
# QF621 Quantitative Trading Strategies


Brandon Lee
Dani Surya Pangestu
Gabriel Woon
Chan Pak Lam
Peter Chettiar
Wong Yong Wen

# Pairs Trading Strategy Using Machine Learning

## Contents

- Problem Statement
- Research Methodology
- Pairs Selection
- Trading Models
  1. OLS
  2. R-OLS
  3. Kalman Filter with Bollinger Bands
  4. Support Vector Machine with Markov Regime Switching


# Problem Statement

# Problem Statement

Implement machine learning in the entire process of the pairs trading framework

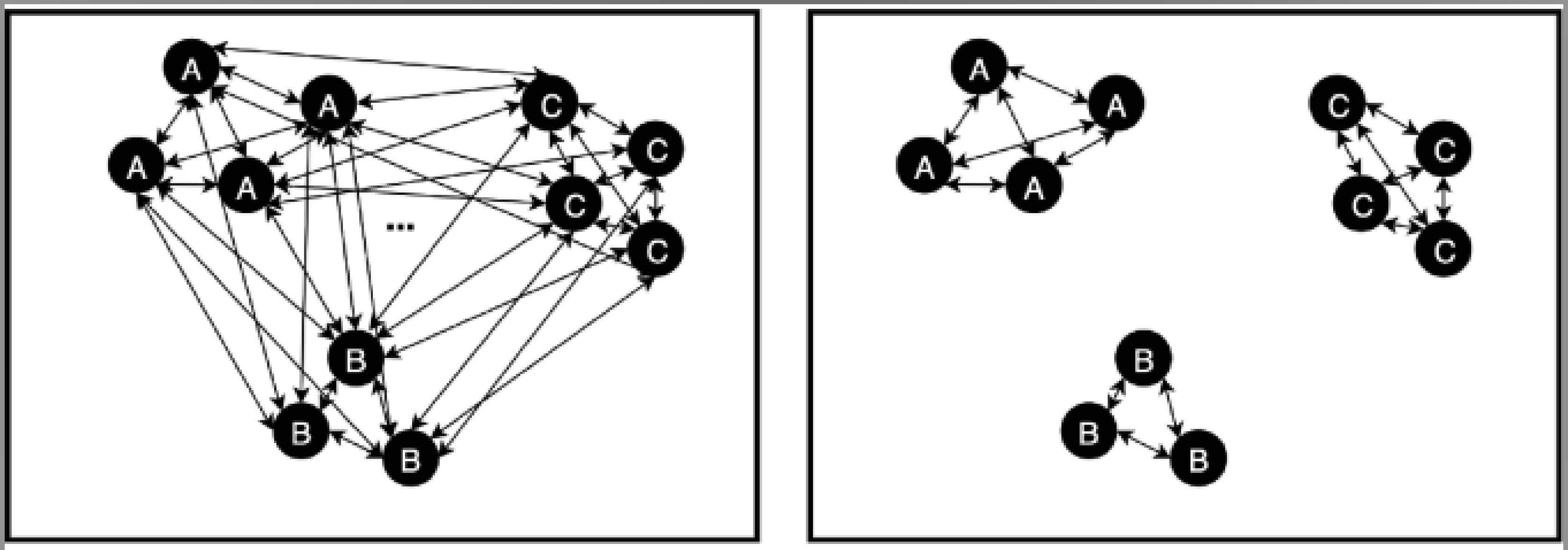
- Pairs Selection
  - Does it outperform the rational/fundamental selection process?
- Trading Model
  - Does the more complex algorithms produce better results?


# Research Methodology

# Research Methodology (Futures Products)

| Metals | Chemical | Energy | Agriculture |
|---|---|---|---|
| Steel | Ethanol | Gas | Soy |
| Non-Ferrous Metal | Fertilizer | Crude Oil | Corn |
| Precious Metal | | Fuel Oil | Wheat |
| | | Heating Oil | Rice |
| | | Diesel | Milk |
| | | Electricity | Steers |
| | | Jet Fuel | Hog |
| | | Sulphur | Wool |
| | | Naphtha | Palm Oil |
| | | Propane | Sugar |
| | | | Cocoa |
| | | | Cotton |

Number of Commodities: 66
Number of Pairs: 2,145




# Pairs Selection (Clustering)

# Rationale

The simplest procedure commonly applied is to generate all possible candidate pairs by considering the combination from every security to every other security in the dataset.

Two different problems arise immediately:

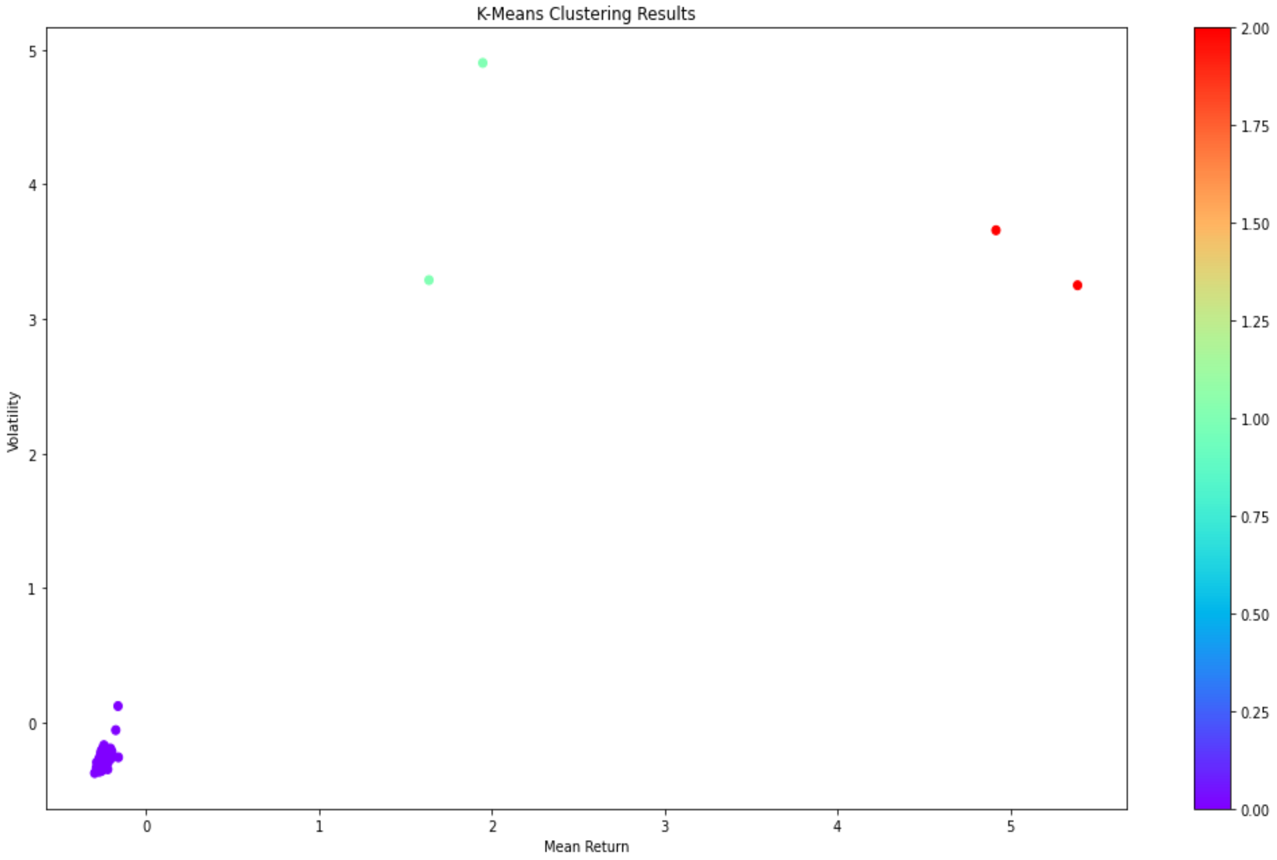
1. First, the computational cost of testing mean-reversion for all the possible combinations increases drastically as more securities are considered.
2. The second emerging problem is frequent when performing multiple hypothesis tests at once and is referred to as the multiple comparisons problem. If 100 hypothesis tests are performed (with a confidence level of 5%) the results should contain a false positive rate of 5%.

