The Paper, the problem

The objective of the study is to formally and statistically test the performance of the optimal statistical arbitrage trading model from Bertram (2010) for the quantitative trading of spreads in the crude oil and refined products markets. Bertram's model was originally modelled and tested on the dual listing of ANZ stock in Australia and New Zealand (ANZ.AU & ANZ.NZ). Cummins and Bucca (2012) then adapted the model and investigated it on oil markets, testing a total of 861 spreads across Brent, WTI, heating oil and gas oil futures.

The paper starts by talking about the energy market and the future contracts considered throughout the paper. Section 3 introduces the optimal statistical arbitrage trading model that is used as the basis for the paper, providing proofs on how to obtain optimal entry and exit thresholds for live spread trading under different objective functions (max expected return, max sharpe ratio).
As the paper involves testing large amounts of implementations of the same model on different spreads, the issue of data snooping bias and multiple hypothesis testing are then introduced. The author then spends time addressing varying methods to control for these issues, ultimately selecting the stepdown procedure and the balanced stepdown procedure ensuring no more than 1% of the tests represent false discoveries.

The paper then goes on to make comments on transaction costs, liquidity in oil and refined products markets, and then finishes with a discussion of results and further comments to the strategy's robustness to transaction costs.

Implementation

To replicate the paper, we must first build Bertram's model, calibrate the OU process, estimate the optimal entry/exit thresholds and then backtest the trading strategy. This must be done for each spread while also using walk-forward analysis techniques to ensure no data snooping occurs.

The process for thoroughly testing each spread will go as follows:

- Sourcing historical data for two given oil contracts and creating a log spread for the two contracts
- Calibrating the above log-spread to the general OU process $ds_t = \alpha(\mu - s_t)dt + \sigma dW_t$
- Storing estimates for α, μ & σ to be used throughout the implementation
- Estimate transaction costs,c, for entering a trade in the specific spread
- Determine optimal entry threshold, a & m, using Bertram's maximum expected return formula with α, σ, c (found previously) as inputs
- Adjust a & m by long run mean, μ, to obtain $\hat{a}$ & $\hat{m}$ as above
- Run an out-of-sample backtest executing trades from $\hat{a} \rightarrow \hat{m}$ as well as the transition back from $\hat{m} \rightarrow \hat{a}$.
- Evaluate performance of backtest results and store results

Computationally, this process must be efficient as it is to be repeated thousands of times. As such, each component was built out as a function to be called upon when needed and more importantly iterated over many times. Each python file has further mathematical decompositions where necessary and validation is completed below.

The strategy

Bertram's trading model starts by assuming the spread between two asset log-price series, denoted by s_t, is given by the following zero-mean OU process:

$$ds_t = -\alpha s_t dt + \sigma dW_t$$

With α,σ > 0 and W_t is a wiener process.

The optimal entry and exit levels of the trading strategy are defined by a and m respectively.

A continuous time trading strategy is defined by entering a trade when, $s_t = a$, exiting the trade when $s_t = m$ and waiting until the process returns to $s_t = a$.

The goal of the strategy is to determine *optimal* levels for a and m such the expected return is maximised. Similarly, a numerical solution can be obtained for maximising sharpe ratio, however Cummins and Bucca only considered the analytical approach of maximising expected return.

Mazimising Expected return

Using Bertram's equation (18) we can find the value for a that maximises expected return for the strategy. Using values obtained for α, σ and transaction costs c, where $a < 0$ and $m = -a$.

$$a = -\frac{c}{4} - \frac{c^2\alpha}{4(c^3\alpha^3 + 24c\alpha^2\sigma^2 - 4\sqrt{3c^4\alpha^5\sigma^2 + 36c^2\alpha^4\sigma})^{1/3}} - \frac{(c^3\alpha^3 + 24c\alpha^2\sigma^2 - 4\sqrt{3c^4\alpha^5\sigma^2 + 36c^2\alpha^4\sigma^4})^{1/3}}{4\alpha}$$

As oil spreads often do not have a zero-mean, a slight adjustment to the model is made. Each spread is fit to the general OU process introducing, μ, long-run mean.

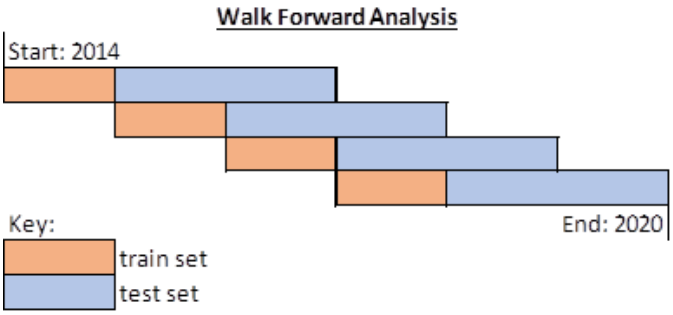
$$ds_t = \alpha(\mu - s_t)dt + \sigma dW_t$$

All calculations are the same, except optimal entry and exit thresholds a and m are adjusted to:

$\hat{a} = a + \mu$ and $\hat{m} = m + \mu$

And to optimise trade performance trades are executed from $\hat{a} \rightarrow \hat{m}$ as well as the transition back from $\hat{m} \rightarrow \hat{a}$.

