# Dispersion tracking and timing

# Cointegration bets

**Rahul Budhadev, Kunal Chakraborty**

10/08/2021

## 1 Executive Summary

Pairs trading is a well-known investment strategy pioneered by Nunzio Tartaglia and team during his time at Morgan Stanley in 1980s. Since then it has been employed as an important long short equity investment tool by hedge funds and institutional investors. At its core, the strategy comprises of two steps. First, it requires the identification of two stocks, for which the price behavior is "similar" or are linked to each other. This indicates that both securities are exposed to similar underlying factors and this information is purely discovered through statistical/ machine learning methods without any serious fundamental analysis of the common factors. This behaviour essentially signifies an equilibrium relationship between the prices of the two securities and is used as an indicator for identifying pairs. Once a pair is identified, the second step of pairs trading strategy kicks in. The underlying premise is that if two securities prices have moved closely in the past then this should persist in the future. Any shift in this equilibrium is short term and should ideally invert in the long run. This presents an interesting opportunity to trade, since upon the identification of any statistical anomaly in the

spread of the securities prices an investor can setup a trade such that she shorts the over performing stock and long the under performing one. The position is exited upon eventual reversal/normalization of the spread.

In our project, we aim to incorporate several innovations in pairs trading strategy. We review the latest trends and novel ideas in this domain and try out some ideas on the existing universe of stocks in the portfolio of Appian Way Asset management. We try our ideas in three different directions, namely pairs identification, entry point of a trade and exit point of a trade. Our emphasis is mostly on application of machine learning in systematically making the above three critical decisions. However, we guide our findings and code output with discretionary intuition and feedback from portfolio managers from Appian Way asset management. As a part of this project, we also created a well documented and well tested code repository with multiple modules. The repository contains a host of econometric tools and functions that can be re-used for a lot of analysis besides the domain of pairs trading. Although in our experiments we were not successful in coming up with a profitable portfolio during the selected time period, however a lot of insights were derived from this exercise which proved very helpful to the fund. Even if the generated returns are currently negative, the results are very promising and with further modifications a profitable portfolio could be certainly created.

# 2 Data collection and initial pre-processing

The following subsection provides a detailed overview of the different data sources used for the project. We also had to carefully pre-process our data so that the results are correctly and robustly estimated.

## 2.a Variables and data sources

We primarily used alpha-vantage Api for fetching our time series data pertaining to stocks. It's a python package that provides access to vast resource of data pertaining to financial time series. Other time series like risk free rate, SP 500 and other macro-economic variables has been obtained from FRED database. We also use some data from Kenneth-French's website related to factors. Although in our final experiments, they don't turn out to be very useful. Rest, our universe of stock tickers is obtained from Appianway's currently managed portfolio/watch list of stocks and commodities. We were provided with a data set which was enriched beforehand by clustering information. More information on the clustering methodology will be provided in the subsequent sections. Appianway relies on a third party organization's (Equity Data Science) platform for their data and reporting needs. The dataset provided to us was enriched with information like **industry, sub-industry, exemplar (cluster identifier), similarity threshold etc**. These fields are very useful in post factor analysis of identified clusters since stocks whose prices move together usually belong in the same sector.

Once we obtain the ticker information from Appianway's proprietary data, we enrich it with daily OHLC prices using alpha vantage api. We also get information about dividend amount and stock split from the api. Out of 836 tickers in the universe, we source data for 819 tickers and then store them in our system. Although our final dataset comprises of more than 20 years of daily prices for stocks, we run most of our experiments post 2010 data.

## 2.b Data pre-processing

Apart from the basic processing of merging stock price information with their static attributes like sector and cluster label, we compute the returns for every ticker for every day. The formula for computing returns:

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

While computation of returns it is assumed that during days where their is a stock split the returns are zero. This is a very simplistic assumption but such incidences are rare in the data and it is safe to make this change. We also don't encounter null values or missing values in our dataset.

### 2.b.1 Rolling beta estimation

Another core aspect of pairs trading strategy is the fact that it is a market neutral strategy, meaning that a particular pair is unaffected by movement of the broader market. In order to setup such a trade, the estimation of market betas is required. Although the beta information is directly obtainable from Bloomberg, we decided to compute the betas on our own. In the classic CAPM world, market beta is defined as:

$$R_i = R_f + \beta_i * (R_m - R_f)$$

Market betas are not static and change with time. Hence we use a 60 day window and compute rolling betas. Using tools from linear regression, the beta parameter could be estimated by the following equation:

$$\hat{\beta}_i = (\tilde{R_m}^T \tilde{R_m})^{-1} \tilde{R_m}^T \tilde{R_i}$$

$$where \tilde{R_m} = market\ excess\ return$$

$$and \tilde{R_i} = excess\ return$$

The risk free rate that we use is the 3 month libor rate. These rates are converted to daily rates using the 30/360 convention.

$$R_{fdaily} = ((1 + R_{fmonthly})^{\frac{365}{90}})^{\frac{1}{365}}$$

.

We store all of this information in one dataframe, which eventually serves as our main file. A snapshot of the data table is provided below. All the columns in this table are the ones that are eventually used to setup the pairs trading portfolio.

| date | 5. adjusted close | 6. volume | 9. Ticker Sym | returns | betas | Sector | Industry Group | Industry | Sub-Industry |
|---|---|---|---|---|---|---|---|---|---|
| 1999-11-01 | 6.7058010118 | 159200.0 | AEM | | | Materials | Materials | Metals & Mining | Gold |
| 1999-11-02 | 6.50759999667 | 126900.0 | AEM | -0.0295566502467389 | | Materials | Materials | Metals & Mining | Gold |
| 1999-11-03 | 6.29288223028 | 87600.0 | AEM | -0.0329949238582385 | | Materials | Materials | Metals & Mining | Gold |
| 1999-11-04 | 6.2433319765 | 129400.0 | AEM | -0.0078740157477561 | | Materials | Materials | Metals & Mining | Gold |

Figure 1: Snapshot of the final table used for analysis after full pre-processing

# 3 Pairs identification methodology

## 3.a Overview

We identify and introduce a suite of innovations, mostly around the pairs selection process. The pairs identification is not solely based on a particular systematic strategy. Instead we have adopted a heuristic approach by constructing a set of systematic and discretionary checks. The first suite of innovations is about creating a robust statistical framework – using rolling correlations and TIC (Theory-Implied Correlation) powered by machine learning algorithms instead of conventional sample correlation. Additionally, we are complementing stationary tests with cointegration. The second category of enhancements is to supplement a purely statistical strategy with fundamental intuitions. In particular, we confine pairs to be those stocks with similar fundamental characteristics, sector classification, and industry exposures – all derived from robust clustering frameworks like hierarchical clustering and affinity propagation. Additionally, empirical checks have been conducted to identify if the selected pairs have their businesses closely aligned with each other, since we expect a stronger co-movement between such pairs.

## 3.b Clustering using affinity propagation

Affinity Propagation creates clusters by sending messages between data points until convergence. Unlike clustering algorithms such as k-means or k-medoids, affinity propagation does not require the number of clusters to be determined or estimated before running the algorithm, for this purpose the two important parameters are the preference, which controls how many exemplars (or prototypes) are used, and the damping factor which damps the responsibility and availability of messages to avoid numerical oscillations when updating these messages.

A dataset is described using a small number of exemplars. 'Exemplars' are members of the

input set that are representative of clusters. The messages sent between pairs represent the suitability for one sample to be the exemplar of the other, which is updated in response to the values from other pairs. This updating happens iteratively until convergence. At that point the final exemplars are chosen, and hence we obtain the final clustering.

The Exemplars based on Affinity propagation were provided to us by Appian Way. This has been utilized in combination with the statistical and qualitative assessments described in subsequent sections to identify robust pairs.

## 3.c Hierarchical clustering

Here, we look for clusters of correlations using the agglomerate hierarchical clustering technique. Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. Its primary advantage over other clustering methods is that you don't need to guess in advance how many clusters there might be. Agglomerate Clustering first assigns each data point into its own cluster, and gradually merges clusters until only one remains. It's then up to the user to choose a cutoff threshold and decide how many clusters are present.

