The project intends to identify leading indicators present in the markets for a given sector of stocks. A leading indicator is a measurable factor that changes before a particular trend starts and thus is widely used to predict future changes. In this project, we try to identify different stocks as the leading indicators for their particular sectors using statistical tests like Granger Causality and Co-integration test. Using these we try to model a time series which we expect would act as a leading indicator for that particular group of stocks. We then use this indicator to predict the sudden changes in the market.
Contents

1 Introduction .................................................. 2
  1.1 Project Aim .................................................. 2
  1.2 Steps Involved .............................................. 2

2 Granger Causality ............................................ 2
  2.1 Introduction .................................................. 2
  2.2 De-trending data ............................................. 2
  2.3 Results ..................................................... 3
  2.4 Inferences ................................................... 3

3 Multiple Correlation and Co-integration ......................... 3
  3.1 Multiple Correlation ....................................... 3
  3.2 Co-integration .............................................. 4
  3.3 Augmented Dickey Fuller Test ............................. 4
  3.4 Results .................................................. 4
  3.5 Inferences ................................................ 4

4 The Leading Indicator .......................................... 5
  4.1 Designing the Indicator .................................... 5
  4.2 Surge and Crisis ........................................... 5
  4.3 When k varies .............................................. 5
  4.4 Percentage Change ........................................ 6
  4.5 Duration Of Change ...................................... 6

5 Conclusion .................................................. 7

6 A suitable indicator ........................................... 7

7 Future Work .................................................. 7
1 Introduction

1.1 Project Aim
We aim to identify leading indicators for stocks in various sectors and compare their performance in predicting sharp changes in the market. We look to design leading indicators based on statistical inference from the daily stock market data and use it to predict sharp falls or rises that might take place in the other stock prices.

1.2 Steps Involved
- We start by normalizing the stock price to the range (0,1). We do this so that the the large change in value of a stock with high price is not seen compared to the changes in other stock prices which are not so high.
- We then apply the statistical tests to identify the leading indicators. We start by looking at Granger Causality. Granger causality is a statistical test to check if one of the time series ‘causes’ the other.
- We then move ahead to looking at our second statistical test. We use co-integration based tests. Co-integration test is used to check whether a particular time series can be expressed as a linear combination of other time series or not.
- After the statistical tests we look to model our leading indicator based on the tests. We look at the possible indicators from the statistical tests and use them to design the leading indicator.
- We then use these leading indicators to predict the sharp rises and falls, called surge and crisis respectively, in the other stock prices.
- Finally, in this project we look at various model parameters and check which of our leading indicators perform by varying those parameters.

2 Granger Causality

2.1 Introduction
Granger Causality is a statistical test to check if a time series is a caused by another time series. It helps one identify if one time series can be helpful in forecasting the other. However Granger causality requires the underlying time series to either be stationary or interrogated of a fixed order. Granger Causality is often helpful to identify the lag values at which the dependent variables will be caused. It takes a null hypothesis that the time series does not cause the other and it is rejected based on the p-value.

2.2 De-trending data
The time series we are using is a stock price data. It is bound to have some trend to at least maintain the interest rates. Thus we look to de-trend the data by applying first differencing. We use the following equation and test the resultant series for stationarity.

\[ \nabla (X_t) = (X_t) - (X_{t-1}) \tag{1} \]

After de-trending the data we get pretty close to a stationary series. We look at two plots for the same. the plots are obtained by single differencing on DB and Bank of America stock market data.
Figure 1: Effect of differencing on stock prices

2.3 Results

For the 8 stocks we looked at in the banking sector the following was the trend observed for causing and being caused. We rejected null hypothesis that X did not cause Y at p<0.05.