Walk forward analysis was also performed when testing and backtesting. The data was split into 4 train sets and 4 out-of-sample validation sets (as below). This was to ensure that no look-ahead bias was occurring while optimising and backtesting.



Critique of Method

The study examines quantitative trading of spreads in the crude oil and refined markets. After thousands of backtests the authors were able to identify profitable strategies that are likely to perform out of sample. Considering their results and discussion, I believe the paper to be successful in testing and evaluating the problem. When Cummins and Bucca selected Bertram's trading strategy it was an appropriate choice at the time. However, after publishing there has been plenty more research into statistical arbitrage and pairs trading. If a similar study were to be conducted today authors may want to consider using jump-diffusion models, partial cointegration or even a machine learning models as the basis for their paper.

Another aspect that the authors could have considered was exploring the trading of crude and refined products intraday or at higher frequencies. This would make the models easier to fit to an OU process and would ultimately find more optimal entry points than when prices are only sampled daily. The authors even comment on this being a potential benefit using higher-frequency data would bring. This was likely not examined due to inability to source quality data for that many contracts at higher frequencies, but if the goal is to determine potential profitable sources of alpha one may want to consider using higher frequency data.

After reviewing and comparing results, I am uncertain how the authors were able to produce such successful strategies without data-snooping. In my own implementation I was very careful to not optimise my parameters on the same data in which I ran my backtests. It is unclear to me if Cummins and Bucca did the same when running their tests. There are plenty of reasons why, structurally, our results would be different and the fact our backtesting datasets did not coincide makes it hard to compare. However, I would have expected some results to be close to replicating the performance of the original backtests but with careful control of my data, I was not able to obtain the results. This can be attributed to alpha decay of the strategy but also may come from the authors optimising on the same data they backtested on. Running my own experiment, performance considerably improved when running a backtest in this manner. If the authors were clearer on how exactly they run their optimisation and backtest this would be a lot clearer. But I am led to believe that they may have not properly controlled for data snooping bias.

Difficulties and further critique

Cummins and Bucca provide a very in-depth article explaining oil markets, spread trading and the model they tested. However, when attempting to recreate their method there were some notable shortcomings in their paper, listed below. Some of my critiques on the paper came from difficulties I had with the implementation, so I have decided to combine the two sections. Below each critique is comments on how I overcame each or any assumptions I had to make when reproducing the paper.

- **Unclear on implementation of Bertram's model –** while they provide a mathematical outline of the model and describe Bertram's model, it is particularly brief. Crucially, it was unclear if the calculated optimal entry threshold was obtained through maximising sharpe ratio or max expected return. As mathematical proofs for both were shown, it was very unclear to understand which method they used and as they did not provide any way to validate. However, I believe they used the analytical solution to maximise expected return. As I mention below though, with no validation or model estimates in the paper it is ambiguous.

Solution: I had to decide on which of the two methods to use for my paper and ultimately decided on maximising expected return. In the introduction of the paper, Cummins and Bucca mention they chose Bertram's model because "analytic solutions exist for the optimal entry and exit levels determined through maximising the expected return per unit time". Going off this quote, I must assume they chose maximum expected return, as only an analytic solution exists for maximising expected return, not sharpe ratio. As no results or calculations were shown I was unable to validate my implementation. As such I had to rely on Bertram's original paper to validate my results.

- **Lack of validation –** replicating the model involved calibrating and implementing a variety of different functions to be used thousands of times. It was critical that these functions and parameter estimates were valid, yet Cummins and Bucca did not provide samples of their parameter estimates. This made validation for this report especially hard. Validation involved having to separately confirm results using other research papers and resources.

Solution: Validation has been completed in another section of this report. Without having anything to use from the paper I largely had to rely on other sources to validate my approach and results.

- **Unclear on spreads used –** in a similar fashion to above, the report involved testing the model of a variety of crude and refined spreads. The paper claims to test a variety of "crack, locational and calendar" spreads but does not offer a comprehensive list or any further insights on to the exact spreads. From piecing together the series names in the appendix spreads are labelled: "GO M02 – HO M01" which would equate to a gas oil second month contract vs. a heating oil front month contract. Extrapolating from there, it appears all possible 1:1 contract spreads were tested from the authors universe of contracts.

Solution: I decided to go with my original assumption of testing all possible 1:1 spreads in my data set. This leaves my report with 300 spreads with a range of calendar, locational and 1:1 crack spreads. It is unclear if the author tested more exotic spreads, but from review of the results in the appendix it appears all are simple spreads.

- **Limited output to review –** the only performance measures the paper covers in the backtests are average daily return, sharpe ratio and average trade length. For a backtest, only reviewing three performance metrics does not fully capture the depth of the performance of a strategy. Results were also only listed for the top performing strategies or were aggregated into summary tables which made it difficult to evaluate the full distribution of results. This became especially challenging when attempting to compare results from my own implementation. In the appendix there are some more results but it would have been very useful to have a wider range of results or some distributions to review.