The Linkage function in the scipy.cluster.hierarchy library does the actual clustering in one line of code, and returns a list of the clusters joined in the format: Z = [stock1, stock2, distance, sample]

There are also different options for the measurement of the distance. The option we will choose is the ward distance measurement, but others are also possible (average, single, centroid, etc.).

It's important to get a sense of how well the clustering performs. One measure is the Cophenetic Correlation Coefficient. This compares (correlates) the actual pairwise distances of all your samples to those implied by the hierarchical clustering. The closer c is to 1, the better the clustering preserves the original distances. Generally c $>$ 0.7 is

considered a good cluster fit. Of course, other accuracy checks are possible.

We used the GICS Sector classification to divide the 819 tickers into the following sectors:

1. Communication Services
2. Consumer Discretionary
3. Consumer Staples
4. Energy
5. Financials
6. Health Care
7. Industrials
8. Information Technology
9. Materials
10. Other
11. Real Estate
12. Utilities

After splitting the dataset based on the above sector classification, a correlation matrix was constructed based on returns of the relevant tickers. Hierarchical clustering was then used on the correlation matrix to create a dendogram. This dendogram identified pairs having minimum distances and showing strong linkages. The generated dendogram for the "Materials" sector is given below.

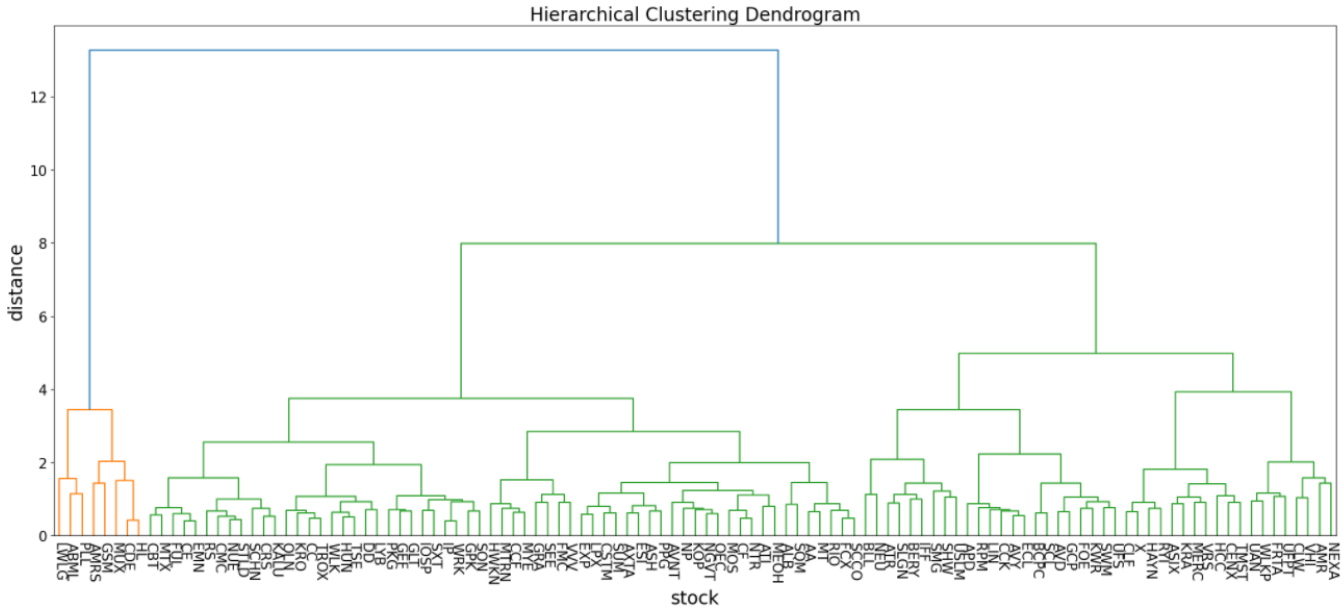


Figure 2: Hierarchical clustering dendogram - Sector: Materials

A cophentic correlation coefficient of 0.7 was set as the threshold for pairs selection. The top 5-10 stock pairs based on minimum dendogram distance were selected from each Sector dendogram tree. Roughly 70-100 selected pairs were passed on for subsequent analysis using rolling correlations, cointegration and qualitative assessments.

## 3.d Rolling correlations

Rolling correlations are correlations between two time series on a rolling window. One benefit of this type of correlation is that you can visualize the correlation between two time series over time.

We computed a 3-month rolling correlation between all stocks along with its mean and standard deviation, to get a historical time-series representation of the co-movement between the two stocks. A threshold of 0.75 for mean and 0.10 for standard deviation was selected to identify pairs that underwent further pairs selection tests.

The stock pairs selected in the hierarchical clustering step were subjected to this assessment. Some examples of high rolling correlated pairs with low standard deviation were

IVV-SPY (Mean: 0.995 and Std Dev: 0.005) and DHI-LEN (Mean: 0.85 and Std Dev: 0.08).

