Week 09/16 Summary

***VPIN/SOI TEAM:***

1. We tested the agreement rate between Lee Ready method and assigning buy/sell from the closest quotes method.

* We improved the speed to calculate the assigning buy/sell from the closest quotes method.
* The agreement rate is ~60% for a 60/1000 AMZN run.

1. We use AMZN 23th data to test the concurrent relationship between price return and SOI. We got the best result with 40+% adjusted R^2 for Lee-Ready method.

* We find the best results by using loop to find the best combination of the three parameters: time bin, threshold value (trades with size above which are excluded), and bucket size.

First decide the range of three parameters to test (start, end, increment)

* Bucket size: (1000,10000,1000)
* Time bin: (15,150,15)
* Threshold size: (1000,5000,1000)

\*Given density of trades size (please see in appendix plot 1)

* Formula: Bucket log price return ~ SOI

\*SOI adjusted by the bucket volume, thus range is [-1, +1]

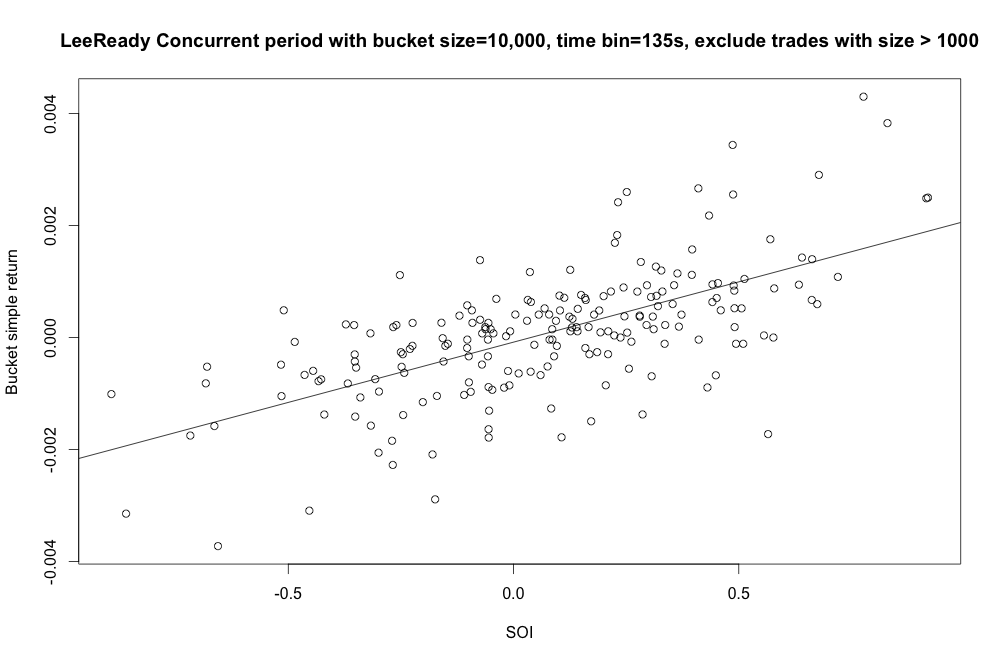
In total, we tested 500 sets of the parameters. 2 of them have R^2 >40%, 54 of them (10.8%) have R^2>30%. Major findings and plots are as the following:

* Best 10 results:

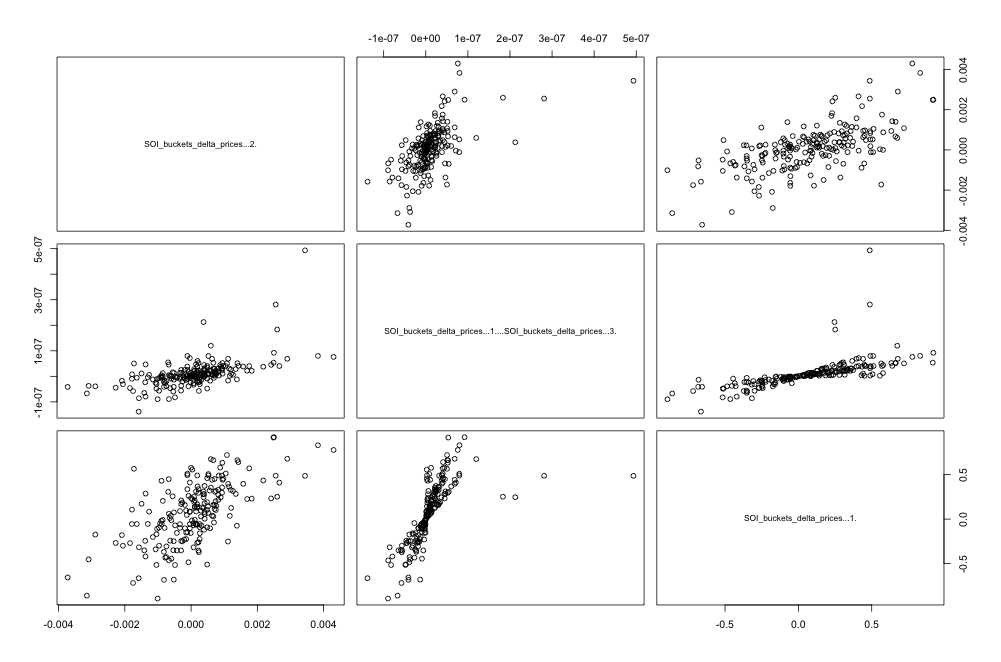
|  |  |
| --- | --- |
| **BucketSize\_TimeBin\_Threshold** | **R^2** |
| 10000 \_ 135 \_ 1000 | 41.30% |
| 10000 \_ 150 \_ 1000 | 41.30% |
| 10000 \_ 120 \_ 1000 | 37.40% |
| 8000 \_ 135 \_ 1000 | 37.10% |
| 9000 \_ 120 \_ 1000 | 36.60% |
| 10000 \_ 105 \_ 4000 | 36.50% |
| 8000 \_ 150 \_ 1000 | 36.20% |
| 10000 \_ 105 \_ 1000 | 36.20% |
| 8000 \_ 120 \_ 1000 | 36.10% |
| 10000 \_ 90 \_ 1000 | 35.60% |

Looking at the best combinations, we find that bucket size over 8000/ time bin over 105/threshold of 1000 generally produce good results. Moreover, it is very interesting that the combination of Bucket Size of 10000 and Threshold 1000 consistently outperforms other combinations. And in this combination, the time bin length is less relevant.

* Take a look at the best result:



Add the cross product of bucket price volatility and SOI as additional independent variable. Variables, as in the sequence below, are: Bucket return, cross product, and SOI.



The plots indicate very good linear relationship between variables.