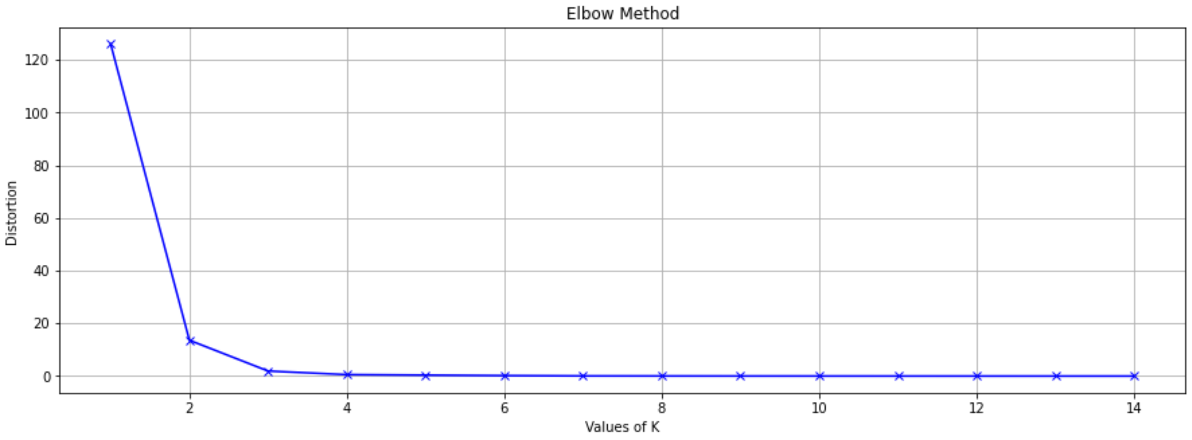
# Clustering Methodology

Pairs Selection Period → 01/01/2009 → 31/12/2017

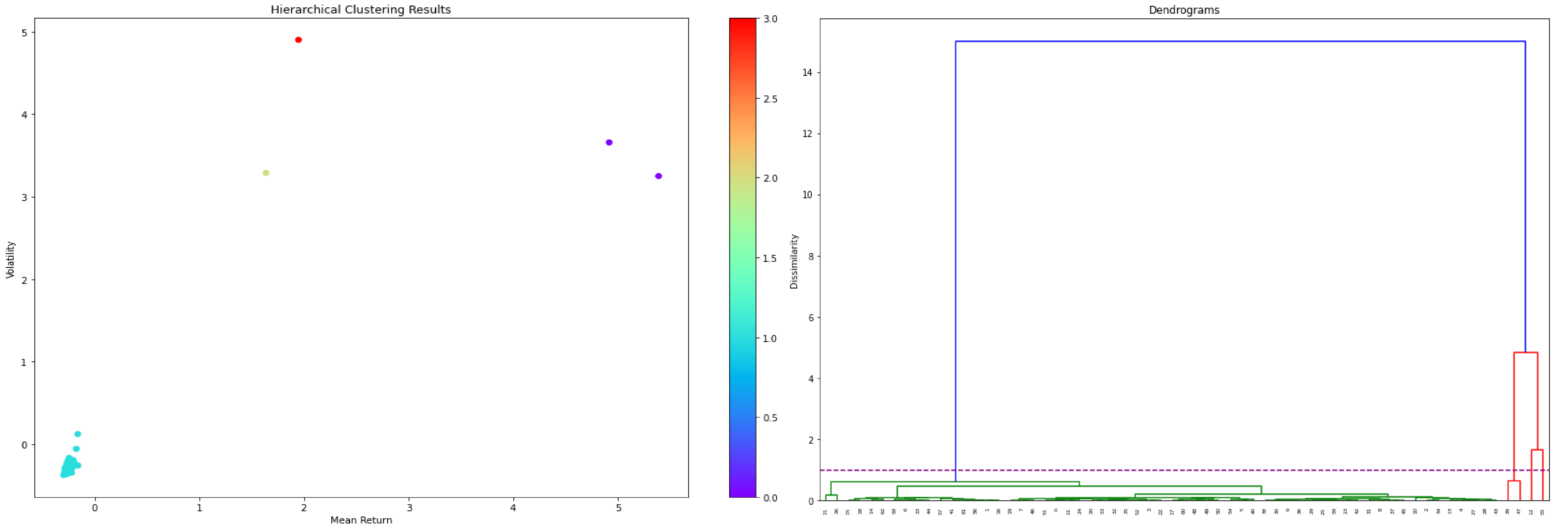
- Unsupervised learning clustering techniques can solve the task at hand:
  - K-Means
  - Hierarchal
  - Affinity Propagation

