Data Preparation

Select 20 stocks in the food industry and use pandas_datareader package in python to download historical stock data from April to December 2019 from Yahoo Finance. The selected stock list is as follows.

Table 1. Stock List

Stock Symbol	**1579.HK**	**0168.HK**	**0291.HK**	**1112.HK**	**2319.HK**
Stock Name	頤海國際	青島啤酒股份	華潤啤酒	H&H國際控股	蒙牛乳業
Stock Symbol	**1717.HK**	**0220.HK**	**1117.HK**	**0151.HK**	**3799.HK**
Stock Name	澳優	統一企業中國	現代牧業	中國旺旺	達利食品
Stock Symbol	**0345.HK**	**1886.HK**	**1583.HK**	**1068.HK**	**0322.HK**
Stock Name	維他奶國際	匯源果汁	親親食品	雨潤食品	康師傅控股
Stock Symbol	**1533.HK**	**6868.HK**	**2218.HK**	**1458.HK**	**0374.HK**
Stock Name	莊園牧場	天福	安德利果汁	周黑鴨	四洲集團

After downloading the data, delete the stock data with missing values. Next, split the training set(2017.04~2017.12) and test set (2018.01~2019.12).Then use three methods for stock matching.

The Distance Method

- **Pair Selection**

First, standardize the stock price. Let $X=\{x_1,x_2,...,x_N\}$be the N samples of a price series. Define

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and

$$\hat{X} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2}$$

To standardize the comparison a normalized series$X' = \{x_1', x_2' \dots x_N'\}$ is formed where

$$X' = \frac{x_i - \bar{X}}{\hat{X}}$$

Second, calculate the distance between two price series X and Y can be defined as

$$d = \sum_{i=1}^{N}(x_i' - y_i')^2$$

Third, each series in the basket is compared with the remaining 19 series and pairs are formed according to the smallest distance criterion.

Table 2. List of Stocks with The Smallest Distance

Stock1	Stock2	Distance
1112.HK	0322.HK	3.5361
1112.HK	3799.HK	3.6603
2319.HK	3799.HK	4.1157
0322.HK	3799.HK	4.2702
0345.HK	6868.HK	4.6834
1112.HK	0345.HK	4.8800
0345.HK	3799.HK	5.0172
1579.HK	3799.HK	5.0256
1579.HK	2319.HK	5.2177
3799.HK	6868.HK	5.3090

- **Setup Rule**

First, calculate the distance between Stock X and Stock Y, the distance is defined

$$distance = x_i' - y_i'$$

Second, construct three characteristic variables **60-day moving average of distance, 5-day moving average of distance, 60-day standard deviation** to standardize the distance.

$$distance'(Zscore) = \frac{5\ day\ MV - 60\ day\ MV}{60 day\ STD}$$

Third, a standard normal distribution has a mean of 0 and a standard deviation 1. If the time series moves k standard deviation beyond the mean, it tends to revert back towards the mean. Using these models, we can create the following trading signals:

1. Whenever the z-score is below -k, we expect the distance to increase. Therefore, we buy 1 share stock B and sell 1 share stock A.
2. Whenever the z-score is above k, we expect the distance to decrease. Therefore, we buy 1 share stock A and sell 1 share stock B.
3. Whenever the z-score is 0, we close the position.

The parameter value k is determined by the model performance in the training set. The transaction fee for each transaction is 0.03% of the transaction amount.

- **Model Test**

Select the three stock pairs with the smallest p-value in the cointegration test, 1112.HK&0322.HK, 1112.HK&3799.HK and 2319.HK&3799.HK.

1. Stock A: 1112.HK, Stock B: 0322.HK

Assuming that the initial amount is 0, set the k value to 1, 1.5 and 2, and use the training set to test the model respectively. When k is **1**, the profit is the highest and the amount reaches **57.92**.

From Figure 1, the two stocks are closer between January and May 2019.

Figure 1. The Standardize Stock Price of 1112.HK and 0322.HK

From Figure 2 that most of the standardized distance are distributed within 1 standard deviations of the mean, and there was a significant deviation in July 2017, December 2017, July 2018 and August 2019.