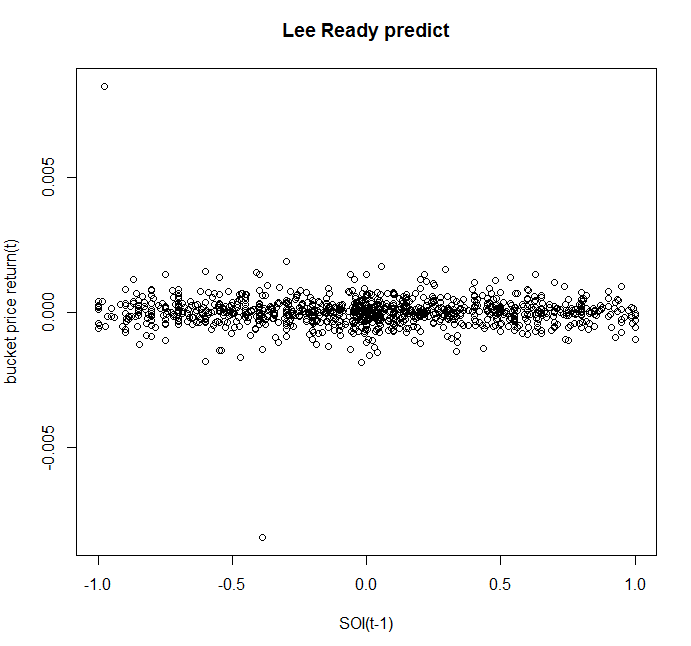
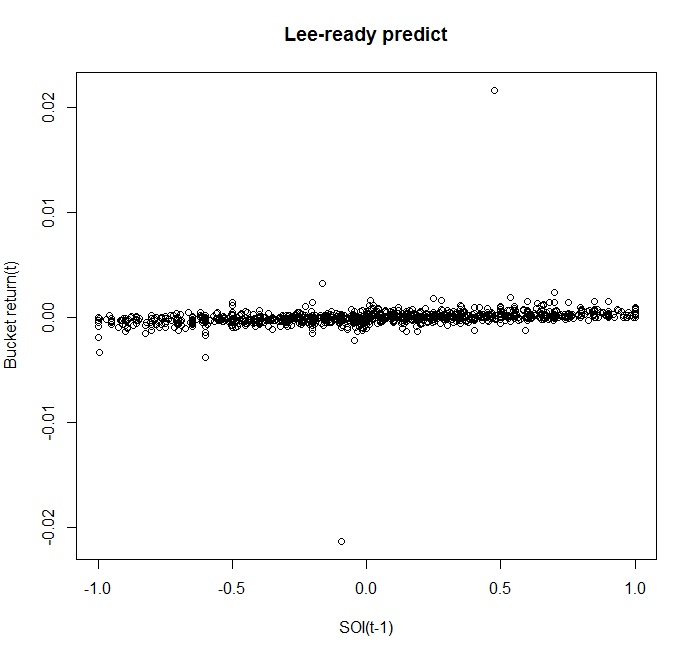
1. Forecast with Lee-ready:

Formula: Bucket returnt ~ SOIt-1

|  |  |
| --- | --- |
| BucketSize\_TimeBin\_Threshold | R^2 |
| 2000 \_ 30 \_ 4000 | **43.6%** |
| 2000 \_ 60 \_ 4000 | **43.6%** |
| 2000 \_ 150 \_ 4000 | **43.2%** |
| 2000 \_ 120 \_ 4000 | **43.1%** |
| 2000 \_ 90 \_ 4000 | **43.1%** |
| 2000 \_ 90 \_ 2000 | **15.4%** |
| 2000 \_ 120 \_ 2000 | **13.8%** |
| 2000 \_ 30 \_ 2000 | **13.0%** |
| 3000 \_ 90 \_ 3000 | **7.8%** |
| 1000 \_ 30 \_ 2000 | **7.7%** |

Though we get some seemingly good result, they are actually inflated by the influential points as the plot below. Thus the result is much weaker than the concurrent one, which is can be intuitively explained, that the order imbalance have direct greater influence on the current bucket.

(2000 \_ 30 \_ 4000) (2000 \_ 90 \_ 2000)