# K-Means Clustering

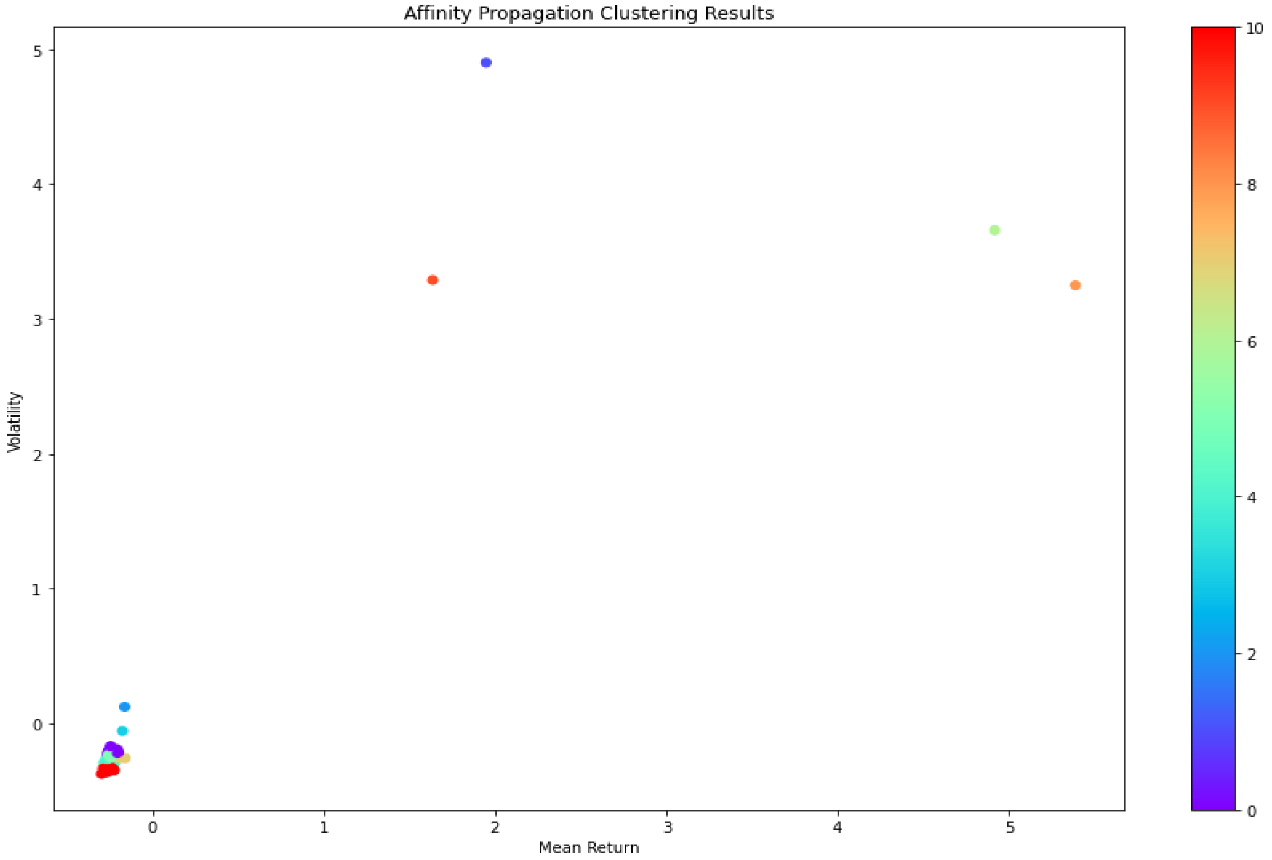


# Hierarchical Clustering



# Affinity Propagation Clustering



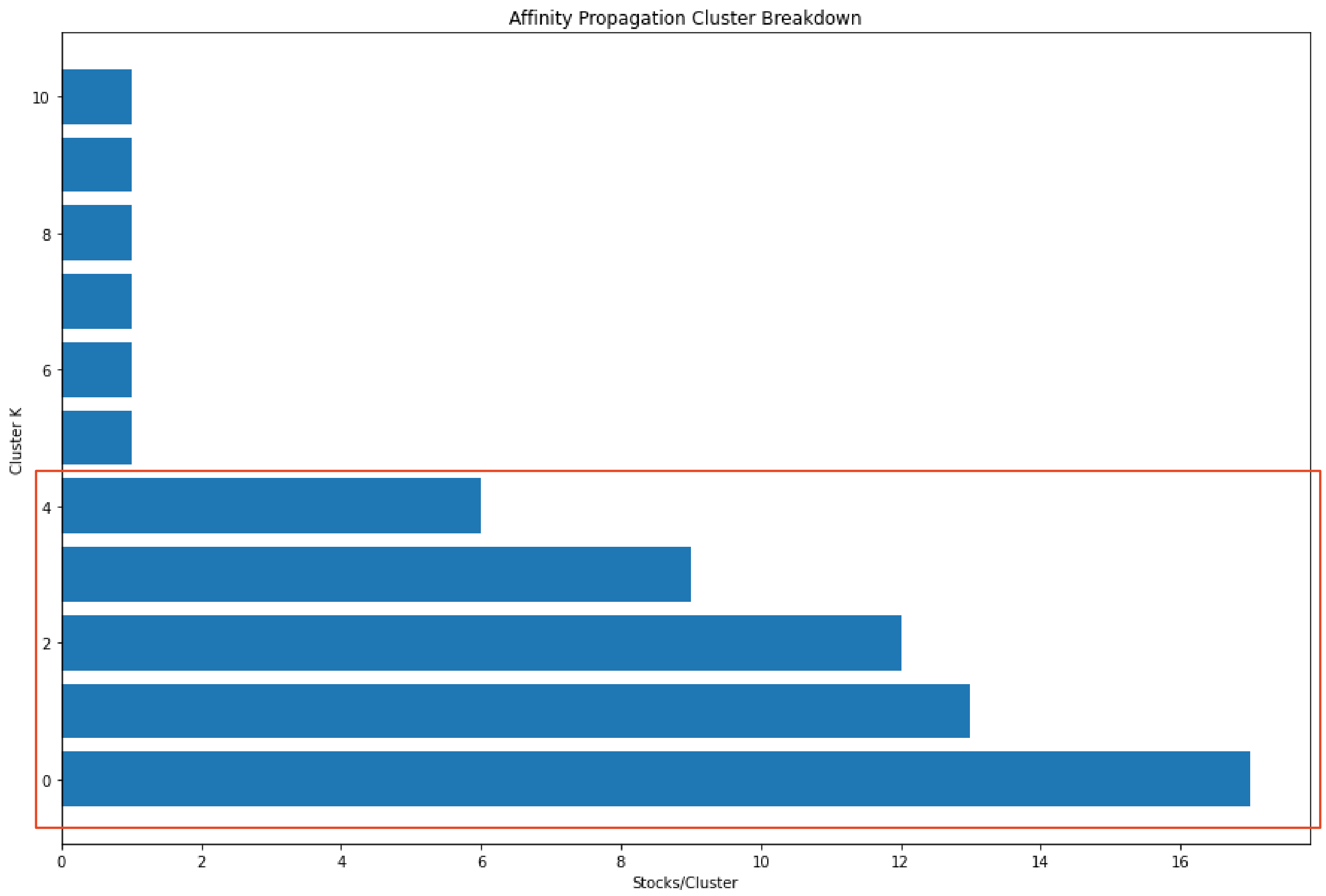
# Clustering Evaluation

Silhouette Score:

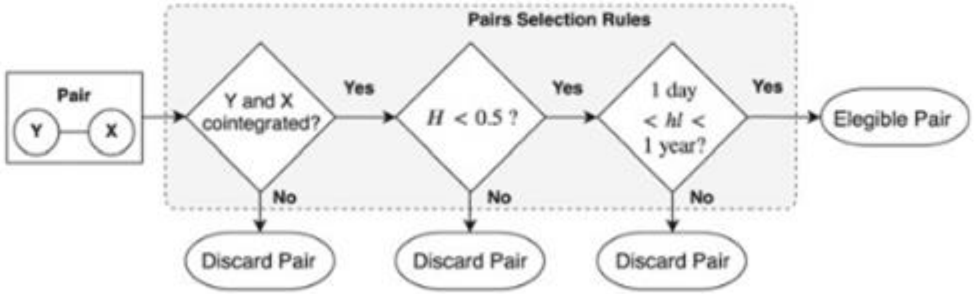
- K-Means Clustering
  0.29743
- Hierarchical Clustering
  0.30728
- Affinity Propagation Clustering
  0.94311

**Number of Clusters: 5**

**Number of Pairs: 662**

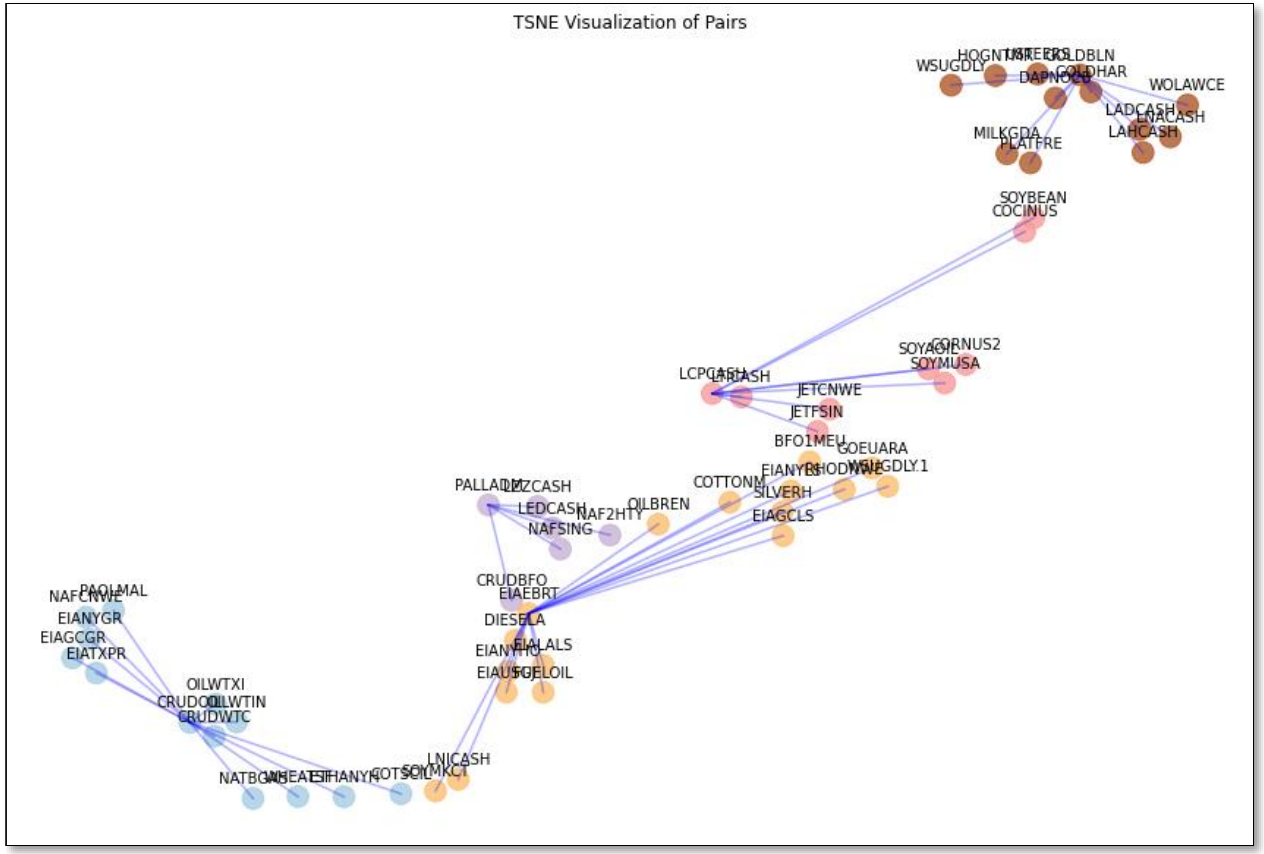


# Absolute Rules of Disqualification (ARODs)



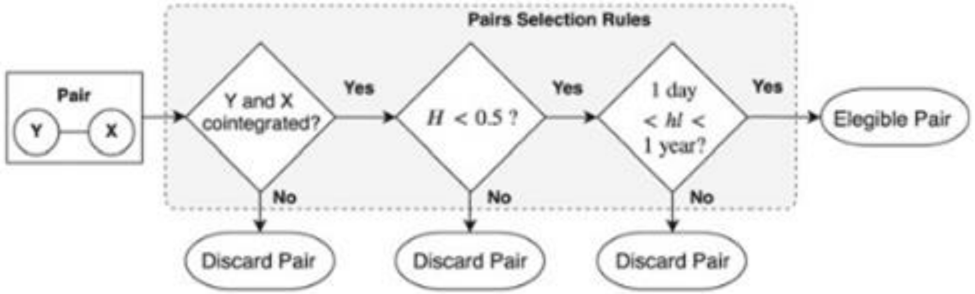

Step 1 - the selection process starts with the testing of pairs, generated from the clustering step, for cointegration using the Engle-Granger test. This cointegration testing method was chosen based on its simplicity.