- **Transaction costs** – Cummins and Bucca mention in their section on transaction costs that the source they used for estimation does not coincide with their test data. This means they potentially overstated or understated their transaction costs for their paper. As transaction costs are an input for entry and exit, threshold optimisation this is an important metric to get right. To compensate they added a section on the strategy's robustness to transaction costs.

Solution: to keep in line with the original paper, I used the same estimates for transaction costs. The main issue Cummins and Bucca had with their estimates were that they were too recent which likely makes it a more accurate estimate for my sample set.

Validation

Validation for the paper implementation is broken up into individual components and individually validated. This ensures accuracy along the entire process and greater confidence once all the components are brought together and tested. Associated python code or other files are listed for each component. The jupyter notebook files also contain further mathematical expositions and notes to supplement the python implementation and output.

Data & creating log spreads

Associated Code or Files: "Full_set.csv"

Purpose: Sourcing crude oil and refined products data and generating 100's of log spreads to use as the key input for the model.

All data for this assignment was sourced from Bloomberg, collecting the first five continuous contracts for Brent, Crude, Gas Oil, Heating Oil and Gasoline. Data was collected from the beginning of 2014 to December 2020. This represents some of most active oil and refined product futures contracts in the market and the basis for this report and the related paper. Bloomberg sources rates from NYMEX and ICE which are the two main exchanges for crude and refined products. Sourcing rates straight from the exchange means we can be sure that prices are accurate and are a true measure of the close price each day.

Log spreads were created using a simple python function to ultimately create 300 spreads. These spreads were then each split up, calibrated by the OU process, then optimal entry and exit thresholds were calculated. A full backtest for each spread was then conducted and results stored in an output file.

Transaction Costs

Purpose: As a key input to estimating optimal entry and exit thresholds transaction costs for each product must be approximated.
In section 5.1 of their paper, Cummins and Bucca detailed the basis for their estimates of transaction costs. They also provided estimated relative transaction costs for each product. For this paper, it is assumed that these costs are reflective of institutional trading transaction costs and are appropriate for inputs in estimating optimal entry thresholds.

Product	Transaction cost
WTI	0.0940%
Brent	0.0289%
Heating oil	0.2482%
Gas oil	0.2482%
Gasoline	0.2482%

Detailed in other parts of the report are comments about robustness of transaction costs and potential limitations of the paper. But for the purpose of inputs for maximising expected return, I believe it was best to use the same estimates as the paper.

Calibrating the OU process

Associated Code or Files: "Validation1-Calibrating the OU.ipynb", "results_set_1.csv", "results_set_2.csv", "results_set_3.csv", "results_set_4.csv"

Purpose: Calibration and estimation of the Ornstein Uhlenbeck (Mejía Vega 2018) process allows for estimation of key parameters (mu, alpha & sigma) to determine optimal entry and exit thresholds.

Ordinary Least Squares (OLS) regression was used to estimate parameters α, μ and σ where:

- α → represents the rate of mean reversion.
- μ → represents the long term mean.
- σ → represent the volatility.

Cummins and Bucca mention they used OLS regression in their paper to estimate parameters but offer no way to validate results. However, referring to Mejia Vega's "*Calibration of the exponential Ornstein–Uhlenbeck process when spot prices are visible through the maximum log-likelihood method. Example with gold prices*", we can validate the OU parameter estimates. Using tables 4 & 5 from the paper, we can transform the regression results into the OU parameters α, μ and σ.

	a	b	Std(η)
Daily	0.9998	0.0012	1.23%
Annual	0.9515	0.3451	19.06%

	μ	α	σ
Paper Results – Daily	7.22	0.0002	1.23%
Paper Results – Annual	7.12	0.0497	19.53%
Validation Result – Daily	**7.215874924**	**0.000166313**	**1.2301%**
Validation Result-Annual	**7.115463917**	**0.049715592**	**19.536%**

With validation complete, to apply in the context of the paper, we can calibrate the OU parameters for each spread using statsmodels OLS regression to estimate regression coefficients a, b and Std(η). After that, using equations below, we can find μ, α and σ to be used throughout the rest of the model.

Below is the mathamatical exposition from my python validation file titled “Validation1- Calibrating the OU.ipynb”. As formulas did not exist in the original paper, I referenced Mejia Vega’s paper on calibrating OU processes to adapt for the purposes of this paper. Calibration follows below:

Calibration using least squres regression

The relationship between consecutive observation S_i, S_{i+1} in linear with a iid normal random term ϵ:

$$S_{i+1} = aS_i + b + \epsilon$$

The relationship between the linear fit and model parameters is given by:

$$a = e^{-\alpha\delta}$$

$$b = \mu(1 - e^{-\alpha\delta})$$

$$sd(\epsilon) = \sigma\sqrt{\frac{1 - e^{-2\alpha\delta}}{2\alpha}}$$

rewriting these equations gives,

$$\alpha = -\frac{\ln a}{\delta}$$

$$\mu = \frac{b}{1 - a}$$

$$\sigma = sd(\epsilon)\sqrt{\frac{-2\ln a}{\delta(1 - a^2)}}$$

Maximising Expected Return

Associated Code or Files: "Validation2- Max Expected Return.ipynb", "results_set_1.csv", "results_set_2.csv", "results_set_3.csv", "results_set_4.csv"

Purpose: Analytic solution to find optimal entry and exit thresholds (a and m) such that expected return is maximised.