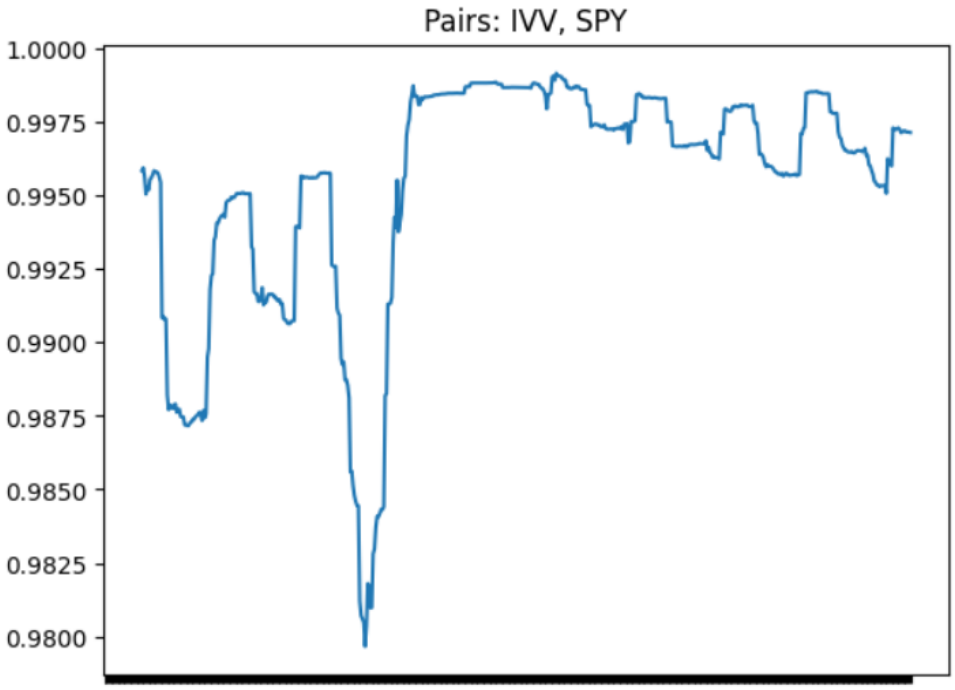


Figure 3: 3-month Rolling correlation between IVV and SPY

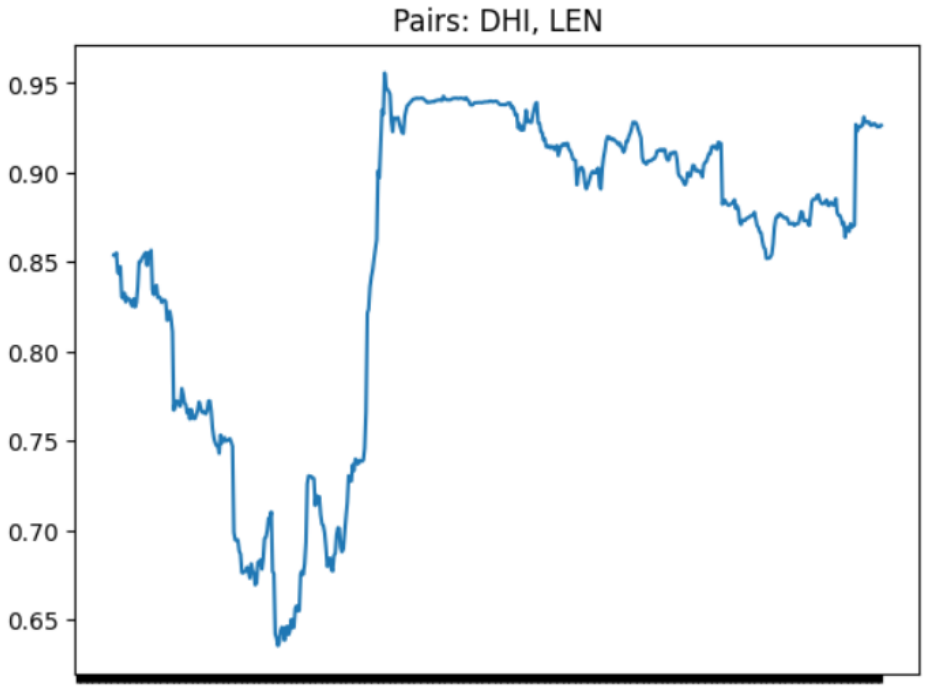


Figure 4: 3-month Rolling correlation between DHI and LEN

DR Horton Inc. (DHI) and Lennar Construction (LEN) are both home construction companies. The iShares Core SP 500 ETF (IVV) seeks to track the investment results of an index composed of large-capitalization U.S. equities. The SPDR SP 500 (SPY) trust is an exchange-traded fund which tracks the SP 500 Index. About 30-50 pairs which passed the rolling correlation assessment were further assessed in subsequent sections.

## 3.e Cointegration

As mentioned above, the core criteria for two stocks to form a pair is based on the idea of similarity in behaviour. There is some flexibility and scope of innovation when it comes to this definition of "similarity". There are two famous statistical techniques that are generally used to assess the degree of similarity between two pairs. The first one is the actual difference between the normalized prices of two stocks. In this method, first the stock prices are normalized and then the euclidean distance between them is measured at every time point. Usually on the basis of some threshold final pairs are selected. The second method is based on the concept of co-integration. Co-integration is a phenomenon wherein the source of non-stationary between two time series is the same and a linear combination of these two series results in a stationary vector.

Price data usually is non stationary and integrated of order 1 I(1). However, in it's most naive definition, when two co-integrating series are regressed on one another (depending on which series error corrects) the residual of the resulting regression is stationary. This is a phenomenal result (Noble price winning discovery by Granger 2003) and provides an interesting opportunity in the context of setting up a pairs trade. Stationary series are vector invariant, meaning that their statistical properties don't change throughout time. Any shock applied to a stationary series would eventually die out. This property is known as "mean reversion" and is one of the key ingredient in the identification of a suitable pair. Two stocks whose prices are co-integrated would have a stationary spread. Any statistical anomaly in this spread presents an opportunity to enter a profitable position since the shock to the spread is expected to die out and revert to it's mean. We emphasize

on this methodology and concept as well in our experiments with pairs trading.

An example of two stocks whose prices are co-integrated is given below. FLS is the ticker for the company Flowserv Corp. The other ticker is GTES, which is representative of Gates industrial corp. Not surprisingly both these companies are in the business of manufacturing pumps, valves, industrial grade fluid control and transmission devices. The y label of the below plot is the normalized price.

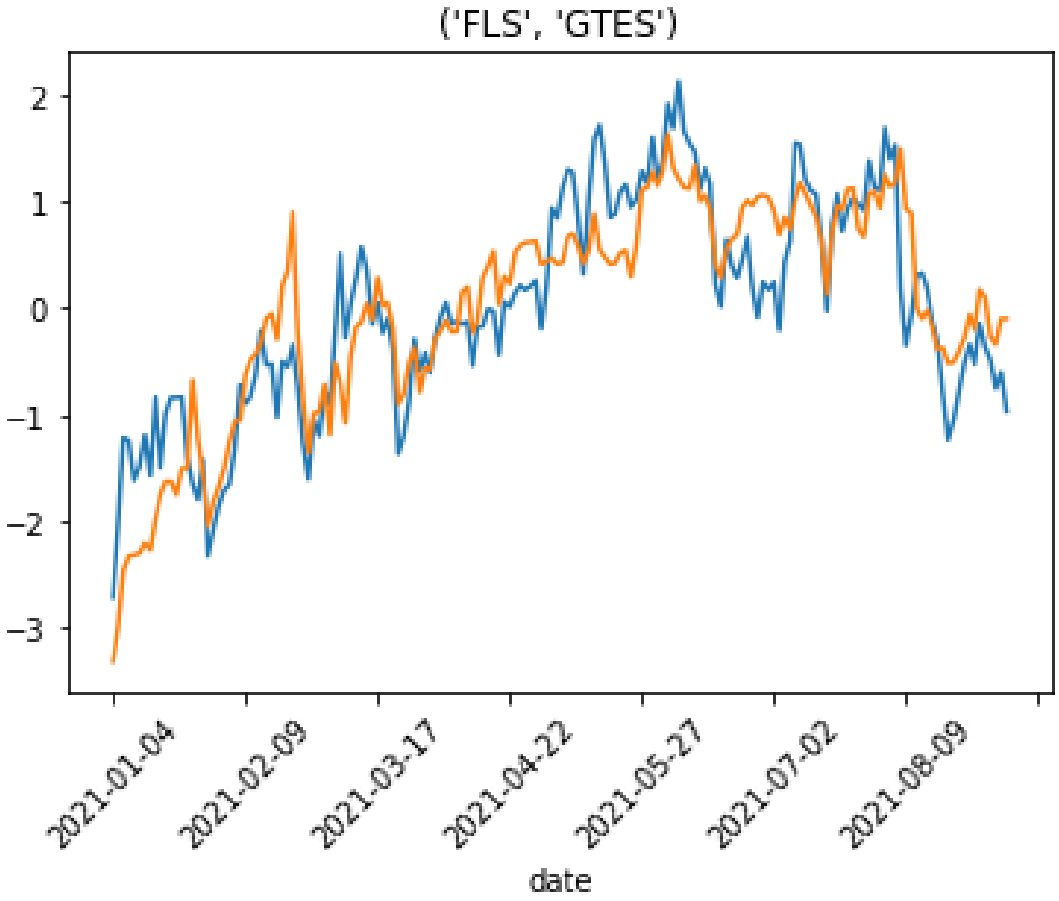


Figure 5: Cointegrated prices of Flowserv Corp and Gates industrial Corp

The stock price of Flowserv Corp error corrects on Gates industrial Corp and hence when we regress FLS price against GTES, the residual series that we obtain is stationary.