<table>
<thead>
<tr>
<th>STOCK</th>
<th>NO. OF TIMES IT WAS CAUSED</th>
<th>NO. OF STOCKS IT CAUSED</th>
</tr>
</thead>
<tbody>
<tr>
<td>jpm</td>
<td>2/7</td>
<td>4/7</td>
</tr>
<tr>
<td>db</td>
<td>5/7</td>
<td>3/7</td>
</tr>
<tr>
<td>hsbc</td>
<td>3/7</td>
<td>6/7</td>
</tr>
<tr>
<td>ms</td>
<td>4/7</td>
<td>3/7</td>
</tr>
<tr>
<td>gs</td>
<td>6/7</td>
<td>3/7</td>
</tr>
<tr>
<td>cs</td>
<td>4/7</td>
<td>5/7</td>
</tr>
<tr>
<td>bac</td>
<td>3/7</td>
<td>3/7</td>
</tr>
<tr>
<td>wfc</td>
<td>4/7</td>
<td>4/7</td>
</tr>
<tr>
<td>Total</td>
<td>31/56</td>
<td>31/56</td>
</tr>
</tbody>
</table>

Table 1: Granger Causality Results

2.4 Inferences

From the above experiment we note that GS and DB are the most likely to be leading indicators for this particular group of stocks. GS causes 6 out of the 7 stocks and DB causes 5 out of the 7 possible. Also we note that some stocks are pretty similar to each other. For example, some stocks like BAC and MS are caused by and cause nearly the same stocks. Overall out of the 56 possible causations 31 satisfy the Granger Causality test, i.e, nearly half of the times some sort of a causation was identified. Also interesting to note is the fact that actually no stock was caused by or caused less than 2 stocks. This implies that this group of stocks is definitely highly related and have strong causations among each other. However, GS and DB are the most probable candidates for being the leading indicator.

3 Multiple Correlation and Co-integration

3.1 Multiple Correlation

Multiple Correlation is a statistical technique which tries to model a dependent variable as a linear function of other independent variables. The loss function that we use is the ordinary least square error. Here, we try to model each stock using only the lagged values for different stocks. It is
important that we do not use the current values of any of the stocks and use only the lagged values. This is because for any stock the current value will be known along with all the other current values. Thus, it doesn’t make sense to predict the value after the value is known. The value of lag used for the test is 1 as we will have 7 independent variables to model the dependent future value of the stock being modeled.

3.2 Co-integration

Co-integration is a statistical test that is used to check whether the linear combination of two time series is stationary or not. It requires that the two underlying time series are integrated of order 1. This holds in our case as has already been shown using differencing.

Since, our time series are also integrated order 1 we try to look at the co-integration coefficient to check the linear model we constructed above using lagged values. For this we use the residuals of the model and check for its To check for co-integration for our time series we look at the residuals from the multiple correlation seen and try to check for their stationarity.

3.3 Augmented Dickey Fuller Test

We use the Augmented Dickey Fuller test to check for the stationarity of the time series. The null hypotheses for the Augmented Dickey Fuller test is that a unit root is present in the time series sample, i.e, the time series is non stationary. The alternate hypotheses thus is that the time series is stationary. We look at the p-values and reject null hypotheses for p<0.05. We use the test to check for stationarity in both the cases: after differencing and for the residuals of the multiple correlation model. the results for this set of tests is presented below.

3.4 Results

We get the following linear models for the same 8 stocks as mentioned above. Also shown are the corresponding p-value for the Augmented Dickey Fuller Test on the residuals of the linear fit.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Significant pieces of Linear Fit</th>
<th>p-Values for residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>jpm</td>
<td>0.41 cs + 0.38 db</td>
<td>0.18</td>
</tr>
<tr>
<td>db</td>
<td>0.66 gs + 0.39 jpm</td>
<td>0.0004</td>
</tr>
<tr>
<td>hsbc</td>
<td>1.08 wfc + 0.2 gs</td>
<td>10^{-4}</td>
</tr>
<tr>
<td>ms</td>
<td>1.16 gs + 0.32 bac</td>
<td>0.012</td>
</tr>
<tr>
<td>gs</td>
<td>0.37 ms + 0.30 db</td>
<td>0.016</td>
</tr>
<tr>
<td>cs</td>
<td>0.30 db + 0.6 jpm</td>
<td>0.027</td>
</tr>
<tr>
<td>bac</td>
<td>0.68 ms + 0.34 gs</td>
<td>0.024</td>
</tr>
<tr>
<td>wfc</td>
<td>0.47 hsbc + 0.16 db</td>
<td>10^{-7}</td>
</tr>
</tbody>
</table>