# Absolute Rules of Disqualification (ARODs)




Number of Pairs: 52

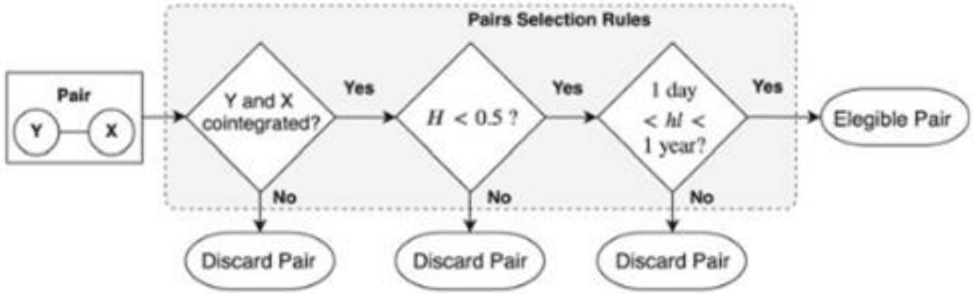
# Absolute Rules of Disqualification (ARODs)




Step 2 - a validation step is implemented to provide more confidence in the mean-reversion character of the pairs' spread. The condition imposed is that the Hurst exponent associated with the spread of a given pair is enforced to be smaller than 0.5, assuring the process leans towards mean-reversion. The Hurst exponent is a measure that quantifies the amount of memory in a specific time series, which in this case means, how mean-reverting is the process under analysis.

```python
def get_hurst_exponent(time_series, max_lag=20):
    """Returns the Hurst Exponent of the time series"""

    lags = range(2, max_lag)

    # variances of the lagged differences
    tau = [np.std(np.subtract(time_series[lag:], time_series[:-lag])) for lag in lags]

    # calculate the slope of the log plot -> the Hurst Exponent
    reg = np.polyfit(np.log(lags), np.log(tau), 1)

    return reg[0]
```

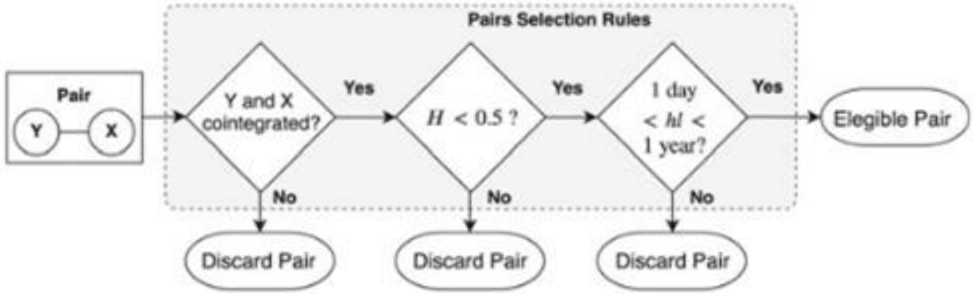
# Absolute Rules of Disqualification (ARODs)




Step 3 - the pair's spread movement is constrained using the half-life of the mean-reverting process. The half-life is the time that the spread will take to mean-revert half of its distance after having diverged from the mean of the spread, given a historical window of data. For medium-term price movements, the spreads that have very short (< 7 day) or very long mean-reversion (> 365 days) periods are not suitable.

```python
def apply_half_life(time_series):
    lag = np.roll(time_series, 1)
    lag[0] = 0
    ret = time_series - lag
    ret[0] = 0

    # adds intercept terms to X variable for regression
    lag2 = sm.add_constant(lag)

    model = sm.OLS(ret, lag2)
    res = model.fit()

    half_life = -np.log(2) / res.params[1]

    return half_life
```

# Absolute Rules of Disqualification (ARODs)

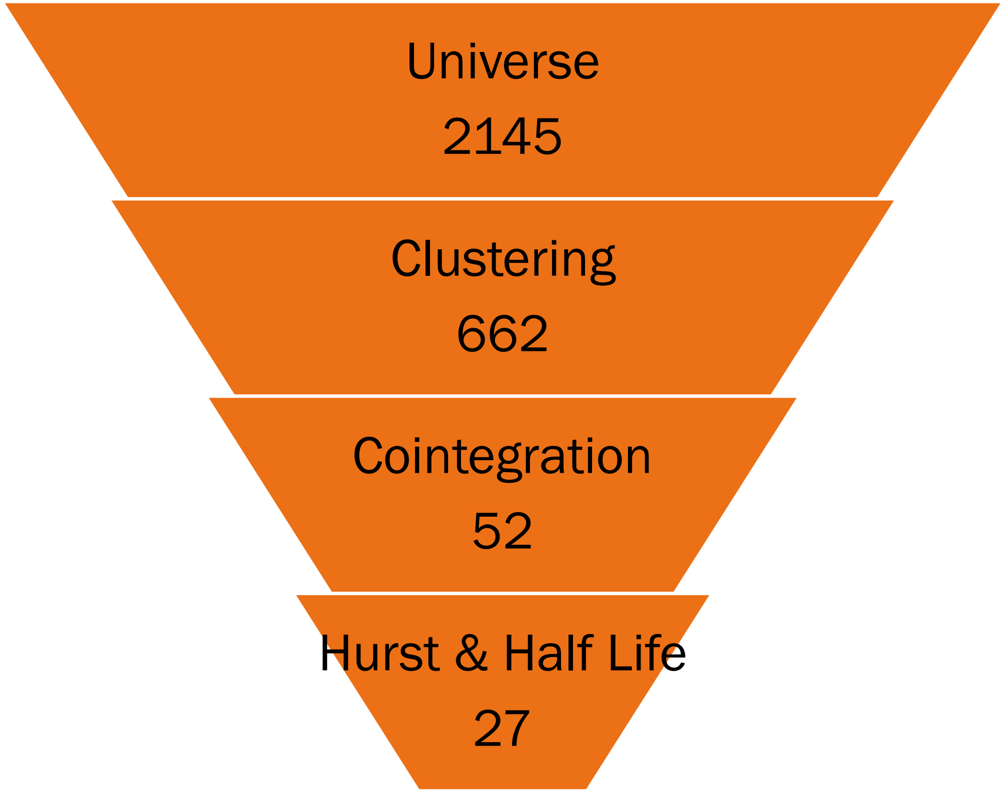



Template Example

| | Leg1 | Leg2 | Hedge Ratio | Hurst Exponent | Half Life |
|---|---|---|---|---|---|
| 0 | EIAEBRT | OILBREN | 0.99142 | 0.10075692 | 2.774739157 |
| 1 | EIAEBRT | EIAUSGJ | 0.02592675 | 0.25199634 | 19.14888229 |
| 2 | EIAEBRT | BFO1MEU | 0.94940347 | 0.27979398 | 8.413260033 |