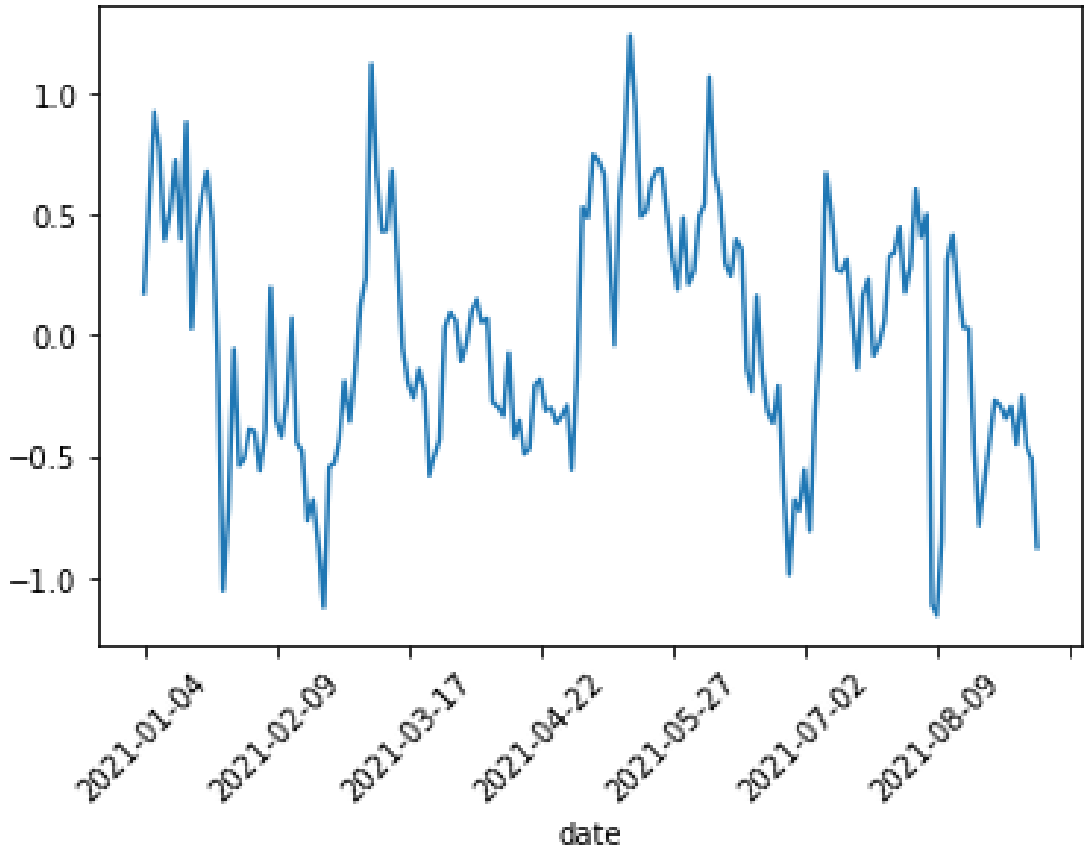


Figure 6: Stationary residual series obtained through a linear combination of FLS and GTES prices

### 3.e.1 Engle-Granger test for cointegration

In order to identify cointegration between two stocks we conduct the two step Engle and Granger test of co-integration. As the name suggests, there are two steps in performing this test. First, one series is regressed against the another. The residual from the regression is stored. In the second step, the residual is tested for stationarity using Augmented dickey fuller test. There are critiques to using this methodology since it is usually not clear which series error corrects on which one. In our approach, we perform both the regression, i.e and then and perform stationary test on both the residuals. The null hypothesis of the ADF test indicates the presence of unit root, so whichever residual series gets strongly rejected for non stationarity becomes the correct choice. Secondly, it

is possible for more than two series to be co-integrated, since the theoretical construct is that existence of a stationary series upon linear combination of multiple integrated I(1) series. For over coming these shortcomings, usually the Johanson's test is preferred. However, we were unable to find any stable implementation of Johanson's test in python and proceeded with the 2 step EG test.

### 3.e.2 Co-integration consistency

One weakness of the co-integration approach is the fact that it is a yes/no phenomenon. Unlike other measures like correlation or euclidean distances, there is no relative "degree" or strength when it comes to determining the extent of cointegration between two pairs. We initially thought of employing different ideas like measuring the speed of convergence through computing impulse response function, or just sorting the pairs using absolute measure of their long run beta. However these approaches are more discretionary in nature and can be used as a side tool to narrow the list of co-integrated pairs. Instead we employ a more robust approach to identify the persistence of co integrating relationship. It is true that trends and patterns change throughout regimes in the financial time series and pairs that are co integrated in all types of business cycles/regimes are ideally good candidates of pairs trading strategy.

To that extent, we estimate a measure known as the "Cointegration consistency" of a pair. In this measure, we move over a rolling window of 'n' days and perform the two step EG test. For every day, we lookback upto the 'n' period and label according to the result of the EG test. Positive results are labeled as 1 whereas negative are labeled as 0. Then taking the ratio of cumulative sum of the consistency measure and dividing it by the length of the available history gives us an idea of how long co-integration has been observed in a particular pair. This information is also very useful because periods where co-integrating relationship do not exist provide an interesting point for analysis. This assessment has been performed for all stock pairs surviving the previous thresholds/assessments.

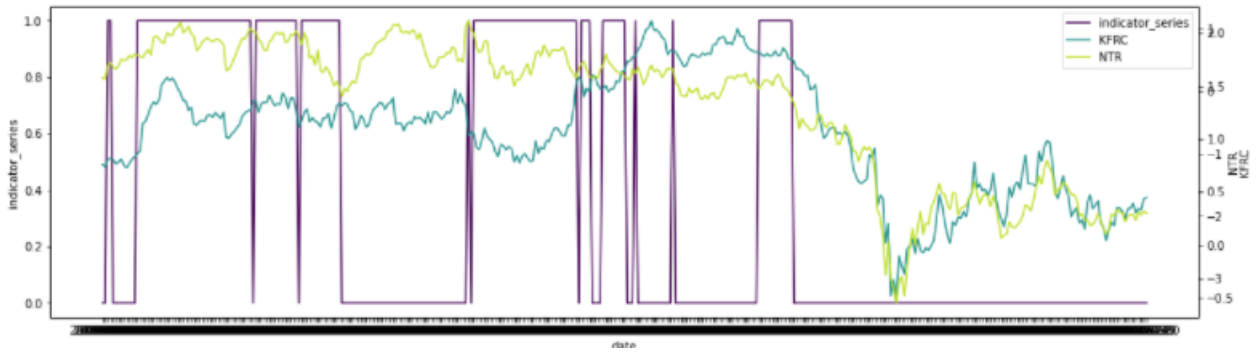


Figure 7: Normalized prices for KFRC-NTR along with indicator series

The above plot shows the entire history of two co-integrated stocks, Kforce Inc (KFRC) and Nurtien limited (NTR). The labels on the y-axis are normalized price whereas the magenta line is the cointegrating indicator. The periods where the indicator line is at the top show strong cointegration behaviour. We utilize such dual axis plots for our investigation of good cointegrated pairs as outlined below.

## 3.f Qualitative assessment

In order to assess the robustness of a pairs trading strategy, in addition to the systematic and statistical analysis done so far, we have embedded discretionary decision making in our pairs selection process. This selection step is based on studying the product/service/business of the concerned stocks and drawing intuitions for long lasting and highly codependent movements. In our example of DHI-LEN in the rolling correlations assessment, we assessed that both are construction companies, thereby passing the statistical as well as qualitative assessments.

Here are some examples of rationale behind the final pairs selection step. Lear Corporation is an American company that manufactures automotive seating and automotive electrical systems whereas Magna International Inc. is a Canadian mobility technology company for automakers (LEA-MGA). Expedia Group, Inc. is an American online travel shopping company for consumer and small business travel whereas Booking Holdings Inc.

is an American travel technology company (EXPE-BKNG). Hormel Foods Corporation manufactures and markets consumer-branded meat and food products whereas Campbell Soup Company, doing business as Campbell's, is an American processed food and snack company (HRL-CPB). Oceaneering International, Inc. is a sub sea engineering and applied technology company based in Houston, Texas, U.S. that provides engineered services and hardware to customers who operate in marine, space, and other environments whereas Cactus Wellhead is the leading manufacturer of pressure control equipment used during drilling, completion and production operations onshore and offshore (OII-WHD). This additional filter ensures that the pairs trading strategy depends on discretionary assessment and is not solely based on technical tests.

## 3.g Final pairs selection

Pairs identification is a key component to create a successful pairs selection strategy. Our multi-faceted approach has helped us develop a methodology that is both systematic and discretionary, and is driven by both statistics and financial intuition. After implementing all the above mentioned filters, we have identified 28 stock pairs which have the potential to create profit through a long-short pairs trading strategy, conditional to identifying optimal entry and exit points.

For example, KBE-KRR successfully cleared all the thresholds for pair selection. The cointegration consistency information of the aforementioned pair from 2019 to 2021 is given below.