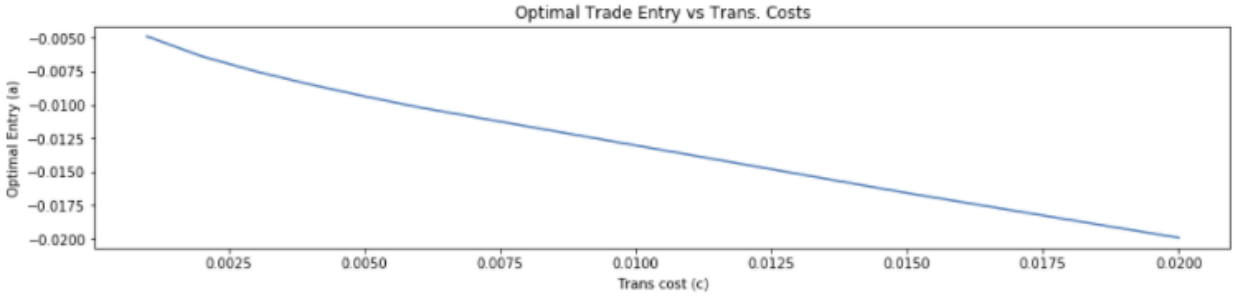
Mazimising Expected return

Using Bertram's equation (18) we can find the value for a that maximises expected return for the strategy. Using values obtained for α, σ and transaction costs c, where $a < 0$ and $m = -a$.

$$a = -\frac{c}{4} - \frac{c^2\alpha}{4(c^3\alpha^3 + 24c\alpha^2\sigma^2 - 4\sqrt{3c^4\alpha^5\sigma^2 + 36c^2\alpha^4\sigma})^{1/3}} - \frac{(c^3\alpha^3 + 24c\alpha^2\sigma^2 - 4\sqrt{3c^4\alpha^5\sigma^2 + 36c^2\alpha^4\sigma^4})^{1/3}}{4\alpha}$$

Using table 1 in Betram's "*Analytic solutions for optimal statistical arbitrage trading*" we can validate our approximation for a. Below is the table with Bertram's results for maximising expected return, with α = 180.967, σ = 0.1538 and varying transaction costs. Also below are results obtained through python implementation validating that the implementation is producing expected results.

c	a (Bertram's solution)	a (python implementation)
0.0010	−0.0048750	-0.00487500478007779
0.0020	−0.0063549	-0.006354930125607594
0.0030	−0.0074910	-0.007491010302362016
0.0040	−0.0084682	-0.008468243601591667
0.0050	−0.0093531	-0.009353129481214492
0.0060	−0.0101780	-0.010178034089253355
0.0150	−0.0165920	-0.016591733968014284
0.0175	−0.0182670	-0.01826659145361514
0.0200	−0.0199340	-0.019934125311568858



As we are able to validate that the output returns accurate solutions, to apply in the context of the paper involves feeding in appropriate inputs. The function takes three arguments c, α and σ. Transaction costs (c) have been discussed and estimation for α and σ have been validated above in calibrating the OU process. Therefore, knowing our inputs to the equation are validated and accurate, while also having cross checked the solution to our equation, we can be confident that our output for estimating optimal entry and exit thresholds will match estimates produced in the original paper by Cummins and Bucca.

Backtesting

Associated Code or Files: "Validation3- Backtester.ipynb", "Main – Full ModelcImplementation.ipynb"

Purpose: Once optimal entry and exit thresholds have been determined, a backtest of the trading strategy must be performed to generate key performance metrics and evaluate results.

A Python backtester function was created for the purpose of testing the model's ability to determine optimal entry and exit thresholds that perform in a simulated trading environment. As the paper did not provide much detail into how it backtested each strategy, an approach was adopted from Ernie Chan's 'Algorithmic Trading'. Bertram's original paper only entered a trade when the process met the entry criteria and exited when the exit criteria was met. It then waited until the process returned to the entry threshold to enter a trade again. Cummins and Bucca extended the strategy to also allow for flipping a long position and shorting when the process met the exit criteria. And finally, to adjust for the non-zero mean in the spread the thresholds were shift by mu – the long term mean of the OU process. This leaves the long/short criteria as:

- Entry criteria (long the spread) = a + mu
- Exit criteria (short the spread) = m + mu

Output and backtest results are produced to be the same as the paper, to compare if generated results are similar. These outputs include – total strategy return, mean daily return, backtested sharpe ratio, trade count and average hold time.

Validation of the backtester involved comparing similar code snippets and manually reviewing output of buy and sell signals. This ensured that the back tester 'traded' as intended and any results produced would be consistent with the backtester function used in the paper.