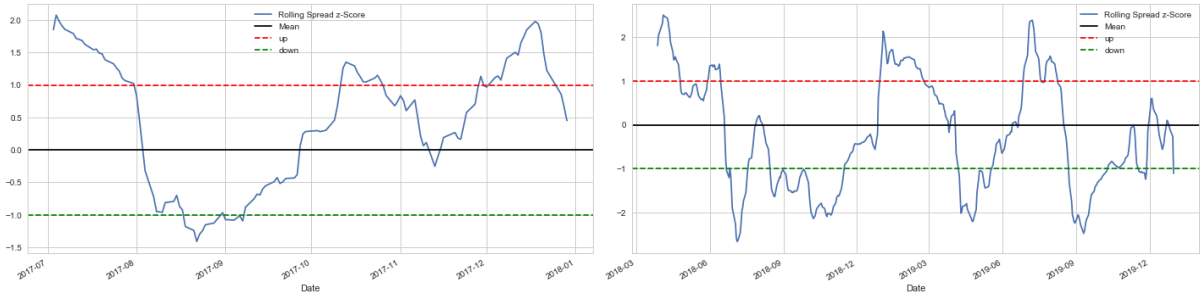

Figure 2. Training Set and Test Set Zscore Chart (1112.HK&0322.HK)

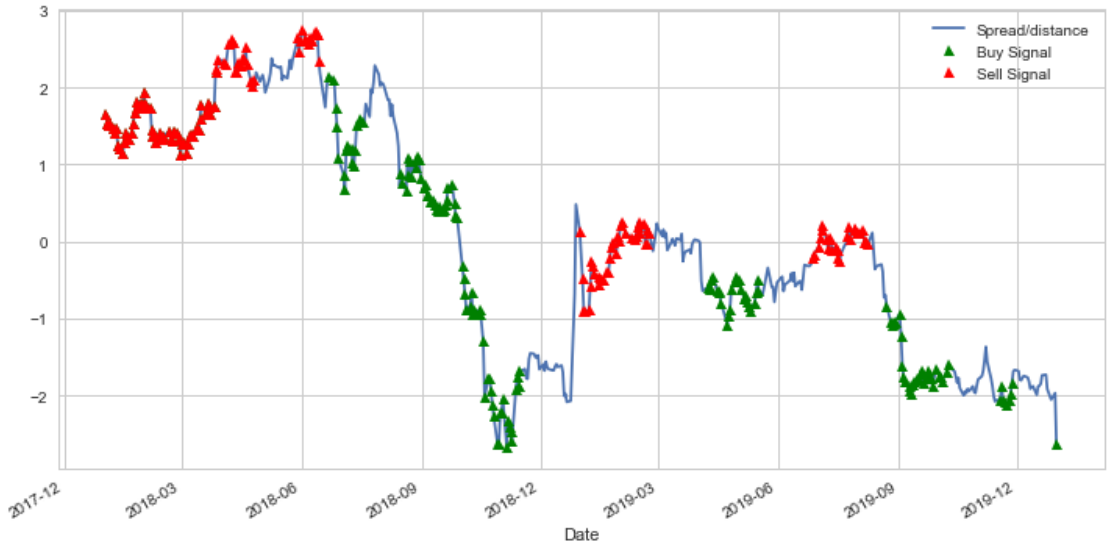

Figure 3. Stock Trading Time Point Distribution (1112.HK&0322.HK)

After calculation, the profit of using the distance strategy in the test set is **164.30**.

2. Stock A: 1112.HK, Stock B: 3799.HK

Assuming that the initial amount is 0, set the k value to 1, 1.5 and 2, and use the training set to test the model respectively. When k is **1**, the profit is the highest and the amount reaches **11.30**. From Figure 4, the price changes of the two stocks are very similar.

Figure 4. The Standardize Stock Price of 1112.HK and 3799.HK

From Figure 5 , there was a significant deviation in September 2017 and December 2017, March 2018, August 2018, and September 2019.

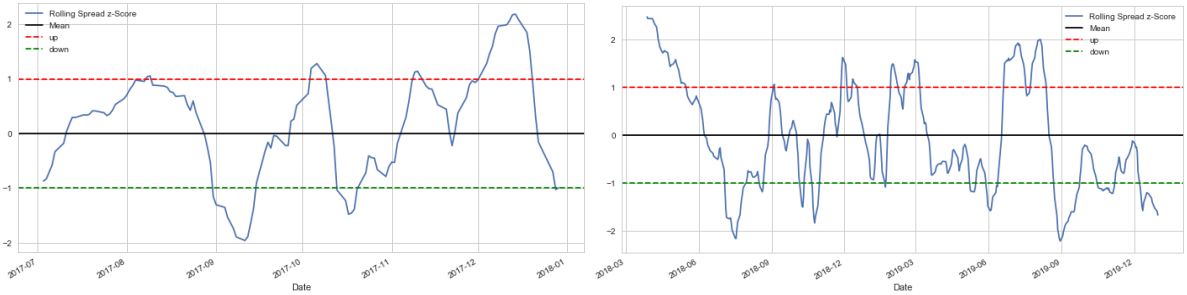

Figure 5. Training Set and Test Set Zscore Chart (1112.HK&3799.HK)