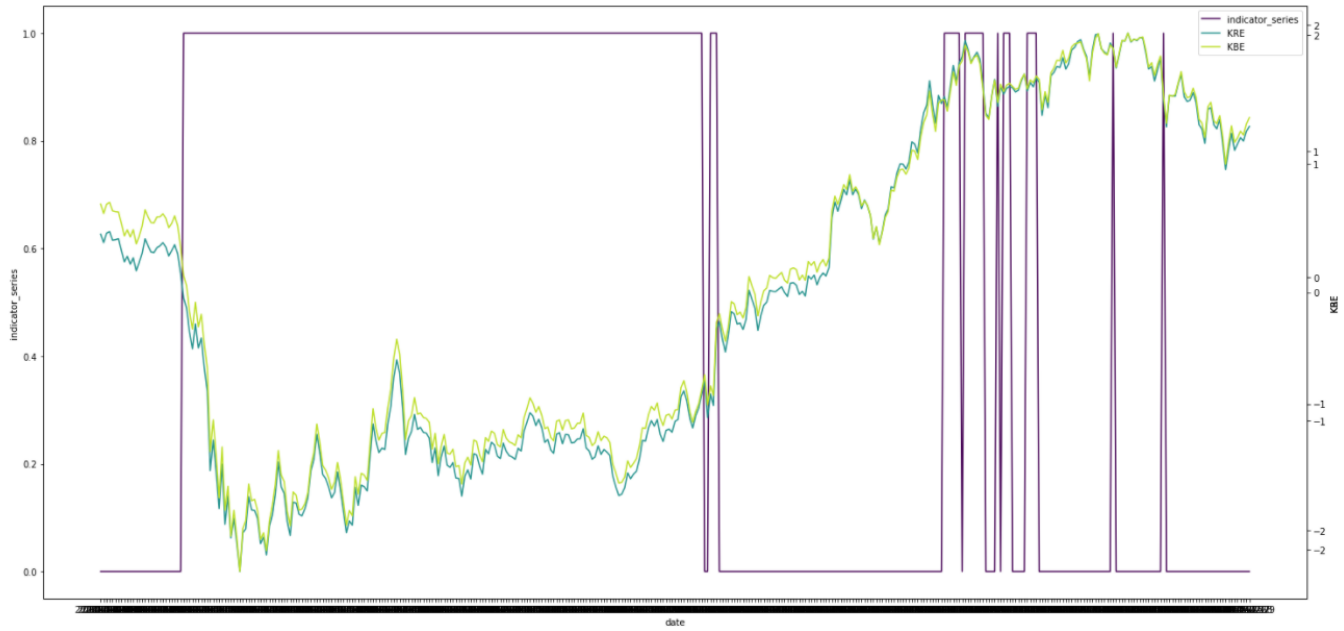


Figure 8: Normalized prices for KBE-KRR along with indicator series

The SPDR® SP® Bank ETF (KBE) seeks to provide investment results that, before fees and expenses, correspond generally to the total return performance of the SP® Banks Select Industry Index (the "Index"). The SPDR® SP® Regional Banking ETF (KRR) seeks to provide investment results that, before fees and expenses, correspond generally to the total return performance of the SP® Regional Banks Select Industry Index (the "Index").

On the contrary, FB-GOOG is an example of a stock pair that failed the cointergration consistency step. (Figure given below).

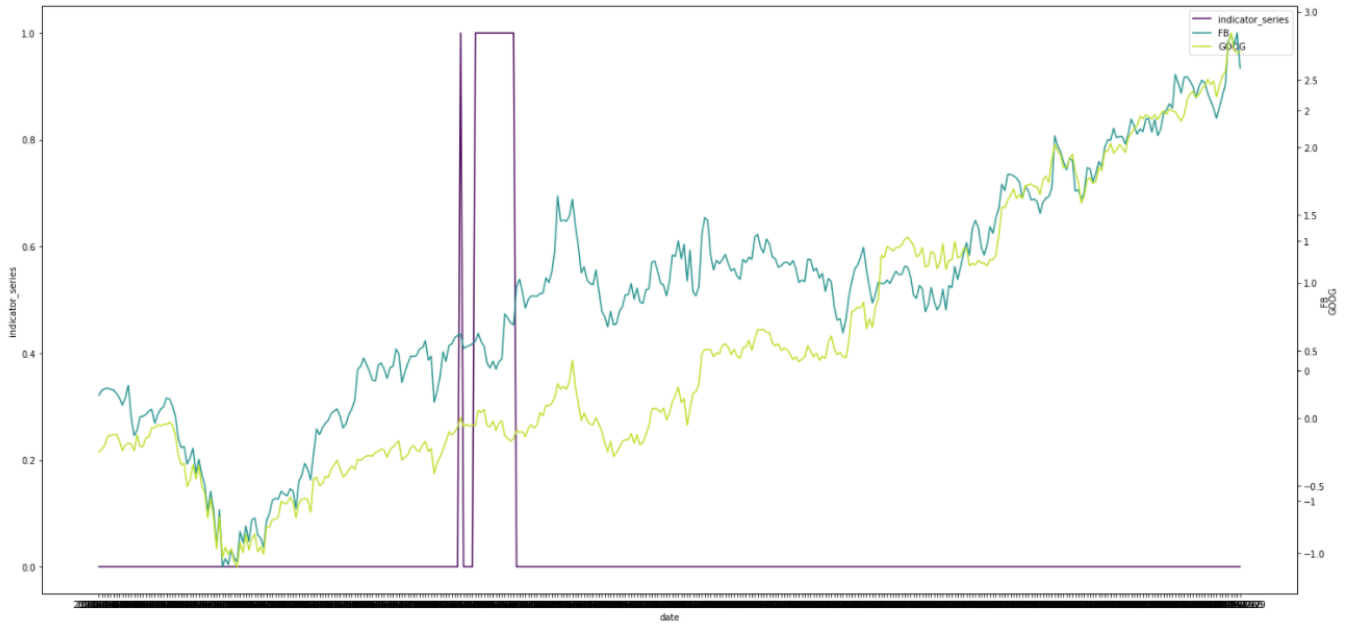


Figure 9: Normalized prices for FB-GOOG along with indicator series

Even though Facebook and Google exhibited a cointegrated relationship in the initial time period, and ticks off all the other statistical and qualitative assessments, it shows no cointegration in recent years and is therefore excluded from our pairs trading strategy.

We have conducted similar assessments on a pair-by-pair basis, and identified 28 cointegrated pairs. We have implemented a holistic framework to identify co-dependent stock pairs, which will further be evaluated using forecasting models. Subsequently, a trading strategy will be constructed using the selected pairs, by identifying optimal entry and exit points for each pair. This will be discussed in the subsequent sections.

# 4 Time-series forecasting

The key innovation idea that we explore as a part of our project is in the execution of pairs trade. In literature there are numerous ways of entering and exiting a pairs trade strategy with most popular one among them being based on thresholds. We formulate the research question "Can a forecasting based trading model achieve better performance?".

## 4.a Overview

In the forecasting based trading model, the main idea is to forecast the spread that is obtained from stationary series after combining the identified pairs. Since the merit of a co-integrating series is that a linear combination of its constituents produces a forecastable series. So given the stationary series, we forecast the next point and compare it with the actual spread. If the difference between the actual spread and forecasted spread is more than historical differences then we enter a trade. There are four steps involved in establishing such a trading model. All of these steps are performed in a rolling fashion over a trading period and training period.

- Generate the residual series on the training period.
- Forecast the next day spread over the trading period.
- Compute the difference between the actual spread and the predicted spread.
- Measure a Z-score on the difference of spreads and compare it with some established thresholds

This method is outlined in the book "A machine learning based pairs trading investment strategy" by Simao Moraes Sarmento and Nuno Horta.

An infographic depicting the time dimension and trading model is provided below.