Results

Associated code or files: "bt_results_1.csv", "bt_results_2.csv", "bt_results_3.csv", "bt_results_4.csv"

Purpose: The final output of the report, reviewed to ensure results are reasonable

Once all the required functions were built, tested and validated, they were iterated through thousands of times to produce the final results of all the backtests.

As the data set for the original paper differs from the data set used in this report, results could not be perfectly replicated. However, results can be compared across sets to see if any similarities exist. Ultimately, with careful consideration certifying each component of the model works as intended, we already have a high degree of confidence that output should be consistent. Therefore, the above sections on calibrating the OU, maximising expected return and backtesting were carefully tested to give a greater degree of confidence in results. The next section of the paper will provide a more in-depth view of the results obtained.

Output

After running the full implementation and backtest of the model on 300 crude and refined product sets across four out-of-sample periods 1200 backtests were conducted. This meant:

- 1200 calibrations of OU process.
- 1200 calculations of optimal entry and exit thresholds.
- 1200 full backtests.

As such, all results cannot be displayed in this single document. However, four results set files have been created as well as four backtest results files.

Result set

The result sets contain all results obtained via calibrating each individual train set. In the file, for each spread, there will be the OU parameters;α, σ& μ, estimated transaction costs c and optimal trade entry threshold a. a sample of the output produced below:

Spread	Transaction Cost	Alpha	mu	sigma	a
('QS5', 'HO4')	0.002482	4.623143957	1.119722074	0.030890517	-0.007946299
('QS5', 'HO5')	0.002482	4.198760288	1.119064327	0.033085418	-0.008530533
('QS4', 'HO4')	0.002482	4.167056926	1.116787081	0.034439167	-0.00876232
('QS5', 'HO3')	0.002482	4.091575629	1.119033841	0.031283531	-0.00830967
('QS4', 'HO3')	0.002482	3.903343275	1.116090712	0.034377868	-0.00892971
('QS4', 'HO5')	0.002482	3.747394012	1.116140185	0.036539154	-0.00938741
('QS3', 'HO4')	0.002482	3.525062282	1.113831346	0.037933803	-0.009789948
('QS3', 'CL5')	0.002482	3.476686171	2.235260513	0.035212491	-0.009390342
('QS2', 'CO5')	0.002482	3.445408499	2.143241816	0.030866898	-0.008685376
('QS3', 'HO3')	0.002482	3.401058134	1.113132431	0.037825621	-0.009881507
('QS2', 'CL5')	0.002482	3.400928286	2.233215115	0.03617722	-0.009613221
('QS4', 'CL5')	0.002482	3.38958534	2.238291509	0.034702631	-0.009379364
('QS3', 'CO5')	0.002482	3.376448921	2.145324455	0.030629706	-0.008698097
('QS1', 'CO5')	0.002482	3.321051481	2.142279047	0.03221826	-0.009017175
('QS3', 'HO5')	0.002482	3.277935553	1.113185794	0.039624185	-0.010287075
('QS5', 'CL5')	0.002482	3.249453026	2.241289456	0.034500925	-0.009468087
('QS1', 'CL5')	0.002482	3.240666547	2.232267603	0.037340827	-0.009950585
('QS4', 'CO5')	0.002482	3.211376669	2.148380067	0.030853307	-0.008871895
('QS5', 'HO2')	0.002482	3.162960531	1.116161038	0.035659918	-0.009743941
('QS4', 'HO2')	0.002482	3.110298716	1.113213368	0.037683312	-0.010135161

Backtest Results

This is the complete final back test results for all 1200 backtests across the four out of sample periods. Each row of the file represents performance measures for a back test measuring:

- Average daily return
- Total strategy return
- Strategy sharpe ratio
- Trade count
- Average hold time (in days)

Sample of each output file below, results will be discussed in a further section.

Discussion & Observations

When comparing my final output with results produced in the paper, there are some notable differences in strategy performance. I have noted in my critique that this may have been due to data snooping bias introduced by the authors. But more simply, the performance may have decayed over time as crude markets have become more efficient and simple alpha strategies no longer work. What is evident though is the significant drop in the performance of the backtests.

Interestingly, the final backtest from the paper in 2010 also happened to be the worst performing. Average trade length was 4-5 times higher and average daily return had dropped noticeably (77% decrease). This sharp reduction in performance may have continued leaving results obtained in early 2000's to be unattainable in modern oil and refined markets.

Trade count / average hold time

The first major point to note is the average trade length is far higher than that of the paper. When I allowed for data snooping, I was able to produce a much lower average hold time due to being able to overfit my data. But with a train test split, the hold time was far longer and average trade count was particularly low. This overall made the strategies perform poorly as the model ideally should have transitioned more between entry and exit thresholds.

Poor performance

Overall performance failed to meet expectations. Most notably, sharpe ratios were fair lower than that reproduced in the paper. In 2009 the average sharpe ratio of all profitable strategies in the paper was 2.43. Out of all the 1200 backtests, I was unable to produce a sharpe ratio greater than 1.39. As noted above, this performance may be in part due to alpha decay.