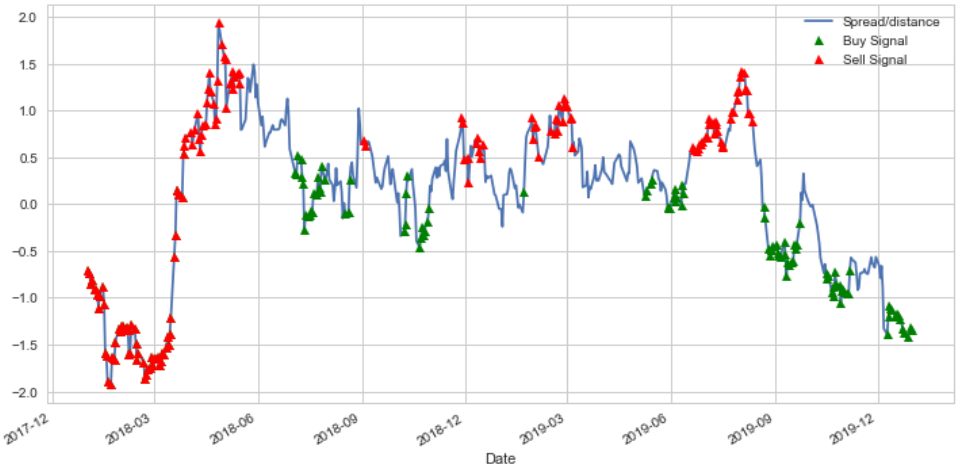

Figure6. Stock Trading Time Point Distribution (1112.HK&0312.HK)

After calculation, the profit of using the distance strategy in the test set is **130.06**.

3. Stock A: 2319.HK, Stock B: 3799.HK

Assuming that the initial amount is 0, set the k value to 1, 1.5 and 2, and use the training set to test the model respectively. When k is **1**, the profit is the highest and the amount reaches **26.39**. From Figure 7, from June 2017 to October 2017, the movement trends of the two stocks were similar. Between November 2017 and May 2018, the price changes of the two stocks deviated and then returned to normal in May 2018. In June 2019, large deviations occurred again.

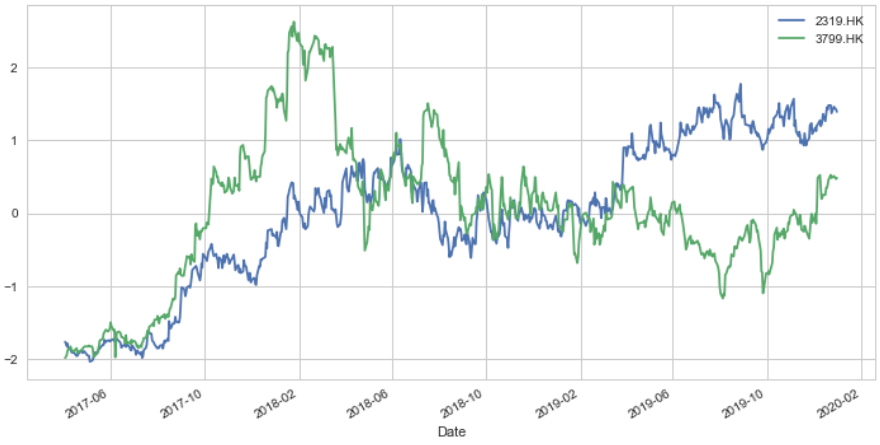

Figure 7. The Standardize Stock Price of 2319.HK and 3799.HK

From Figure 8, there was a significant deviation in September 2017, December 2017, August 2018 and February 2019.

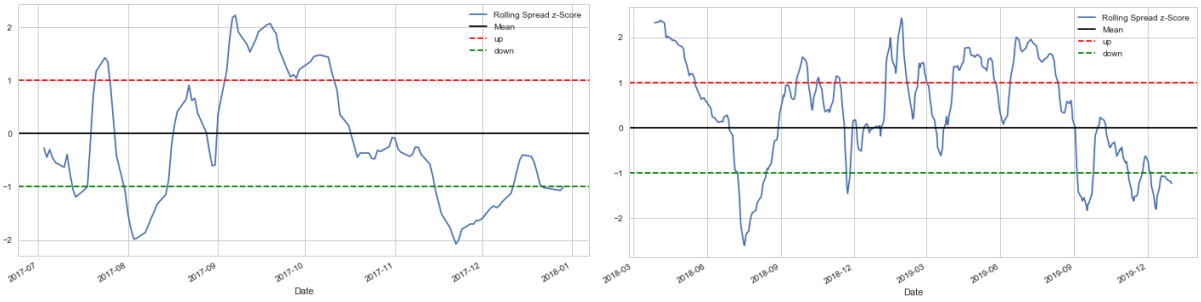

Figure 8. Training Set and Test Set Zscore Chart (2319.HK&3799.HK)

Figure 9. Stock Trading Time Point Distribution (2319.HK&3799.HK)

After calculation, the profit of using the distance strategy in the test set is **96.91**.