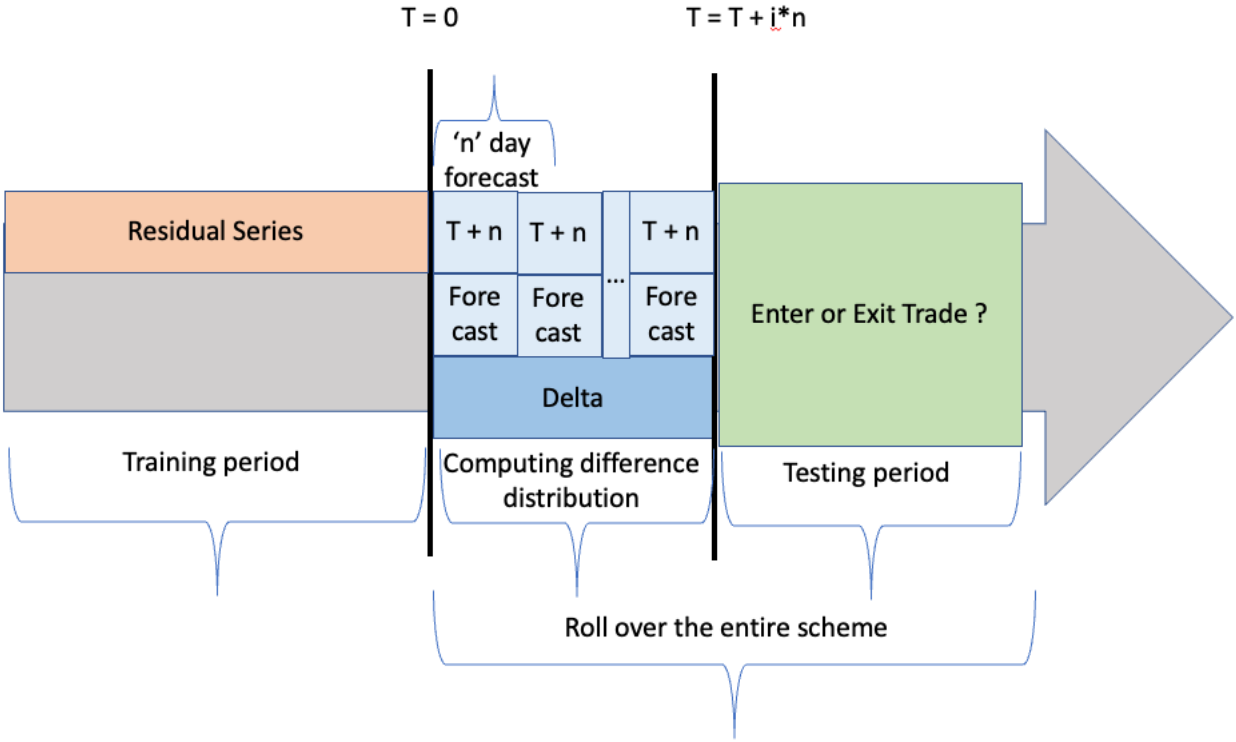


Figure 10: timeline of the executed trading strategy

Note that we take an additional i*n days to generate the distribution of the difference of spreads between the forecasted spread and the actual spread. This distribution is then used for computation of Z score.

## 4.b Forecasting methodologies

After extracting the residual series, the next step is to forecast the next day spread. There are no restrictions when it comes to this area, as the number of options available are numerous. Since for co-integrated series, the residual series is stationary, simplistic forecasting methods can also be utilized. In our work we explore two forecasting techniques, namely we fit an Auto-regressive process of order 1 and we also tryout pmdarima package for automatically identifying and forecasting using an ARMA(p,q) process.

### 4.b.1 AR(1)

This is the most simplistic time series process that could be fit to be stationary process. We use statsmodels's linear regression module and perform the following regression on the residuals

$$R_t = \phi * R_{t-1} + c$$

The forecasts are not very accurate but it provides us a baseline to compare. A snapshot of AR(1) forecast looks like the following.

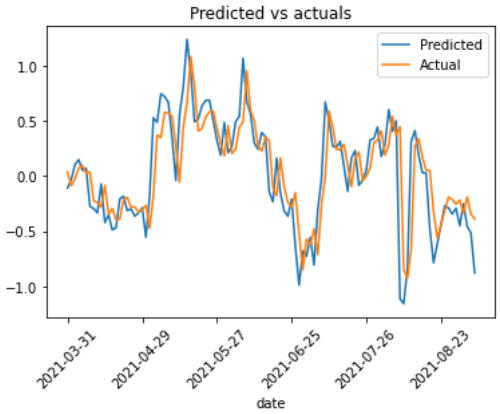


Figure 11: AR(1) Forecasting of residual of FTS and GTES

It can be seen that the forecast essentially lags behind the actual series. The simplest AR1 forecast is simply the AR1 coefficient multiplied by the last time series value available.

### 4.b.2 Auto ARIMA

The stationary time series generated could have any AR or MA orders. Fitting an AR1 model on top is quite a simplistic assumption. This approach could be enhanced by fitting an ARMA model. Fitting an ARMA model could be done through maximum likelihood estimation. Usually the order of the ARMA model is decided through data visualization and analysis however a systematic heuristic could be developed by a grid of AR and MA

orders and it is traversed and the values minimizing a particular metric is chosen as the final value. Usually the metric chosen are AIC/BIC and this methodology is followed by the pmarima package in python. We use the same for our experiments. We don't observe very stark difference as far as forecasting performance is concerned.

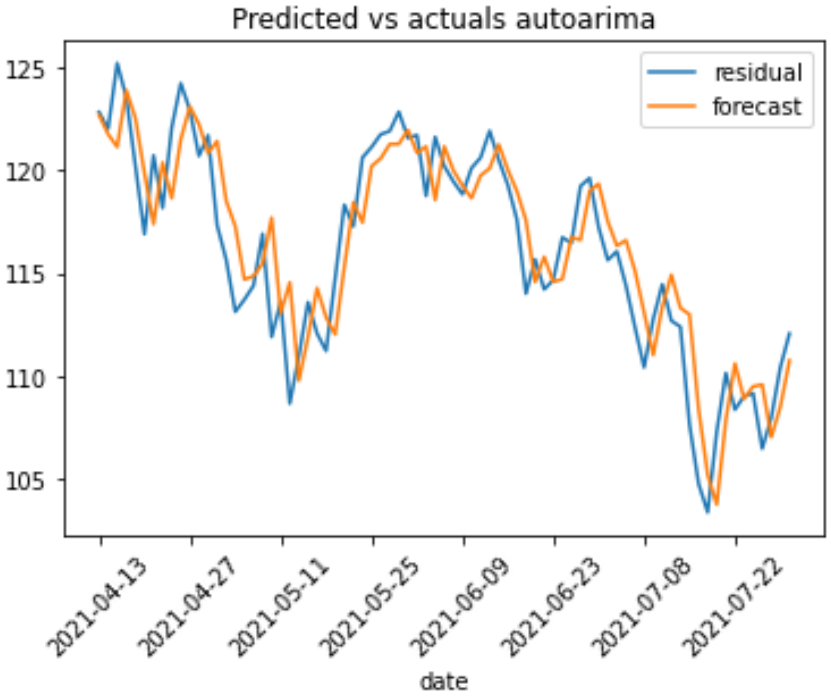


Figure 12: ARMA Forecasting of residual of Analog Devices, Inc. and Microchip Technology Inc.

# 5 Pairs trading strategy

In this section we dive deeper into the trading model we described above. Post the estimation of forecasted spread, we move on to the next step of the trading model where we decide the optimal entry and exit points of the trade

## 5.a Overview

The underlying premise of the forecasting based trading model is the deviation of the difference between the predicted spread and the actual spread. This is close to the difference of difference method popularly used in economic studies. Once we generate our forecasted spread, we compute the difference with the actual spread. We perform this for some 'n' days to generate a difference of spread distribution.