**Number of Pairs: 27**

# Final Pairs




| | |
|---|---|
| CRUDOIL-EIANYGR | EIAEBRT-OILBREN |
| CRUDOIL-EIAGCGR | EIAEBRT-EIAUSGJ |
| CRUDOIL-OILWTXI | EIAEBRT-BFO1MEU |
| GOLDBLN-GOLDHAR | EIAEBRT-GOEUARA |
| LCPCASH-JETCNWE | EIAEBRT-EIALALS |
| LCPCASH-LTICASH | EIAEBRT-EIANYHO |
| PALLADM-LZZCASH | EIAEBRT-EIANYLS |
| LNICASH-RHODNWE | EIAEBRT-EIAGCLS |
| LCPCASH-GOEUARA | EIAEBRT-DIESELA |
| WHEATSF-NAFCNWE | EIAEBRT-FUELOIL |
| WHEATSF-EIANYGR | CRUDOIL-OILWTIN |
| WHEATSF-EIAGCGR | CRUDOIL-NAFCNWE |
| WHEATSF-NATBGAS | CRUDOIL-CRUDWTC |
| | CRUDOIL-ETHANYH |


# Trading Models

# Trading Model 1
# Ordinary Least Squares (OLS)

***Trading parameters*:**

- Training period from 2017-01-01 to 2017-12-31 (1 year)
- Test period from 2018 onwards till Apr 2022 (~4.33 years)
- Static Hedging (using the same hedge ratio throughout the testing period)
  - Long/Short the spread if the spread is smaller/greater than the average spread volatility derived in training period
  - If the spread is greater/smaller than 2.5x spread vol, we will liquidate our position (part of our risk management measure)

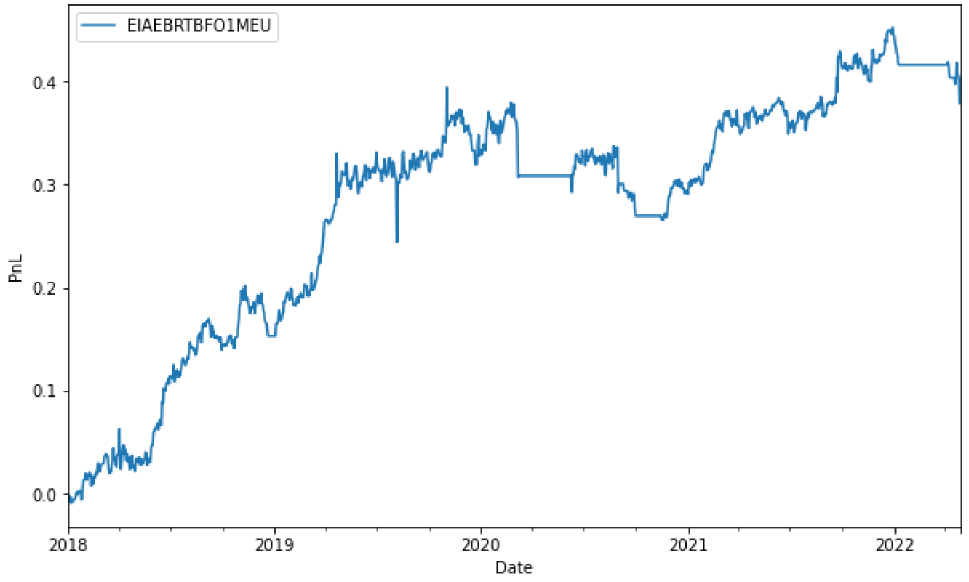
# Trading Model 1
# Ordinary Least Squares Result

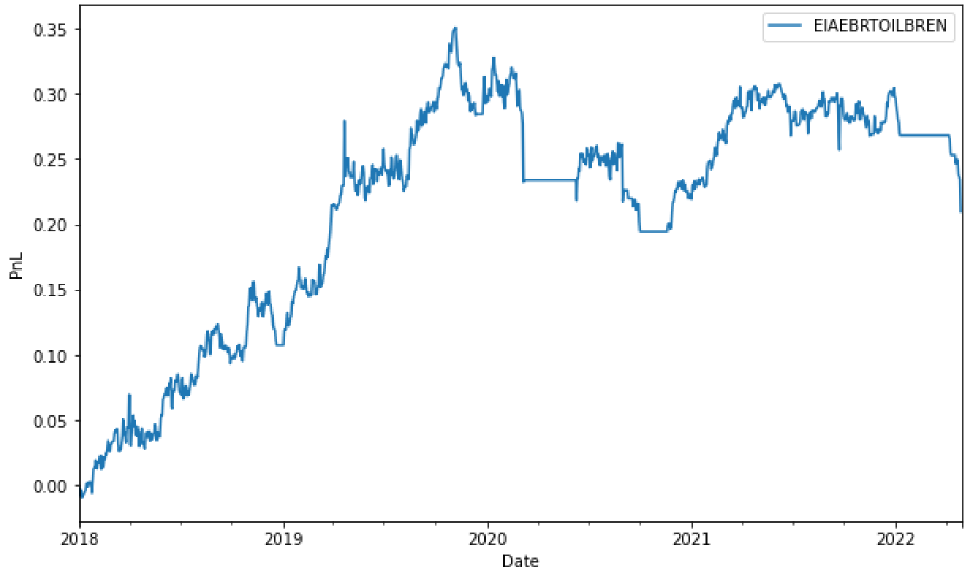
Out of <u>27</u> pairs identified, **only** <u>2</u> pairs gave positive PnL result.

| Pair Name | Cum. Return | Sharpe ratio | Sortino ratio | Max drawdown |
|---|---|---|---|---|
| EIAEBRT-OILBREN | 0.209887 | 0.46816 | 0.04090 | -0.14966 |
| EIAEBRT-BFO1MEU | 0.378797 | 0.75566 | 0.06808 | -0.12654 |

| Name | Type | Sub-Type | Ticker |
|---|---|---|---|
| Europe Brent Spot FOB U$/BBL Daily | Energy | Crude Oil | EIAEBRT |
| Crude Oil BFO M1 Europe FOB $/BBl | Energy | Crude Oil | OILBREN |
| Crude Oil BFO M1 Europe FOB $/Bbl | Energy | Crude Oil | BFO1MEU |