In summary, these three stock pairs performed very well under the distance strategy, and performed better than the training set in the test set. The reason is that besides the better distance strategy, the distance trend of the stock pairs in the test set is similar to the training set.

The Cointegration Method

- **Pair Selection**

Use the coint function in the statsmodels package to check the cointegration between stock prices. The select criterion of coint function is based on the degree of cointegration which measured by **Engle-Granger test**. The function returns the t statistic of the residual unit root test and the approximate asymptotic p value based on MacKinnon.

Screen out the stock pairs with p value less than 0.05, the result is as follows.

Table 3. List of Stocks Satisfying the Test of Cointegration

Stock1	Stock2	score	p-value
1717.HK	1533.HK	-4.0888	0.0053
0291.HK	0345.HK	-3.9730	0.0078
0345.HK	6868.HK	-3.7837	0.0143
1112.HK	0322.HK	-3.7662	0.0150
0345.HK	0322.HK	-3.6616	0.0205
2319.HK	1533.HK	-3.5200	0.0307
1886.HK	1068.HK	-3.5148	0.0311
0291.HK	1068.HK	-3.4866	0.0336
0322.HK	3799.HK	-3.3579	0.0472
1112.HK	3799.HK	-3.3443	0.0489

- **Setup Rule**

First, fit the model using the training set data to get the beta value. Then use the following formula to construct a spread sequence in the test set.

$$Spread = Stock\ Price_B - beta * Stock\ Price\_A$$

Second, construct three characteristic variables **60-day moving average of spread, 5-day moving average of spread, 60-day standard deviation** to standardize the spread.

$$Spread'(Zscore) = \frac{5\ day\ MV - 60\ day\ MV}{60 day\ STD}$$

Third, a standard normal distribution has a mean of 0 and a standard deviation 1. If the time series moves k standard deviation beyond the mean, it tends to revert back towards the mean. Using these models, we can create the following trading signals:

1. Whenever the z-score is below -k, we expect the spread to increase. Therefore, we buy

1share stock B and sell **beta** share stock A.

2. Whenever the z-score is above k, we expect the spread to decrease. Therefore, we buy **beta** share stock A and sell 1 share stock B.

3. Whenever the z-score is 0, we close the position.

The parameter value k is determined by the model performance in the training set. The transaction fee for each transaction is 0.03% of the transaction amount.

- **Model Test**

Select the three stock pairs with the smallest p-value in the cointegration test, 1717.HK&1533.HK, 0291.HK&0345.HK and 0345.HK&6868.HK.

1. Stock A: 1533.HK, Stock B: 1717.HK

Assuming that the initial amount is 0, set the k value to 1, 1.5 and 2, and use the training set to test the model respectively. When k is **2**, the profit is the highest and the amount reaches **34.91**. From Figure10, from March to October 2017, the stock prices of the two stocks changed similarly. 1533.HK had a larger increase at the end of 2017, and 1717.HK ushered in a larger increase in July 2019.

Figure 10. The Stock Price of 1717.HK and 1533.HK

From Figure 11 ,there was a significant deviation in August 2017 and June 2019. Therefore, the operation is mainly concentrated in this time period.

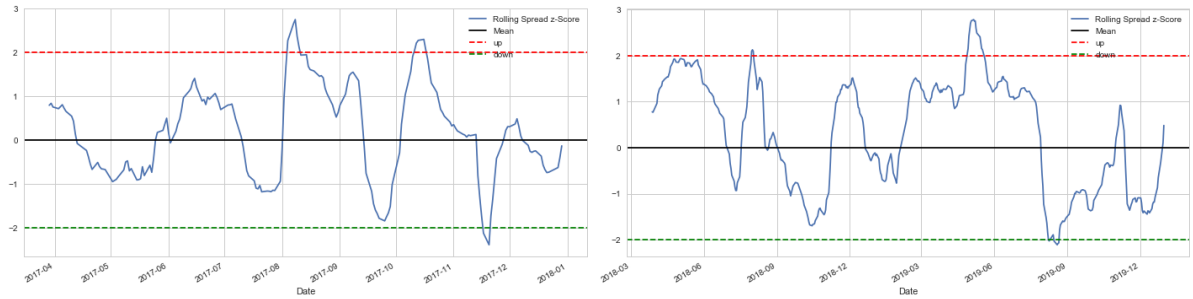

Figure 11. Training Set and Test Set Zscore Chart (1717.HK&1533.HK)

Figure 12. Stock Trading Time Point Distribution (1717.HK&1533.HK)

After calculation, the profit of using the cointegration strategy in the test set is **72.63**.