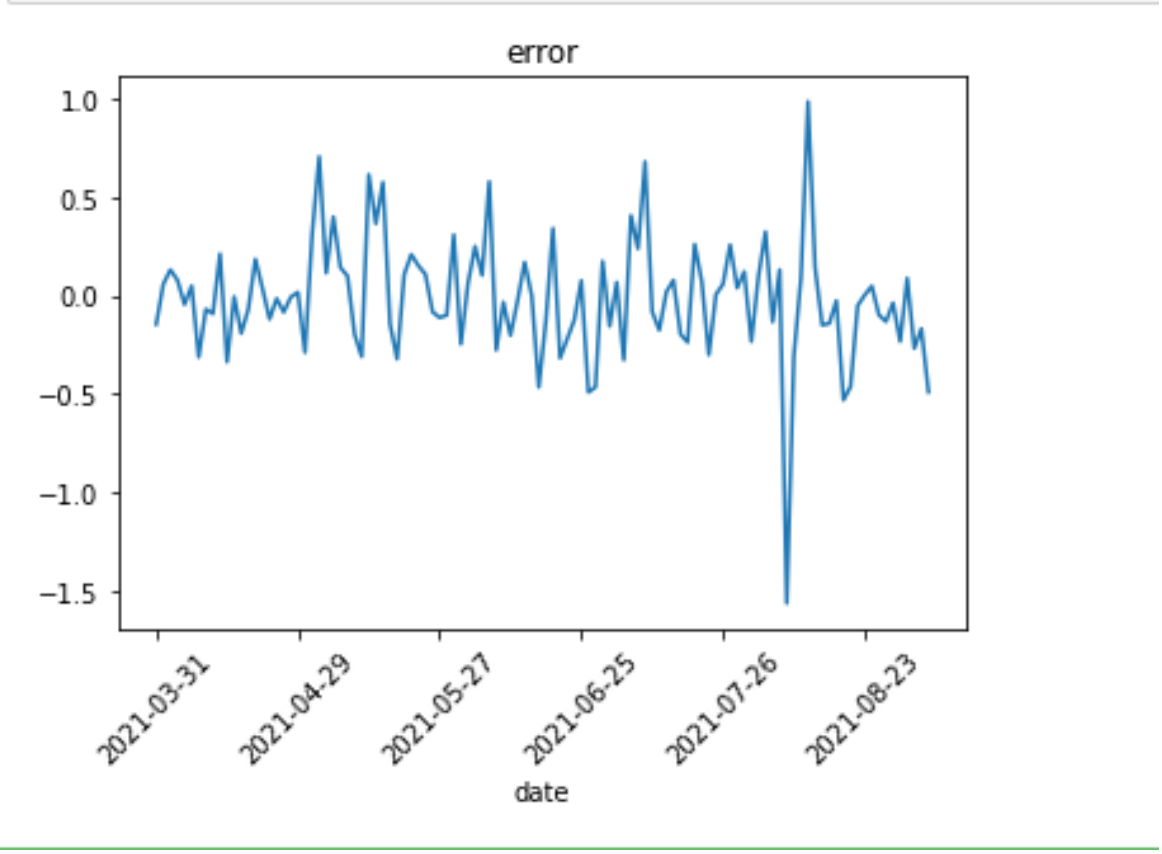


Figure 13: Difference between forecasted spread and actual spread

As it can be seen in the above picture, the difference of difference series is stationary with somewhat fixed upper and lower bounds. However, we do see a case where there is huge deviation and such points are ideally the entry points of our trade. The systematic process to identify such entry points are explained in detail in the following subsection.

## 5.b Optimal entry-exit points

The difference in difference distribution serves as the primary feed to compute mean and deviation for estimation of Z score. Instead of the traditional way of using mean and standard deviation for computation of Z score, we use median and median absolute deviation. Although the Z score obtained from such statistic does not conform to the true meaning of Z score but it gives us a good metric to place thresholds on. The benefit of using median and median absolute deviation over their traditional counterpart is the fact that these metrics are unaffected by the presence of outliers. We experiment over different thresholds and this is hyper-parameter that can be tuned over a validation period to come up with the optimal parameters. As soon as there is a breach in the z score threshold we enter a trade. The trade reversal threshold is also a paramater and can be tuned accordingly

## 5.c Trade Setup

Depending on the definition of the spread (i.e $Y - \beta X$ or $X - \beta Y$) and direction of the spread breach, we take our positions. The direction of the spread is determined by the nature of the equilibrium relationship. If X error corrects on Y, then the spread definition would be $Y - \beta X$ and vice verca. As far as the direction of the spread breach is concerned, in the definition given above, if the spread breach is positive, i.e the Z score of $Y_t - \beta X_t > threshold$, the trading signal for Y would be short and long for X. This indicates that Y is the overperforming stock and needs to be shorted whereas X is the underperforming stock and should be longed. The similar logic would be applied if the breach of the spread where in opposite direction.

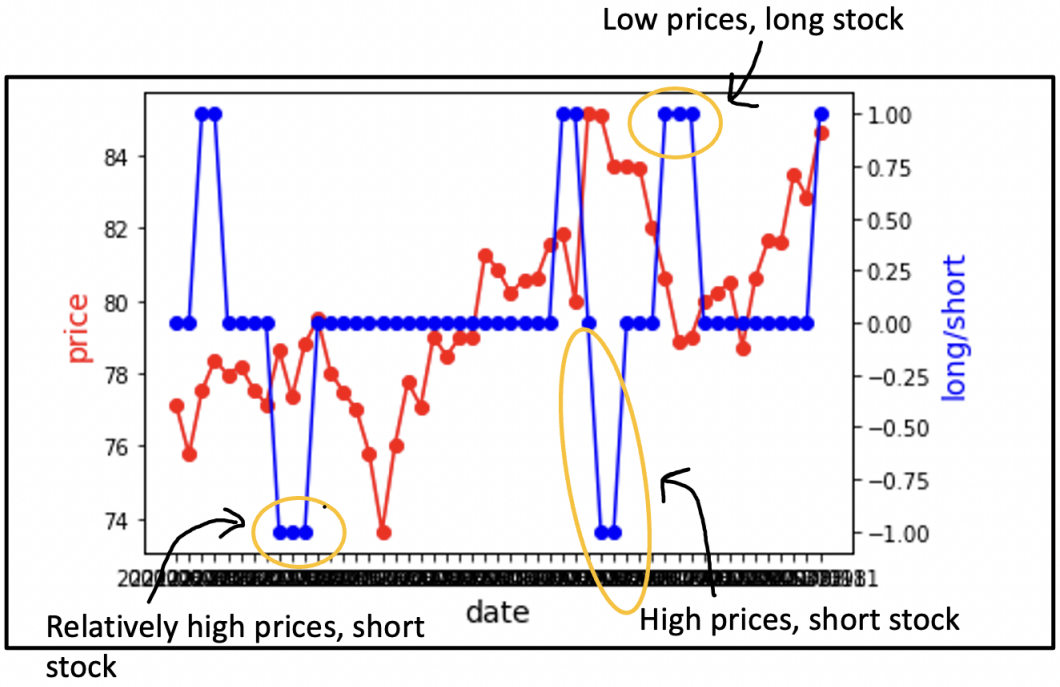


Figure 14: Stock price of Gates Industrial Corp PLC along with trading signal

As we can see from the image above, when the stock price is low, the trading signal is to long the stock and when the stock price is relatively high, it is to short the stock. Although this cannot be viewed in isolation since, the long/short signal depends on the direction of the spread.