Fundamentally, both pairs are in the same industry (Energy- Crude Oil)



# Trading Model 2
# Rolling OLS (R-OLS)

***<u>Trading parameters</u>*:**

- OLS regression run based on rolling 252 trading days
  - Dynamic hedge ratio
- Long/Short the spread if the spread is smaller/greater than the rolling spread volatility (based on past 252 days)
- If the spread is greater/smaller than 2.5x spread vol, we will liquidate our position (part of our risk management measure)

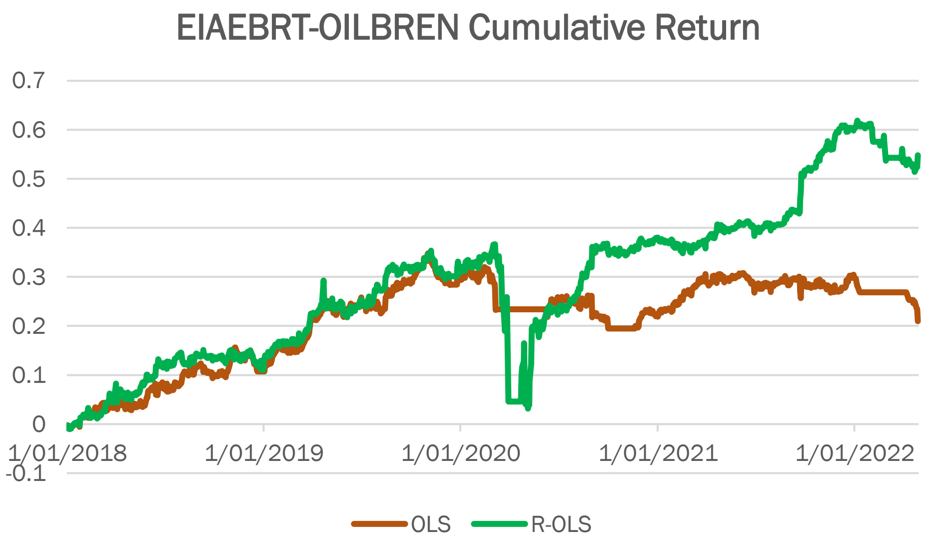
# Compare OLS vs Rolling OLS (R-OLS)

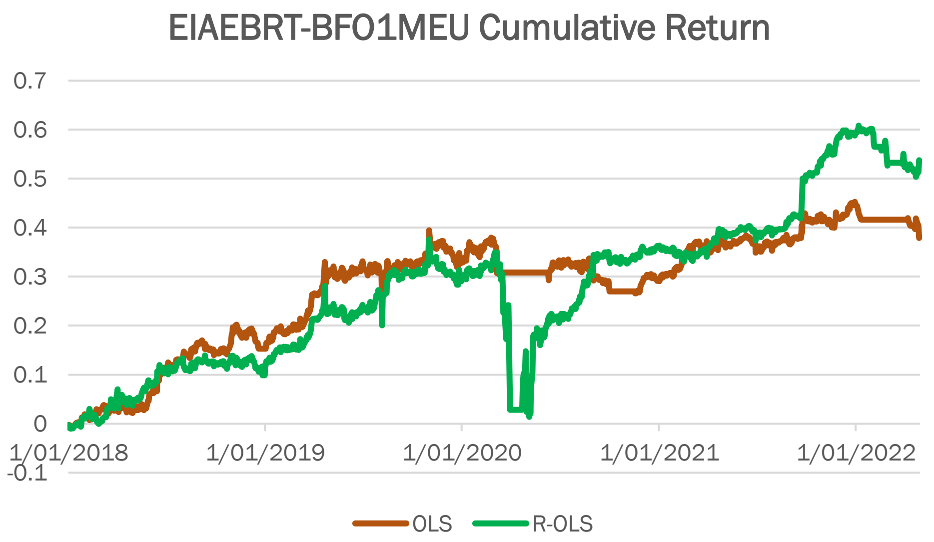
**OLS** :Out of 27 pairs identified, **only** 2 pairs produced positive PnL
**Rolling OLS**: Out of 27 pairs identified, 20 pairs produced positive PnL

| Pair Name | Cum. Return | Sharpe ratio | Sortino ratio | Max drawdown |
|---|---|---|---|---|
| EIAEBRT-OILBREN (OLS) | 0.209 | 0.468 | 0.041 | -0.149 |
| EIAEBRT-OILBREN (R-OLS) | 0.548 | 0.716 | 0.0560 | -0.313 |
| EIAEBRT-BFO1MEU OLS | 0.378797 | 0.75566 | 0.06808 | -0.12654 |
| EIAEBRT-BFO1MEU (R-OLS) | 0.537 | 0.674 | 0.0532 | -0.333 |

| Name | Type | Sub-Type | Ticker |
|---|---|---|---|
| Europe Brent Spot FOB U$/BBL Daily | Energy | Crude Oil | EIAEBRT |
| Crude Oil BFO M1 Europe FOB $/BBl | Energy | Crude Oil | OILBREN |
| Crude Oil BFO M1 Europe FOB $/Bbl | Energy | Crude Oil | BFO1MEU |

Fundamentally, both pairs are in the same industry (Energy- Crude Oil)



# Trading Model 3
# Kalman Filter with Bollinger Bands

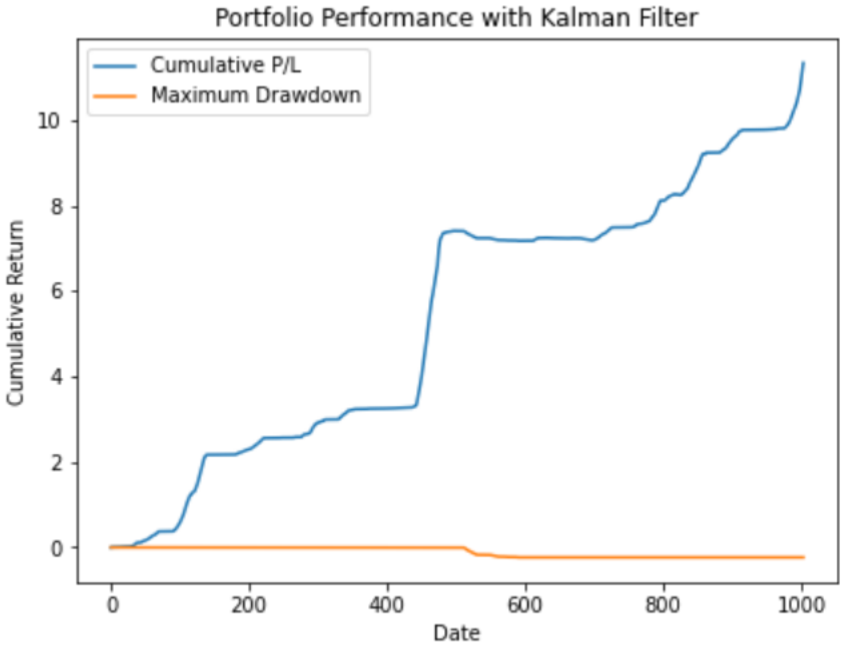
## ***Why Kalman Filters are important?***

It removes **arbitrariness** of other methods.