Consistent across back tests

On a positive note, there seems to be a fair amount of consistency across each four backtest batches. There are no large outliers and an even distribution of results. Unfortunately, as the authors of the original paper never showed any distributions of results or results for unprofitable strategies, I was unable to compare. But overall, each backtest is not too dissimilar from one another which was very promising.

Associated Files

Due the large amounts of output produced by the backtests, output has been arranged into 8 csv files. There a four parameter estimate files corresponding to the four test sets and four backtest result files containing the collated backtested results from 1200 backtests. High level summary of the backtests will be displayed below.

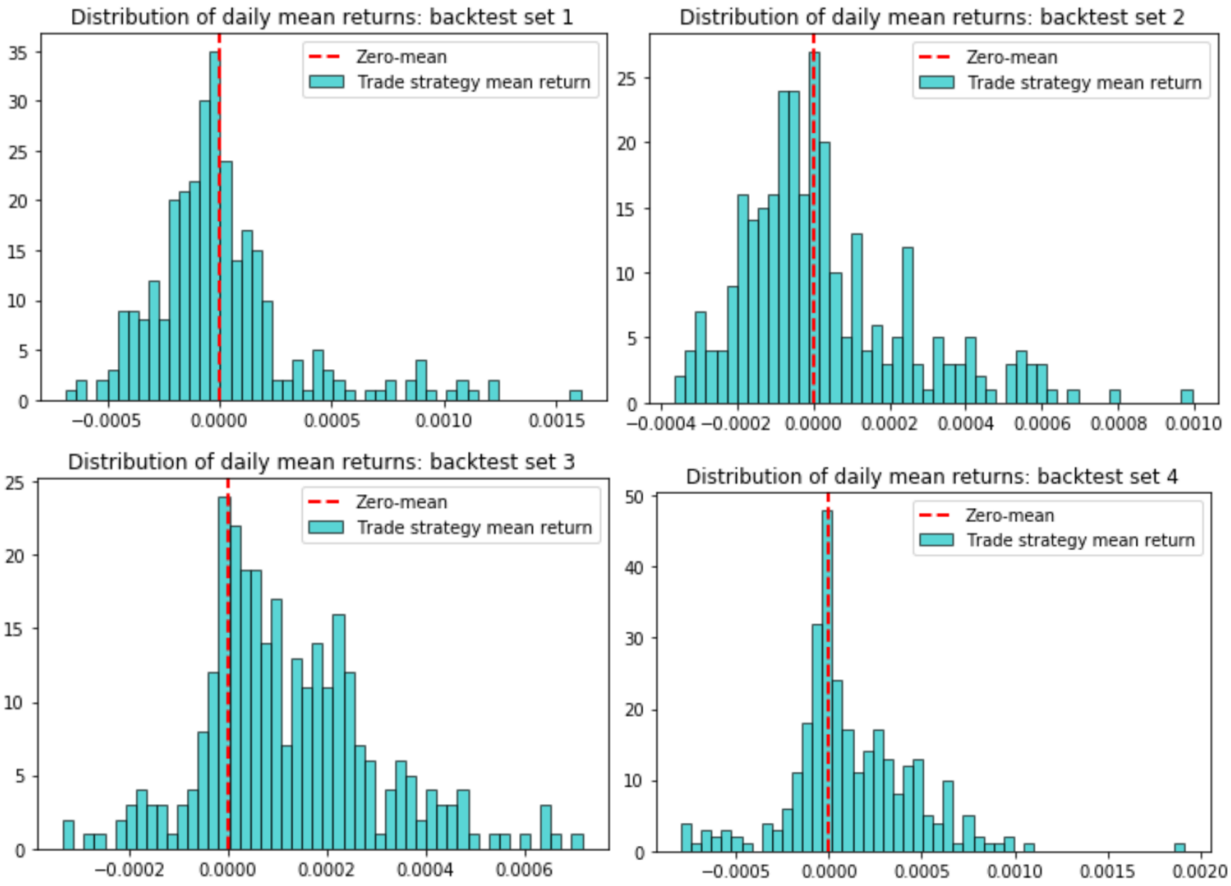
Parameter Estimates	Backtest Results
"results_set_1.csv"	"bt_results_1.csv"
"results_set_2.csv"	"bt_results_2.csv"
"results_set_3.csv"	"bt_results_3.csv"
"results_set_4.csv"	"bt_results_4.csv"

Backtest Results Output

Performance Summary	Backtest Results Set 1	Backtest Results Set 2	Backtest Results Set 3	Backtest Results Set 4
Number positive strategies	119	136	206	142
% positive	39.67%	45.33%	68.67%	47.33%
Average daily mean	0.000507%	0.002555%	0.012203%	0.011393%
Average sharpe	-0.0556764	-0.001903	0.182018	0.07613
Average total return	0.30499%	1.53550%	7.33240%	6.79333%
Average trade count	6.46	3.54	2.55	6.00
Average hold time	146.388	234.3	207.77	159.25

All result sets produce an average daily mean greater than zero across the backtests. However, most have more negetive strategies than positive showing there is a right skew in the data, where top strategies tend to outperform the rest. Backtest 3 and 4 are the most successful, producing average total returns far greater than that of 1 and 2, while also producing an average sharpe ratio greater than zero across all backtests.