2. Stock A: 0345.HK, Stock B: 0291.HK

Assuming that the initial amount is 0, set the k value to 1, 1.5 and 2, and use the training set to test the model respectively. When k is **2**, the profit is the highest and the amount reaches **32.63**. From Figure 13, from March 2017 to January 2018, the stock prices of the two stocks changed similarly. Between February and June 2018, the volatility of the two stocks diverged, but returned to normal in July.

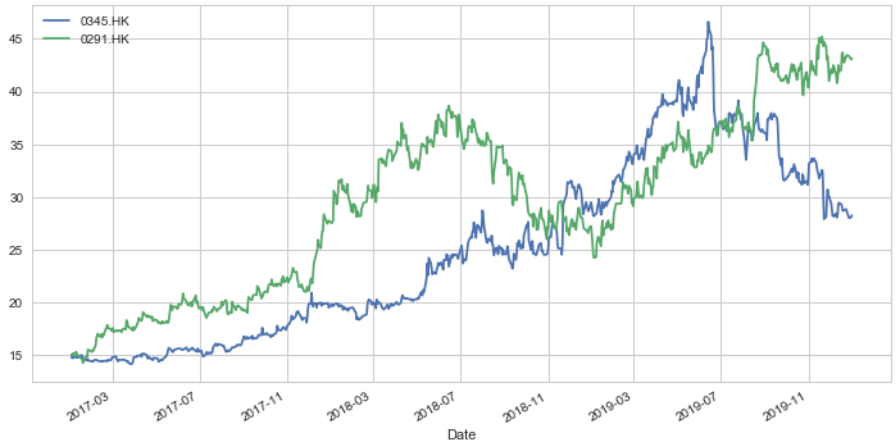

Figure 13. The Stock Price of 0345.HK and 0291.HK

From Figure 14, there was a significant deviation in December 2017, October 2018 and June 2019.

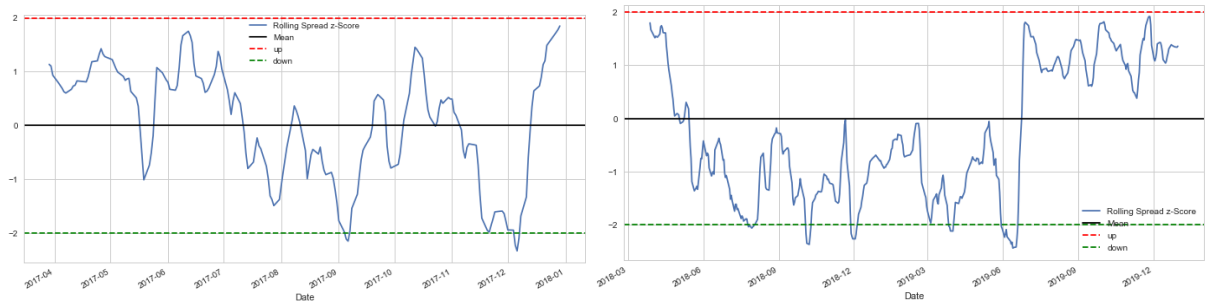

Figure 14. Training Set and Test Set Zscore Chart (0345.HK & 0291.HK)