## ***Trading Parameters:***

- Trading Period of 1st January of 2018 to 29th April 2022
- Equal allocation of capital for each pairs
- Daily stop loss of 5%
- We will use Bollinger Band to make decision to enter or exit the trade.
  - Using the data of 125 days before the trade
  - Short the spread when it exceeds the mean +1.5 Standard Deviation

# Trading Model 3
# Kalman Filter with Bollinger Bands




| Profitable Pairs |
|---|
| EIAEBRT-OILBREN |
| EIAEBRT-EIAUSGJ |
| EIAEBRT-BFO1MEU |
| EIAEBRT-EIALALS |
| EIAEBRT-EIANYHO |
| EIAEBRT-EIANYLS |
| EIAEBRT-EIAGCLS |
| EIAEBRT-DIESELA |
| EIAEBRT-FUELOIL |
| CRUDOIL-NAFCNWE |
| CRUDOIL-ETHANYH |
| CRUDOIL-EIANYGR |
| CRUDOIL-EIAGCGR |
| CRUDOIL-OILWTXI |
| LCPCASH-JETCNWE |
| LCPCASH-GOEUARA |
| WHEATSF-NAFCNWE |
| WHEATSF-EIANYGR |
| WHEATSF-EIAGCGR |

- **19 pairs** provide a positive PnL using Kalman Filter.

| | |
|---|---|
| **Cum.Return** | 958.8% |
| **Sharpe Ratio** | 0.445 |
| **Sortino Ratio** | 7.51 |
| **Max. Drawdown** | 2.78% |

# Trading Model 4
# Support Vector Machine with Markov Regime Switching

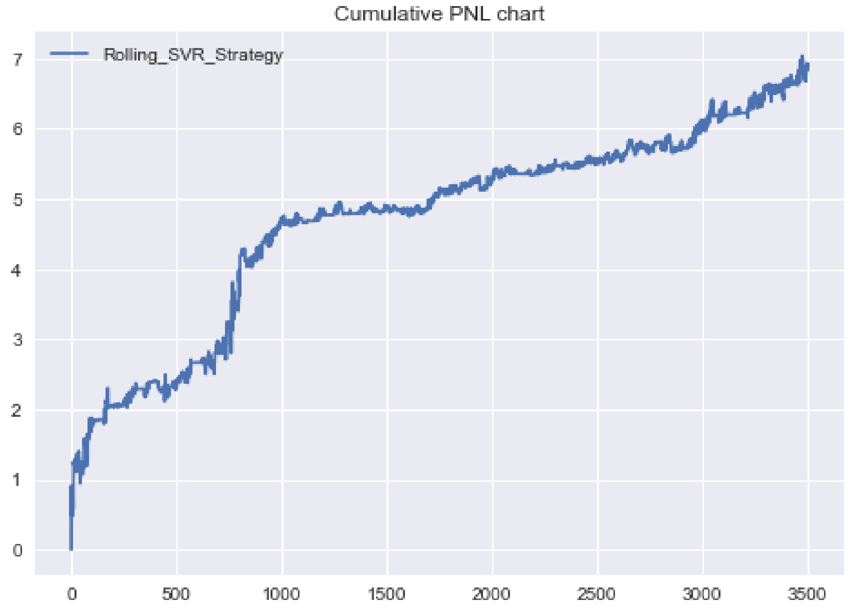
Identify change in regime using Markov switching regime model

| Final Pairs |
|---|
| EIAEBRT-OILBREN |
| EIAEBRT-BFO1MEU |
| EIAEBRT-GOEUARA |
| EIAEBRT-EIALALS |
| EIAEBRT-DIESELA |
| CRUDOIL-NAFCNWE |
| LCPCASH-JETCNWE |
| LCPCASH-LTICASH |
| PALLADM-LZZCASH |
| LNICASH-RHODNWE |
| LCPCASH-GOEUARA |
| WHEATSF-NAFCNWE |
| WHEATSF-EIANYGR |
| WHEATSF-EIAGCGR |
| WHEATSF-NATBGAS |

Generate dynamic hedge ratio using Support Vector Regression, 2-step Engle-Granger method

# Trading Model 4
# Support Vector Machine with Markov Regime Switching

| Trade Position | Trade Position (numerical) | PnL |
|---|---|---|
| BUY | 1 | (100∗price_S2) − (100∗hedge_ratio∗price_S1) |
| SELL | -1 | (100∗hedge_ratio∗price_S1)−(100∗price_S2) |
| Hold | 0 | NA = 0 |




| Performance Metrics | |
|---|---|
| Cumulative Return | 685% |
| Sharpe Ratio | 0.5833 |
| Max Drawdown | -51.91% |

# Conclusion

| Component | Complexity | Machine Learning Method | Success Factor | Proof of Concepts |
|---|---|---|---|---|
| Pairs Selection | Medium | Clustering | Reduced unique pairs from 2145 to 27 | ✓ |
| Trading Model 1 | Low | Ordinary Least Square | 2 positive PnL with very low cumulative return | χ |
| Trading Model 2 | Low | Rolling Ordinary Least Square | 20 positive PnL with very low cumulative return | χ |
| Trading Model 3 | High | Kalman Filter with Bollinger Band | 19 positive PnL with high cumulative return | ✓ |
| Trading Model 4 | High | Support Vector Machine with Markov Regime Switching | 11 positive PnL with high cumulative return | ✓ |

# Conclusion

- Pairs Selection
  - Does it outperform the rational/fundamental selection process?
  - ANS: **Yes**

- Trading Model
  - Does the more complex algorithms produce better results?
  - ANS: **Yes**

# FAQ