The distributions of daily mean returns for each backtests set further confirm these observations:



Further results for the top performing strategies in each backtest set below. A full set of the performance of all 1200 backtests are available in backtest files listed above.

Backtest set 1 – results top 20 performing strategies

Strategy Pair	Strategy Sharpe Ratio	Average Daily Return	Total Strategy Return	Trade Count
('QS5', 'HO4')	1.389337407	0.12192%	73.27413%	41
('XB5', 'XB1')	1.374153601	0.16067%	96.56146%	5
('QS4', 'HO3')	1.335757612	0.12130%	72.90368%	41
('QS5', 'HO3')	1.259989973	0.11365%	68.30164%	47
('QS3', 'HO2')	1.1466064	0.10871%	65.33178%	37
('QS4', 'HO2')	1.102281358	0.10388%	62.43195%	42
('QS2', 'CO5')	1.060300201	0.10712%	64.37908%	38
('QS4', 'HO4')	1.024456741	0.09097%	54.67159%	23
('QS2', 'HO2')	0.904776592	0.08671%	52.11063%	31
('QS2', 'CO4')	0.870683755	0.08923%	53.62761%	28
('QS1', 'CO4')	0.830141359	0.09259%	55.64520%	36
('QS3', 'HO3')	0.819181541	0.07527%	45.23991%	27
('QS5', 'HO2')	0.816575769	0.07673%	46.11299%	27
('QS1', 'HO1')	0.779143163	0.09032%	54.28458%	26
('QS3', 'CO5')	0.729080355	0.07245%	43.54125%	26
('QS2', 'HO3')	0.721128649	0.06724%	40.41347%	21
('QS5', 'XB1')	0.606981564	0.08800%	52.88865%	11
('QS4', 'XB1')	0.587203805	0.08565%	51.47397%	11
('QS1', 'CO3')	0.521438774	0.05889%	35.39439%	28
('QS3', 'HO4')	0.504375639	0.04544%	27.30668%	11

Backtest Set 2 – results top 20 performing strategies

Strategy Pair	Strategy Sharpe Ratio	Average Daily Return	Total Strategy Return	Trade Count
('XB5', 'CO5')	1.211319363	0.07911%	47.54465%	5
('QS4', 'XB2')	1.013172595	0.09931%	59.68823%	11
('XB5', 'CL3')	0.90575257	0.06390%	38.40613%	6
('XB5', 'CO4')	0.893594509	0.05876%	35.31490%	3
('XB5', 'CO3')	0.885148199	0.05880%	35.34072%	3
('XB5', 'CO2')	0.861025105	0.05821%	34.98721%	3
('XB5', 'CL2')	0.831602013	0.06057%	36.40378%	3
('XB5', 'CL4')	0.827393496	0.05709%	34.31271%	5
('XB4', 'CO3')	0.81897673	0.05340%	32.09576%	3
('XB4', 'CO4')	0.813202103	0.05287%	31.77306%	3
('XB5', 'CO1')	0.799160302	0.05589%	33.59258%	3
('XB5', 'CL5')	0.767965255	0.05216%	31.34775%	5
('QS3', 'XB2')	0.697563699	0.06880%	41.35091%	11
('XB5', 'HO2')	0.676662995	0.04398%	26.43125%	1
('XB4', 'CL5')	0.656767187	0.04460%	26.80514%	7
('XB5', 'HO3')	0.633750034	0.04119%	24.75661%	1
('XB2', 'CL1')	0.631750427	0.05416%	32.55038%	3
('XB5', 'HO1')	0.629537495	0.04205%	25.26962%	1
('QS3', 'XB4')	0.61314368	0.05275%	31.70323%	5
('XB5', 'HO4')	0.61223569	0.04034%	24.24487%	1

Backtest set 3 – results top 20 performing strategies

Strategy Pair	Strategy Sharpe Ratio	Average Daily Return	Total Strategy Return	Trade Count
('HO5', 'XB2')	0.811971776	0.06538%	39.29084%	6
('XB3', 'CL5')	0.805644353	0.05475%	32.90528%	9
('HO4', 'XB2')	0.740416391	0.06036%	36.27541%	6
('XB3', 'CO5')	0.707306409	0.04625%	27.79578%	6
('HO3', 'XB5')	0.698932556	0.04337%	26.06608%	2
('XB3', 'CL2')	0.691617294	0.05053%	30.36699%	11
('XB3', 'CL4')	0.659711596	0.04545%	27.31753%	9
('HO4', 'XB5')	0.650879865	0.04025%	24.19249%	2
('QS5', 'XB2')	0.648949472	0.07172%	43.10224%	6
('XB3', 'CL1')	0.628297812	0.04784%	28.75241%	11
('QS2', 'XB2')	0.591419166	0.06639%	39.89979%	4
('QS4', 'XB2')	0.587953594	0.06512%	39.13933%	6
('XB3', 'HO1')	0.583675977	0.04272%	25.67619%	1
('HO4', 'XB4')	0.582033332	0.03794%	22.80161%	2
('HO5', 'XB5')	0.571627021	0.03537%	21.25737%	2
('QS1', 'XB2')	0.547966465	0.06335%	38.07216%	4
('XB3', 'HO2')	0.547557944	0.04019%	24.15216%	1
('XB5', 'CO1')	0.538977437	0.03412%	20.50800%	2
('HO1', 'CL5')	0.527354371	0.02667%	16.03050%	0
('HO5', 'HO1')	0.526869392	0.00869%	5.22367%	1