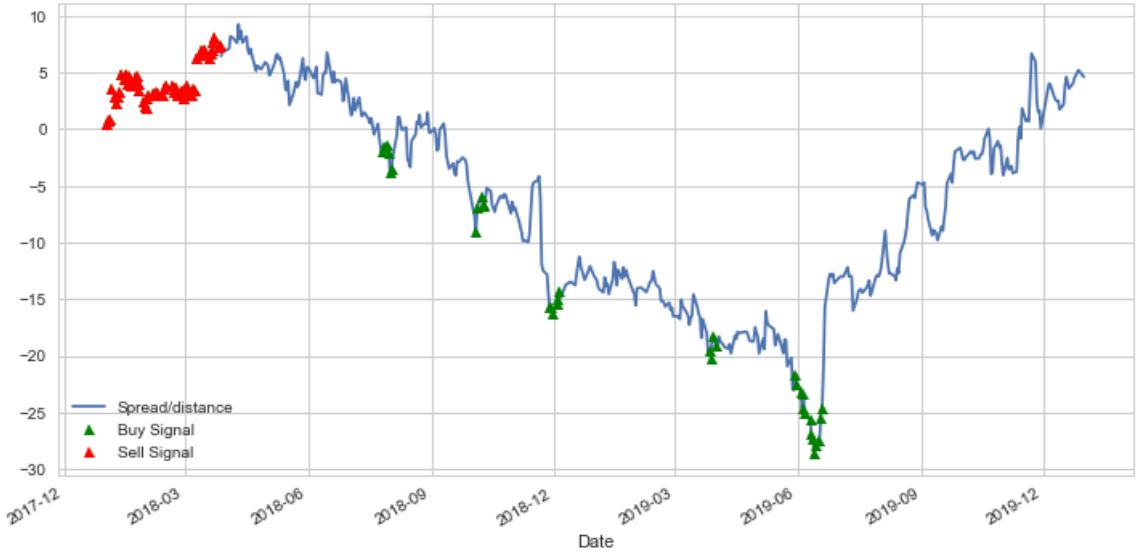

Figure 15. Stock Trading Time Point Distribution (0345.HK & 0291.HK)

After calculation, the profit of using the cointegration strategy in the test set is **107.84.**

3. Stock A: 0345.HK, Stock B: 6868.HK

Assuming that the initial amount is 0, set the k value to 1, 1.5 and 2, and use the training set to test the model respectively. When k is **2**, the profit is the highest and the amount reaches **46.02**. From Figure 16, before June 2018, the price movements of the two stocks were very similar, but the price of 0345.HK continued to rise for a period of time, while the price of 6868.HK did not change significantly.

Figure 16. The Stock Price of 0345.HK and 6868.HK

From Figure 10,there was a significant deviation in July 2017, December 2018, July 2019 and October 2019.

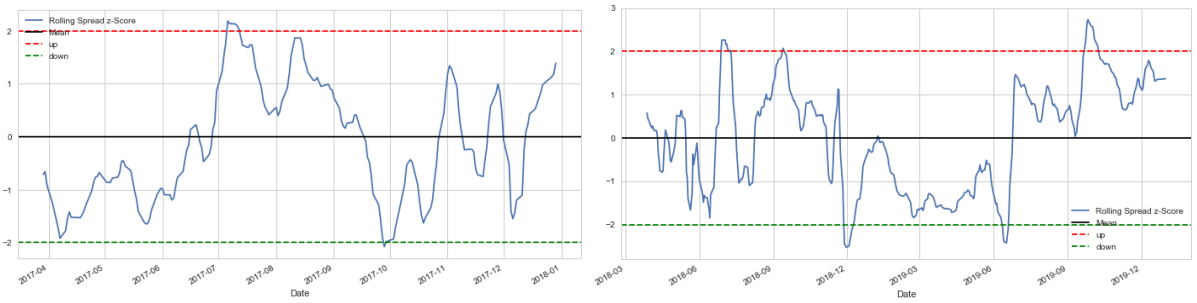

Figure 17. Training Set and Test Set Zscore Chart (0345.HK & 6868.HK)