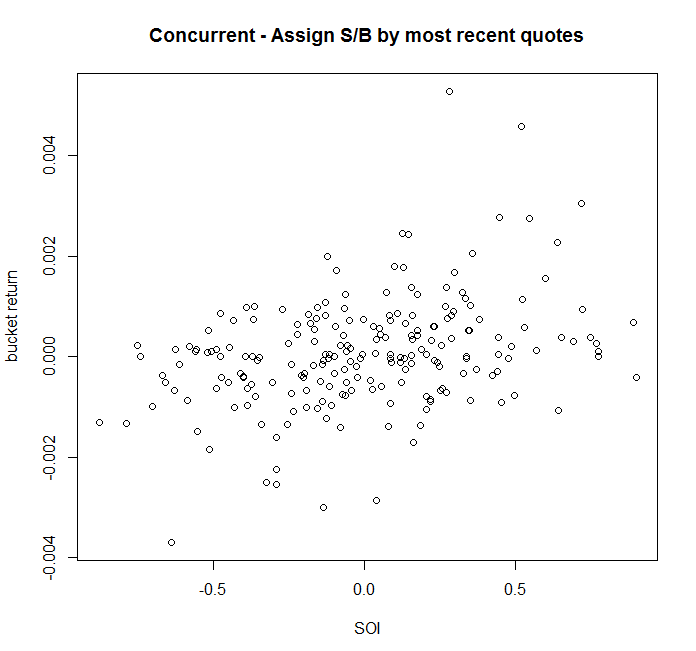


Exclude the influential points, the relationship between the SOI(t-1) and Price return(t) is randomly distributed around zero. This indicates that the predict power of Lee-Ready method calculated SOI is weak.

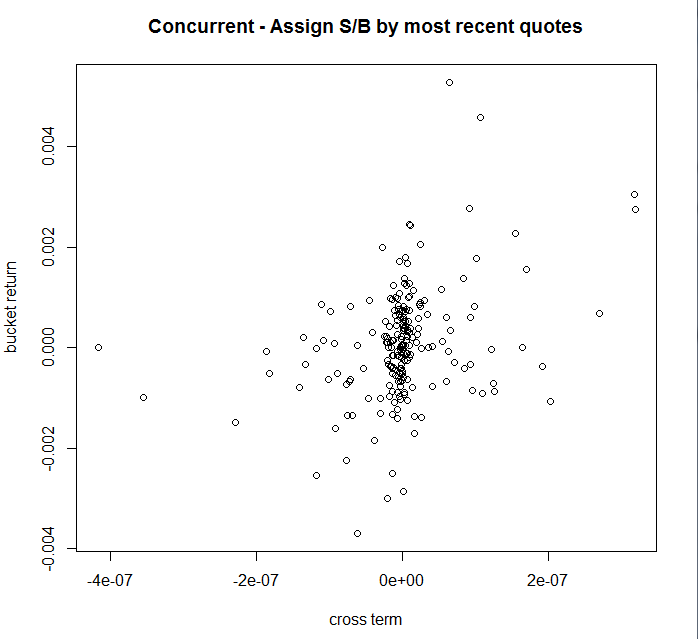
1. Use same approach as in 2 for assigning buy/sell by quotes.

|  |  |
| --- | --- |
| **BucketSize\_TimeBin\_Threshold** | **R^2** |
| 10000 \_ 60 \_ 3000 | 12.5% |
| 6000 \_ 30 \_ 1000 | 11.4% |
| 6000 \_ 150 \_ 1000 | 9.5% |
| 10000 \_ 60 \_ 5000 | 9.5% |
| 5000 \_ 60 \_ 1000 | 9.5% |
| 8000 \_ 60 \_ 5000 | 9.2% |
| 7000 \_ 60 \_ 5000 | 8.9% |
| 6000 \_ 60 \_ 1000 | 8.9% |
| 4000 \_ 60 \_ 1000 | 8.8% |
| 10000 \_ 90 \_ 5000 | 8.6% |

Best result plot as the follow:



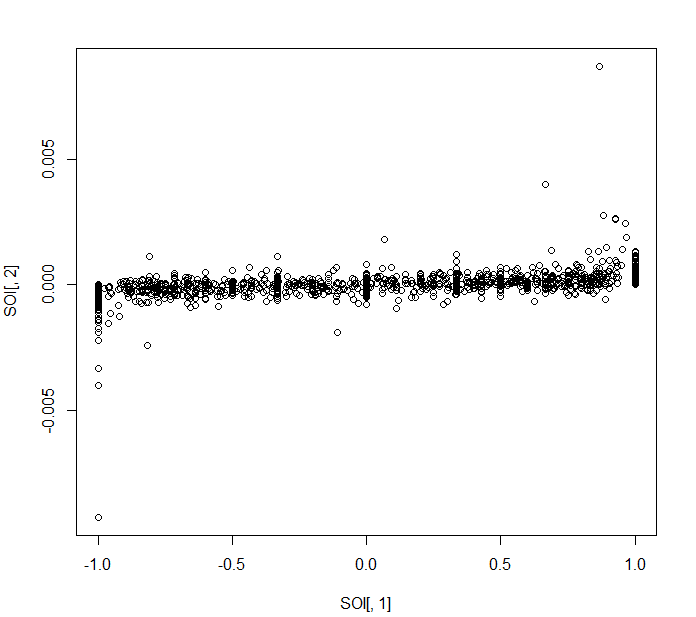
Including the cross term doesn’t help increase the adjusted R^2.