Table 2: Multiple Correlation and Co-integration Results

3.5 Inferences

Most of the commodities we looked at satisfied the Co-integration Test. This implies that most of them can be satisfactorily written as a linear combination of the lagged values of remaining stocks. On the other hand, MS and GS had high positive coefficients in most of the linear regressions showing that they are good candidates for the Leading Indicators. Thus we conclude the Multiple Correlation and Co-Integration test by noting that the possible candidates for a leading indicator would be GS and MS.
4 The Leading Indicator

4.1 Designing the Indicator

Having done the statistical tests and having seen their results we move on to designing the indicator based on those results. We design a leading indicator as follows:

- Randomly choose a weight $k$ in the range $(0,1)$.
- Pick two stocks and assign them weights $k$ and $1-k$.

Then we move to checking which of the combinations gave us the best results in terms of predicting the sharp changes in other stocks. We now look at some factors we used in our definition of surge and crisis and their effect on overall prediction.

4.2 Surge and Crisis

We define Surge as a condition when the price of a commodity rises more than a certain threshold percentage, we call this percentage as percentage change, in a certain threshold duration, where the duration is called duration of change. Similarly, we define Crisis as a condition when the price of a commodity falls more than a certain threshold percentage in a certain threshold duration.

4.3 When $k$ varies

(a) Different values of $k$ for $k(GS)+(1-k)(DB)$
(b) Different values of $k$ for $k(GS)+(1-k)(MS)$
(c) Different values of $k$ for $k(DB)+(1-k)(MS)$
(d) Different values of $k$ for $k(HSBC)+(1-k)(JPM)$

Figure 2: F1 values on prediction of surge and crisis

In the above graphs we have given $k$ weight to the left stock and $(1-k)$ to the right. It is plotted against the $F1score = 2 \ast precision \ast recall / (precision + recall)$. 
We can see that varying $k$ clearly shows us which of the two in contention are a better indicator or if a mix of two can give us a better result. Looking at the results we can infer that MS and DB are better indicators than GS and a combination of MS and DB gives us a very good leading indicator.

### 4.4 Percentage Change

Percentage Change, as described above, is used in the definition of crisis and surge. A higher percentage change should give us fewer cases where such crisis or surge are identified and a fewer number of cases when they actually happen. For the following graphs we use GS and $(DB*0.42+MS*0.58)$ as our leading indicators.

![Figure 3: Effect of percentage change on identification](image1.png)

(a) F1 score for (MS+DB) Leading Indicator with % change
(b) F1 score for GS Leading Indicator with % change

Here we see that for lesser values the score is high and it stabilizes as we go ahead. This maybe because of the fact that a small change of 1% is too frequent giving us a good overall result.

### 4.5 Duration Of Change

The duration of prediction is another parameter to define a crisis/surge. Note that we have fixed the prediction distance to 1 as we only want to see if the leading indicator can predict 1 day in advance. Since, the leading indicator is itself a stock price expecting one price to lead the other by a bigger margin is not justified.

![Figure 4: Effect of duration of change on identification](image2.png)

(a) F1 score for GS Leading Indicator
(b) F1 score for MS+DB Leading Indicator

Figure 4: Effect of duration of change on identification
Based on these graphs we can say that the performance of the leading indicator: DB*0.42 + MS*0.58 was better than that of GS as an indicator alone. The test was done at per change = 2

5 Conclusion

- Applying granger causality can give us an idea of the leading indicator as most stock prices data are stationary after first differencing.
- The co-integration result showed that for most stock prices we can satisfactorily express them as a linear combination of lagged values of other stock.
- We looked at the overall effect in identification of surge and crisis of the different variables involved.

6 A suitable indicator

Leading indicators for stock prices are usually made from other macroeconomic factors like GDP, Global Oil Prices, Currency Strength and Interest Rates. Using stock prices as leading indicators does give some results but it will definitely not be able to capture the reasons from which the market changes as it is itself a result of those changes. Also such a leading indicator is bound to vary with time and different stocks can act as leading indicators at different times. The above presentation had results averaged out for a period of 4 years.

7 Future Work

- A similar approach can be used to find the leading indicators in other sectors of the markets as well.
- Inclusion of a few macroeconomic factors along with stock prices and their effect can be studied.
- An on-line algorithm to look at the data and find the leading indicator can be studied upon as the leading indicator changes with time.
References