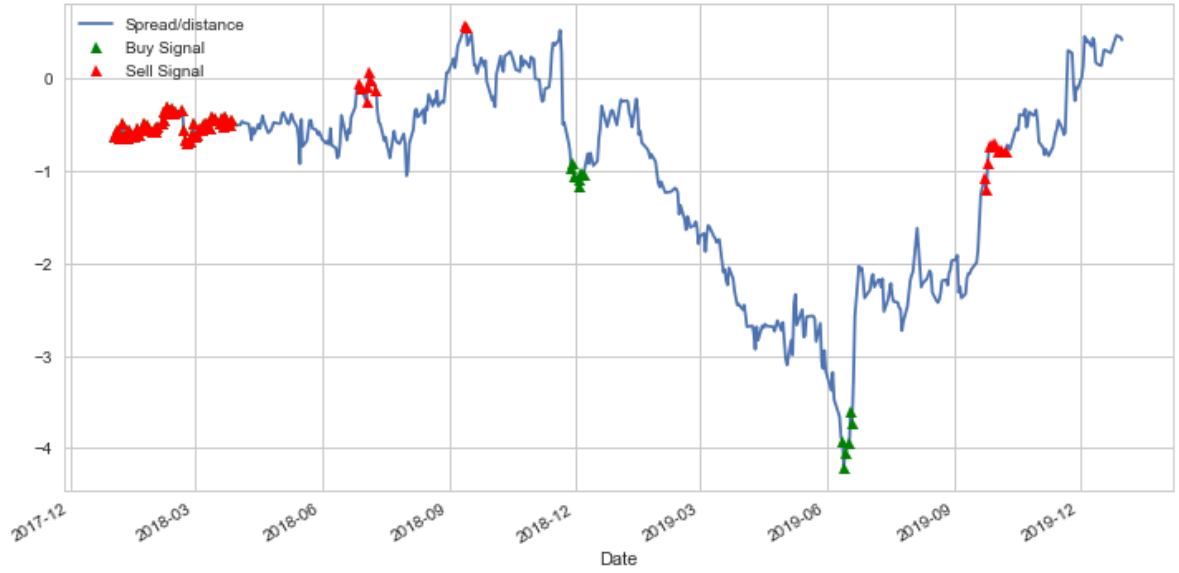

Figure 18. Stock Trading Time Point Distribution (0345.HK & 6868.HK)

After calculation, the profit of using the cointegration strategy in the test set is **0.66**.

In summary, from the above figure that 1533.HK & 1717.HK and 0345.HK & 0291.HK perform well in the cointegration strategy, and the performance of the test set is even better than the training set. But 0345.HK & 6868.HK 's performance has declined because the share price of 0345.HK rose significantly from 2018 to 2019, and there is a weak or even no cointegration relationship with the stock price of 6868.HK.

Hurst Exponent

- **Pair Selection**

Let X_t and Y_t be two price series and $Z_t = X_t - Y_t$ be the difference between the two prices called "spread". Then use hurst function to computes hurst exponent.

As we know, when hurst exponent equals 0.5, the process is unpredictable. When hurst exponent is greater than 0.5, the process is trending. When hurst exponent is less than 0.5, the process is mean reversion.

Therefore, we pick out stock pairs with hurst exponent less than 0.5. The top 10 result is as follows.

Table 4. List of Stocks (Hurst Exponent)

Stock1	Stock2	Hurst
2319.HK	1533.HK	0.0182
1068.HK	0374.HK	0.0330
1886.HK	0374.HK	0.0444
2218.HK	1458.HK	0.0514
1068.HK	1458.HK	0.0586
1458.HK	0374.HK	0.0591
6868.HK	1458.HK	0.0602
1117.HK	0374.HK	0.0651
1886.HK	1458.HK	0.0700
0151.HK	3799.HK	0.0750

- **Setup Rule** (Same as cointegration strategy)

First, fit the model using the training set data to get the beta value. Then use the following formula to construct a spread sequence in the test set.

$$Spread = Stock\ Price_B - beta * Stock\ Price\_A$$

Second, construct three characteristic variables **60-day moving average of spread, 5-day moving average of spread, 60-day standard deviation** to standardize the spread.

$$Spread'(Zscore) = \frac{5\ day\ MV - 60\ day\ MV}{60 day\ STD}$$

Third, a standard normal distribution has a mean of 0 and a standard deviation 1. If the time series moves k standard deviation beyond the mean, it tends to revert back towards the mean. Using these models, we can create the following trading signals:

1. Whenever the z-score is below -k, we expect the spread to increase. Therefore, we buy 1share stock B and sell **beta** share stock A.
2. Whenever the z-score is above k, we expect the spread to decrease. Therefore, we buy **beta** share stock A and sell 1 share stock B.
3. Whenever the z-score is 0, we close the position.

The parameter value k is determined by the model performance in the training set. The transaction fee for each transaction is 0.03% of the transaction amount

- **Model Test**

Select the three stock pairs with the smallest hurst exponent in the cointegration test, 2319.HK&1533.HK, 1068.HK&0374.HK and 1886.HK&0374.HK.

1. Stock A: 2319.HK, Stock B: 1533.HK

Assuming that the initial amount is 0, set the k value to 1, 1.5 and 2, and use the training set to test the model respectively. **No matter what k is, profit is negative.**

From Figure19, 1533.HK's price soared in December 2017, then returned to normal and continued to decline.2319.HK price maintains upward trend.

Figure 19. The Stock Price of 1533.HK and 2319.HK

2. Stock A: 1068.HK, Stock B: 0374.HK

Assuming that the initial amount is 0, set the k value to 1, 1.5 and 2, and use the training set to test the model respectively. When k is **1.5**, the profit is the highest and the amount reaches **1.02**.

From Figure 20, the overall price trends of the two stocks are similar.

Figure 20. The Stock Price of 1068.HK and 0374.HK

From Figure 21, there was a significant deviation in July 2017 and November 2017. Between

2018 and 2019, zscore changed drastically, so trading operations were intensive.