Backtest set 4 – results top 20 performing strategies

Strategy Pair	Strategy Sharpe Ratio	Average Daily Return	Total Strategy Return	Trade Count
('CO1', 'CL2')	1.058200724	0.19131%	115.93546%	16
('CO2', 'CL3')	0.798630104	0.07239%	43.86561%	12
('XB5', 'CO2')	0.743036871	0.08503%	51.52963%	7
('XB4', 'CO1')	0.728576809	0.09734%	58.99038%	7
('QS3', 'HO2')	0.728471268	0.06609%	40.04808%	13
('QS1', 'HO1')	0.652238801	0.07239%	43.87080%	17
('QS2', 'HO2')	0.618513101	0.05827%	35.31207%	11
('XB5', 'CO3')	0.604562638	0.06445%	39.05592%	5
('XB1', 'CO1')	0.596919999	0.10959%	66.41432%	16
('HO3', 'XB5')	0.58227016	0.06519%	39.50401%	9
('QS1', 'CO4')	0.573835524	0.07291%	44.18335%	11
('HO2', 'XB5')	0.552831071	0.06393%	38.73899%	9
('XB5', 'CL3')	0.536754687	0.07555%	45.78356%	9
('CO4', 'CL5')	0.528526307	0.03074%	18.62721%	9
('QS2', 'CO2')	0.523974905	0.06657%	40.34386%	11
('QS4', 'HO3')	0.521900004	0.04495%	27.23758%	15
('XB4', 'CL3')	0.504283693	0.07332%	44.43039%	12
('XB4', 'CO2')	0.499465001	0.05879%	35.62943%	5
('QS1', 'CL3')	0.479258531	0.08223%	49.82867%	6
('QS1', 'XB2')	0.475962036	0.09305%	56.38804%	11

Conclusion

After careful implementation and validation of results, I believe I successfully replicated Cummins and Bucca's quantitative spread trading of crude oil and refined products. As discussed, while my final results were not as successful as the original paper, this was not due to poor or improper implementation. This study was able to perform 1200 full back tests across a wide range of oil spreads including calendar, crack and locational spreads. My results, as a whole, underperformed the original paper. However, a number of successful strategies were produced with each set of backtests producing an average daily mean greater than zero. Some performers in each backtest set managed to achieve sharpe ratios greater than one, which is far from the original but still shows some promising results that could be considered for a live trading implementation.

References

Bertram, W. 2009, 'Optimal trading strategies for Itô diffusion processes', *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 14, pp. 2865-2873, viewed 10 February 2020, <https://doi.org/10.1016/j.physa.2009.04.004>.

Bertram, W. 2010, 'Analytic solutions for optimal statistical arbitrage trading', *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 11, pp. 2234-2243, viewed 10 February 2020, <https://doi.org/10.1016/j.physa.2010.01.045>.

Chan, E. 2013, *Algorithmic Trading: Winning Strategies and Their Rationale*, Wiley, New Jersey, USA.

Clegg, M. & Krauss, C. 2018, 'Pairs trading with partial cointegration', *Quantitative Finance*, vol. 18, no. 1, pp. 121-138, viewed 10 February 2020, <https://doi.org/10.1080/14697688.2017.1370122>.

Cummins, M. & Bucca, A. 2012, 'Quantitative spread trading on cruise oil and refined products market', *Quantitative Finance*, vol. 12, no. 12, pp. 1857-1875, viewed 10 February 2020, <https://doi.org/10.1080/14697688.2012.715749>.

Hilpisch, Y. 2020, *Python for Algorithmic Trading*, O'Reilly Media, Inc, California, USA.

Mejía Vega, C.A. 2018, 'Calibration of the exponential Ornstein-Uhlenbeck process when spot prices are visible through the maximum log-likelihood method. Example with gold prices', *Advances in Difference Equations*, vol. 2018, no. 1, pp. 1-14, viewed 10 February 2020, <https://doi.org/10.1186/s13662-018-1718-4>.

Stübinger, J. & Endres, S. 2018, 'Pairs trading with a mean-reverting jump–diffusion model on high-frequency data', *Quantitative Finance*, vol. 18, no. 10, pp. 1735-1751, viewed 10 February 2020, <https://doi.org/10.1080/14697688.2017.1417624>.

Wu, J. 2015, *A pairs trading strategy for goog/googl using machine learning*, Stanford, viewed 10 February 2020, <http://cs229.stanford.edu/proj2015/028_report.pdf>.