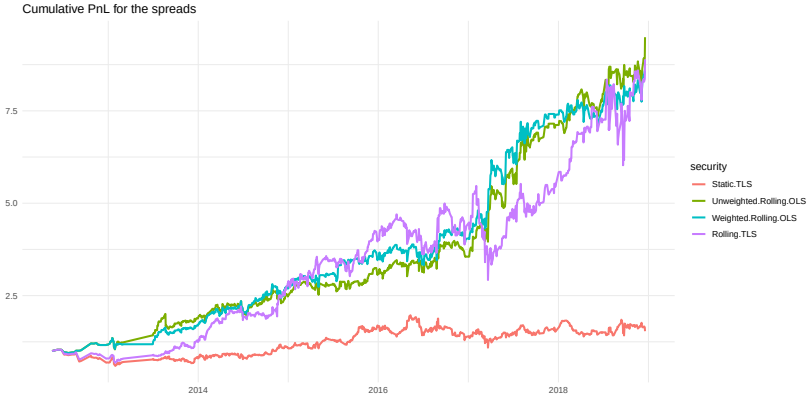
σ^2_{st} : Variance of the stationary spread

σ^2_{non-st} : The variance of the non-stationary component of the spread, the common factor spread.

If we have the non-stationary variance as close to zero as possible, then the SNR is large, which would mean higher signal-to-nosie ratio and thus a better fit for tradability.

For our pair **DTK~IMH** during the trading period, the SNR is: 2.5164, whereby an SNR above 1 is considered good for tradability purposes.

For the four spread series, the trading cumulative profit and loss is illustrated below, where we open positions when the spread deviates past the two standard deviation mark, and exit the trades on reversion to mean.



The summary statistics for the three rolling strategies are shown below:

	Static TLS	Unweighted Rolling OLS	Weighted Rolling OLS	Rolling TLS
Annualized Return	0.1147	0.7652	0.7208	0.7364
Annualized Std Dev	0.5268	0.3993	0.3615	0.5447
Annualized Sharpe (Rf=0%)	0.2177	1.9164	1.9939	1.3520

To benchmark the performance of the three rolling pairs strategies, we benchmark the strategies using the NSE All Share Index, which is a market proxy and benchmark in the Nairobi Securities Exchange.

	Static TLS to NASI	Unweighted Rolling OLS to NASI	Weighted Rolling OLS to NASI	Rolling TLS to NASI
Alpha	0.0010	0.0025	0.0023	0.0028
Beta	-0.0086	0.1265	0.1008	0.0336
Beta+	0.0140	0.1457	0.1172	0.0428
Beta-	-0.0129	-0.0055	0.0044	0.0285
R-squared	0.0000	0.0099	0.0077	0.0004
Annualized Alpha	0.2854	0.8641	0.8011	1.0048
Correlation	-0.0051	0.0997	0.0877	0.0194
Correlation p-value	0.8715	0.0016	0.0056	0.5399
Tracking Error	0.6154	0.4834	0.4582	0.6241
Active Premium	-0.0411	0.6103	0.5658	0.5815
Information Ratio	-0.0668	1.2626	1.2348	0.9318
Treynor Ratio	-13.3532	6.0484	7.1536	21.8915

From the above statistics: the unweighted rolling OLS performs better in terms of annualized returns and sharpe-ratio as compared to the weighted rolling OLS, TLS AND static TLS. The four spreads trading strategies have excess returns(alpha) above the NASI[18]. It is also evident that the strategies are "market neutral" both overally(*beta*) and in bull(*beta+*) and bear(*beta-*) markets, since the beta[19] is almost 0.

[18] NSE All Share Index

[19] An assets's/strategy's beta is the measure of the strategy's/asset's market risk, and is usually obtained by regressing the asset's returns with the market returns. A beta equals 0 implies little-to-no market risk, hence the term "market-neutral"

CONCLUSION

The study compares various hedge-ratio construction methods based on regression analysis, using stocks listed in the NSE. The framework used which starts from pairs selection, then spread modelling and finally trading rules was used throughout this study. The static hedge ratios computed using a one-time regression do not appear to be stationary except in the case of TLS, and are thus eliminated during the trade backtesting. Of the four spread series, the un-weighted rolling OLS performs the best with 76% returns. It should however be noted that the high returns could diminish a bit when transaction costs and slippages are accounted for.

This study also concludes that indeed the returns generated from the pairs trade are market neutral, when compared to the NSE All Share Index.

Therefore, we conclude that in the Nairobi Securities Exchange, statistical arbitrage opportunities indeed exist, and could be exploited profitably by arbitrageurs who seek arbitrage opportunities, hedgers who seek to hedge their trading positions, and speculators who bet on mean-reversion aspect of spreads between stocks.

RECOMMENDATIONS

Future work on the statistical arbitrage on NSE should try incorporating other methods of computing hedge ratios, such as the ratio based approaches and kalman-filtering approaches.

Future work on statistical arbitrage in NSE should also try trading this strategy in the futures markets, and commodity markets.

Future work on the same should study more trading rules such as optimal entry and exits and comparison to a simple entry and exit strategy, to see if the results are significantly different.

Future work on statistical arbitrage in the NSE should study possible inter-global markets cointegration, by investigating pairs which span across different markets, and studying if profitable trading is possible.

Future work on statistical arbitrage in the NSE should study possibility of constructing such portfolios using Exchange Traded funds.

REFERENCES

1. 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.

2. Vidyamurthy, G. (2004). Pairs Trading: quantitative methods and analysis (Vol. 217). John Wiley & Sons.

3. Chan, E. P. (2021). Quantitative trading: how to build your own algorithmic trading business. John Wiley & Sons.

4. Chan, E. (2013). Algorithmic trading: winning strategies and their rationale (Vol. 625). John Wiley & Sons.