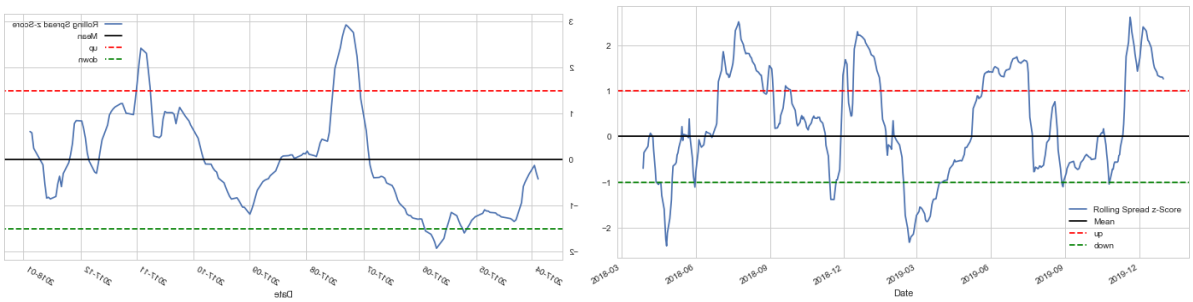

Figure 21. Training Set and Test Set Zscore Chart (1068.HK &0374.HK)

Figure 22. Stock Trading Time Point Distribution (1068.HK &0374.HK)

After calculation, the profit of using the hurst exponent in the test set is **254.14**.

3. Stock A: 1886.HK, Stock B: 0374.HK

Since 1886.HK has been suspended since April 2018, no simulation is conducted.

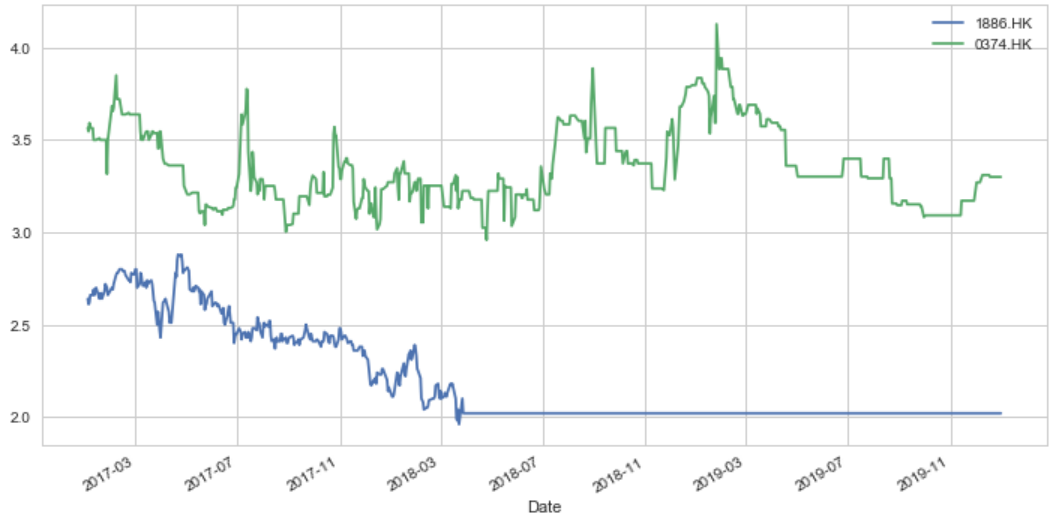

Figure 23. The Stock Price of 1886.HK and 0374.HK

In summary, judging from the performance of the above three pairs of stocks, hurst strategy performed poorly.

In the first stock pair, the returns are even negative. The reasons may be the following. From 2017 to 2018, the trends of the two stocks were similar, but between 2018 and 2019, the trends of 1533.HK and 2319.HK showed a large deviation, which was inconsistent with the previous ones.

Conclusion

Compared with the three strategies, the distance strategy performs best. The parameter k that perform best in the test set also perform best in the training set, and all have positive profits.

The cointegration strategy is next, the parameter k that performs best in the training set is not optimal in the test set. A pair of stocks performed poorly in the training set, with a profit of almost 0.

The Hurst strategy performed the worst, with even negative profits in the test set.

Table 5. Profit Comparison of Different Strategies

Distance Strategy			
Stock	k	Training Set Profit	Test Set Profit
1112.HK&0322.HK	1	57.92	164.3
1112.HK&3799.HK	1	11.303	130.06
2319.HK&3799.HK	1	26.39	96.91
Cointegration Strategy			
1717.HK&1533.HK	2	34.91	72.63
0291.HK&0345.HK	2	32.63	107.84
0345.HK&6868.HK	1	46.02	0.66
Hurst Strategy			
2319.HK&1533.HK	-	-	-
1068.HK&0374.HK	1.5	1.02	254.14
1886.HK&0374.HK	-	-	-