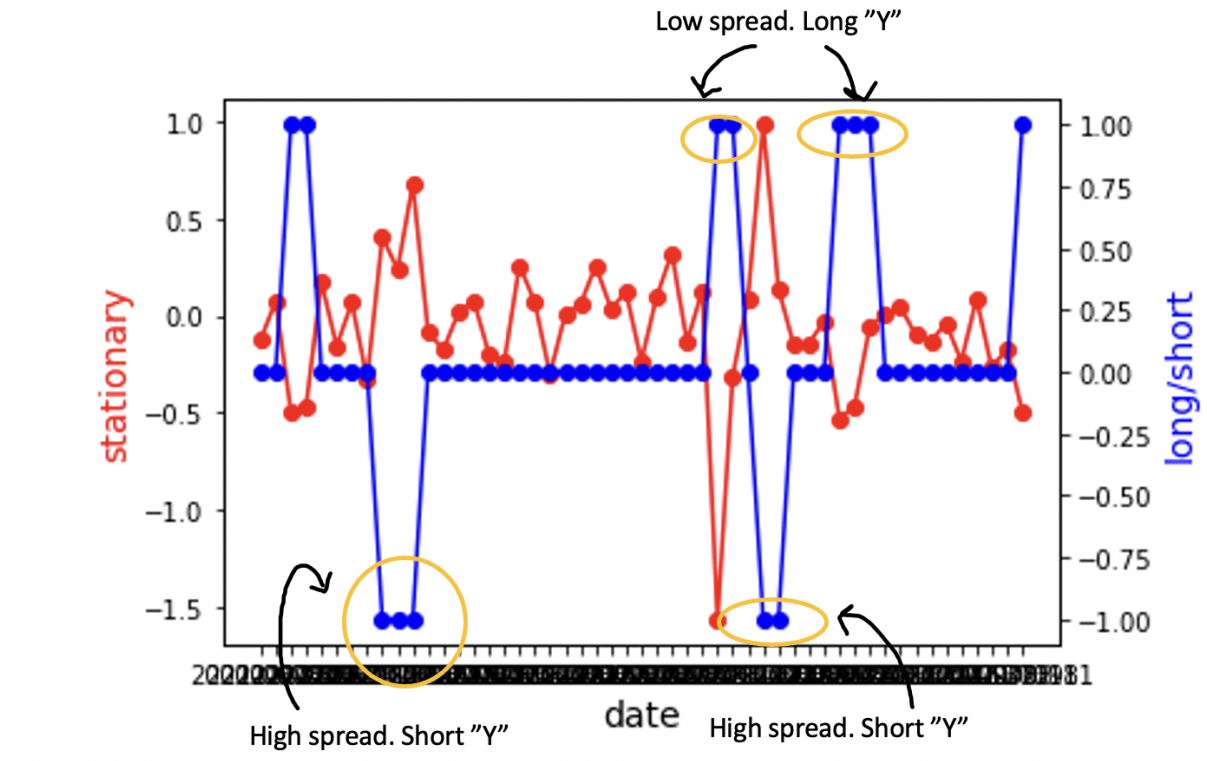


Figure 15: Diff of Diff spread amongst Flowserv Corp and Gates Industrial Corp PLC with trading signal

In above figure, the intuition of the signals is very clear. The trading signals generated are with respect to "Y" or Gates Industrial Corp and whenever the spread breach is positive, it indicates an overperformance of Y with respect to the equilibrium deviation. Hence the trading signal generated is short. Similarly when the spread breach is negative, the price of Gates industrial Corp is too low compared to the equilibrium threshold and hence the position has to be long.

# 6 Results

We chose the top pairs using the heuristic defined above and trade them using the above strategy. The trading setup is as follows:

1. Training period length : 180 days
2. Forecasting horizon : 30 days
3. Distribution computation horizon : 30 days
4. Trading horizon : 60 days

We start at the beginning of the 2019-01-01 and proceed till 2020-08-31. We don't optimize our portfolio amongst the different pairs, however it is something that can be easily done. The overall portfolio could be made beta neutral by assigning an average beta value per pair. Our results are shown below.

Even though the overall portfolio profits are negative, the results are encouraging. We do make positive profits for a lot of the pairs. One risk of the pairs trading strategy is that it is contrarian in its view and the positions taken are against short term momentum. This can sometimes lead to large losses as it could be seen for the first pair. Such returns can be avoided by implementing a stop loss strategy. More investigation needs to be done on the performance of the trade and the trading signals generated. Unlike the pair selection process, the trading process is completely systematic and does not have the benefit of discretionary analysis. Deeper analysis into the trading process is warranted.

Besides, we have barely scratched the surface when it comes to utilizing different forecasting methods. There are very clear two next steps that can be pursued post this result.

1. Optimize the portfolio and make it market factor neutral. This can be achieved by utilizing the rolling betas that we computed. Using a simple optimzation package

| | |
|---|---|
| LRCX-MKSI | -23.479247 |
| LEA-MGA | -6.824521 |
| ITT-PH | -6.094797 |
| HRI-WCC | -5.701347 |
| CRS-KALU | -5.399674 |
| CVX-XOM | -2.992277 |
| CSX-NSC | -2.689174 |
| IP-WRK | -2.678251 |
| ESE-MWA | -2.341569 |
| GGG-NDSN | -2.163873 |
| PTEN-HP | -1.818188 |
| ADI-MCHP | -1.557611 |
| GRC-MLI | -1.504526 |
| OII-WHD | -1.427970 |
| KRE-KBE | -1.335108 |
| VTV-IWD | -0.864696 |
| NUE-STLD | -0.825681 |
| ON-NXPI | -0.720274 |
| RYN-WY | -0.546303 |
| NFLX-FB | -0.054833 |
| IGV-FDN | 0.101826 |
| SCCO-FCX | 0.156228 |
| CDE-HL | 0.287798 |
| HRL-CPB | 1.242028 |
| DHI-LEN | 1.288064 |
| IVV-SPY | 2.324638 |
| IWF-VUG | 3.675761 |
| EXPE-BKNG | 7.614180 |

Figure 16: Total profit per 1 unit of stock of pairs purchased

like cvxpy, a market neutral markowitz portfolio can be created at the beginning of every trading period with different weights given to different pairs. Then the pairs trading strategy could be entered into to gain additional returns on top of created portfolio

2. Better forecasting methods can be utilized like LSTMs or Fbprophet package. AR1 and ARMA methods are very simplistic. It would also be interesting to see, if the stationary residual series can be modeled as a function of other features as well like macro-economic variables etc. This would allow an investor to incorporate some

business regime information in the forecasting as well.

# 7 Conclusion

Pairs trading strategies have been widely used since the 1980s, however, innovations are constantly being explored within the industry to maximize profits in such a trading strategy. There is room for innovation in all three aspects of this strategy - namely the pairs selection process, the forecasting methodology and the entry/exit points of the trading strategy. We have explored all these three aspects, however we have only implemented a bird's eye view and there are endless possibilities of creating a successful pairs trading strategy.

Our focus has primarily been towards implementing a holistic framework, that has both systematic and discretionary aspects in the pairs selection process. We believe that the robustness of the pairs identification process is an integral part for a successful pairs trading strategy. Our assessments can be further extended to incorporate pairs that are market neutral (Beta differential less than a certain threshold) as well as neutral to certain factor exposures. Recently, Graph theory and Page Rank algorithm have been used to identify robust stock pairs and it would be very interesting to explore this avenue.

The difference of difference method is an innovation that has been explored in this project, however, in terms of the forecasting methodologies used, we could take it up a notch by implementing models like LSTM or Fbprophet. Different methodologies can also be explored to identify the trading strategy itself and optimal entry/exit points. A stop loss strategy can be additionally implemented to minimize losses for cases where this strategy fails due to unprecedented momentum of the selected pairs. Portfolio construction/optimization can add another level of complexity to ensure profits at a portfolio level.

Various aspects of our strategy uses certain hyperparameters/threshold that can be tuned through rigorous cross-validation and testing. This can certainly add another level of robustness in estimating profit/loss through our implemented trading strategy. The initial

results are promising and these avenues should most certainly be explored to make most of our proposed pairs trading strategy.

# 8 References


1. Pairs trading quantitative method and analysis by ganapathy vidyamurthy

2. A machine learning based pairs trading investment strategy by Simao Moraes Sarmento

3. Krauss, Christopher (2015) : Statistical arbitrage pairs trading strategies: Review and outlook, IWQW Discussion Papers, No. 09/2015, Friedrich-Alexander-Universität Erlangen-Nürnberg, Institut für Wirtschaftspolitik und Quantitative Wirtschaftsforschung (IWQW), Nürnberg

4. Galenko, Alexander and Popova, Elmira and Popova, Ivilina, Trading in the Presence of Cointegration (October 22, 2007). Available at SSRN: https://ssrn.com/abstract=1023791 or http://dx.doi.org/10.2139/ssrn.1023791

5. Gatev, Evan and Goetzmann, William N. and Rouwenhorst, K. Geert, Pairs Trading: Performance of a Relative Value Arbitrage Rule (February 2006). Yale ICF Working Paper No. 08-03, Available at SSRN: https://ssrn.com/abstract=141615 or http://dx.doi.org/10.2139/ssrn.141615

6. Innovations in pairs trading Wolfe research

7. Thomaidis N.S., Kondakis N., Dounias G.D. (2006) An Intelligent Statistical Arbitrage Trading System. In: Antoniou G., Potamias G., Spyropoulos C., Plexousakis D. (eds) Advances in Artificial Intelligence. SETN 2006. Lecture Notes in Computer Science, vol 3955. Springer, Berlin, Heidelberg. https://doi.org/10.1007/1175291277

8. Pairs selection and outranking: An application to the SP 100 index, European Journal of Operational Research, Elsevier, vol. 196(2), pages 819-825, July.

9. http://alkaline-ml.com/pmdarima/

10. https://www.alphavantage.co/documentation/