And using quotes based classification methodology produces worse results both in terms of concurrent period regression and one period ahead prediction. This experiment shows Lee Ready is perhaps more accurate in assessing buy/sell direction.

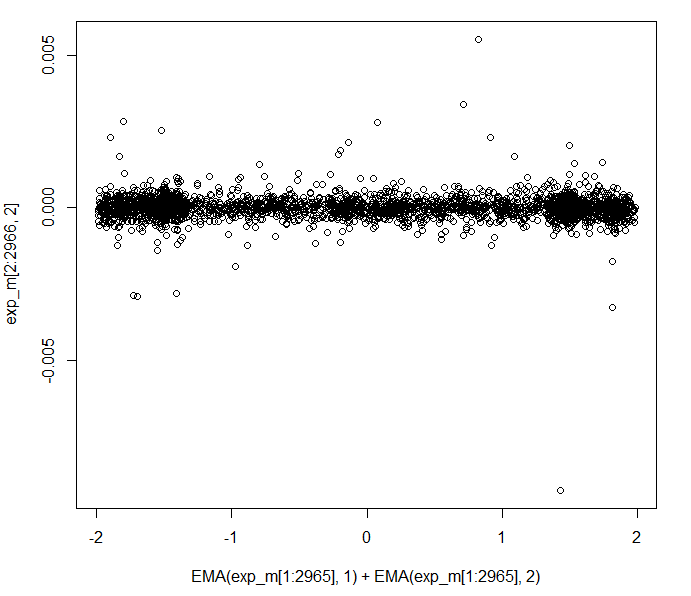
1. Time bucket concurrent analysis:

If we digress from VPIN methodology of calculating SOI, and instead define SOI using time buckets, we find the highest concurrent regression R^2 dropped to 0.233, with a time bucket size of 150s. Note when we choose the time bucket bin size, if the time bin size is large, we are unlikely obtain strong prediction results even if we obtain meaning same period association between SOI and price change.



From the graph, we can conclude that the association between SOI calculated using time buckets and concurrent period price change is not strong. And using an extended model

Using the same time bucket size of 150s, the model gives an R^2 that is virtually 0. Below is the plot:

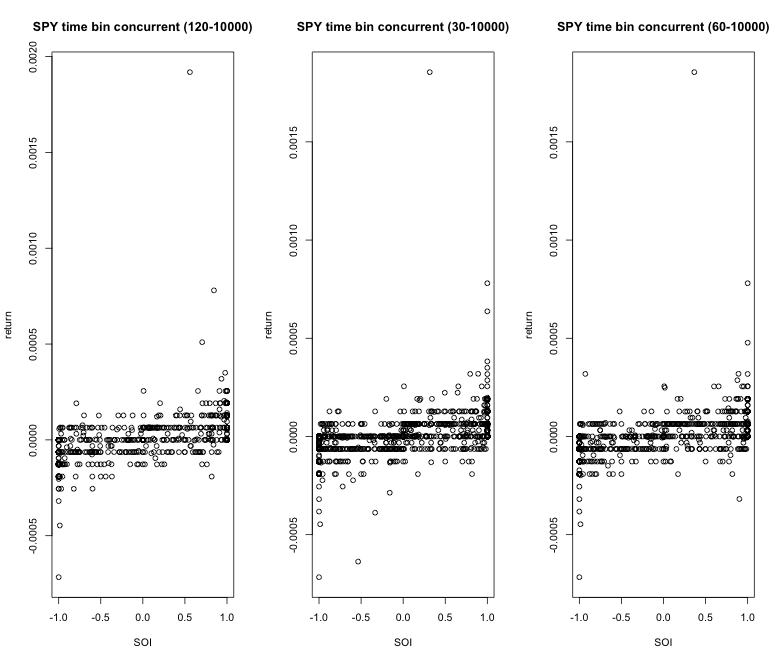


And we conclude this extended model does not have any predictive power, either.

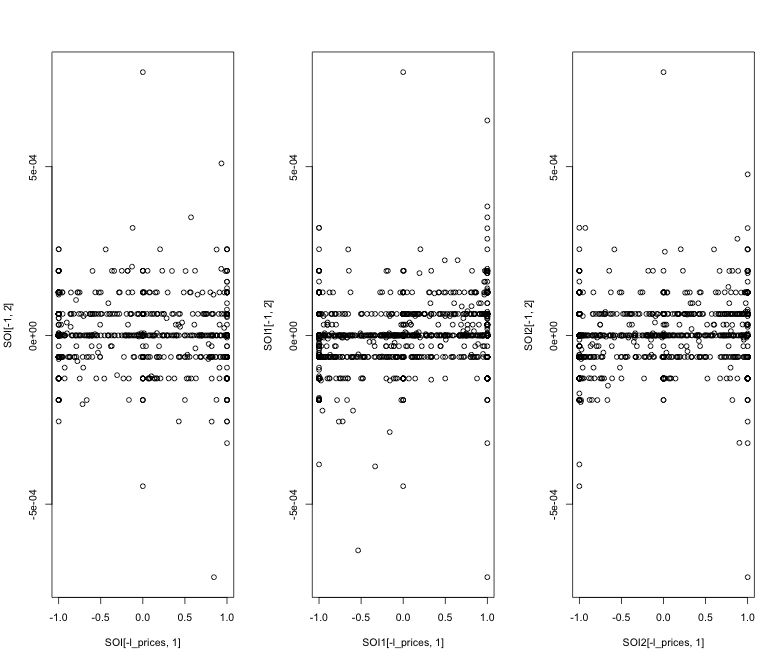
1. Time bin concurrent analysis on SPY. Due to the huge dataset quantity, we are only using the first 20K trades to assess feasibility of forecasting.

|  |  |
| --- | --- |
| **Time\_Threshold** | **R^2** |
| 120 \_ 10000 | 23.3% |
| 30 \_ 10000 | 23.0% |
| 60 \_ 10000 | 22.8% |
| 90 \_ 10000 | 22.8% |
| 180 \_ 10000 | 22.8% |
| 150 \_ 10000 | 21.1% |
| 30 \_ 20000 | 5.2% |
| 120 \_ 20000 | 4.7% |
| 90 \_ 20000 | 4.6% |
| 180 \_ 20000 | 4.6% |

The scatterplots of three best results:



Forecast:



Although the contemporaneous regression does have significant R^2, the 1 period forecast yield R^2 ~ 0. Thus we can conclude SOI does not have much value in forecasting one period ahead price changes.

Conclusion:

Even if SOI does have some contemporaneous significance with price changes, such effect evaporates immediately in the next period shown by ~0 R^2 in all the forecasting attempts. For securities with thinly traded volume, and when price changes are relatively infrequent, in order to obtain a significant contemporaneous regression result, we must either 1) increase the volume size (for volume bucket SOI), or 2) increase the time bucket size (for time bucket SOI). For frequently traded SPY, note the contemporaneous significance is not affected much by the time bucket size at all.

We would also be tempting to conclude that a larger bucket size is likely to increase contemporaneous R^2, but reduce forecasting accuracy; yet given there is no forecasting power in SOI at all, such intuition is not of much use at this point.

Appendix:

Plot 1: AMZN trades size density

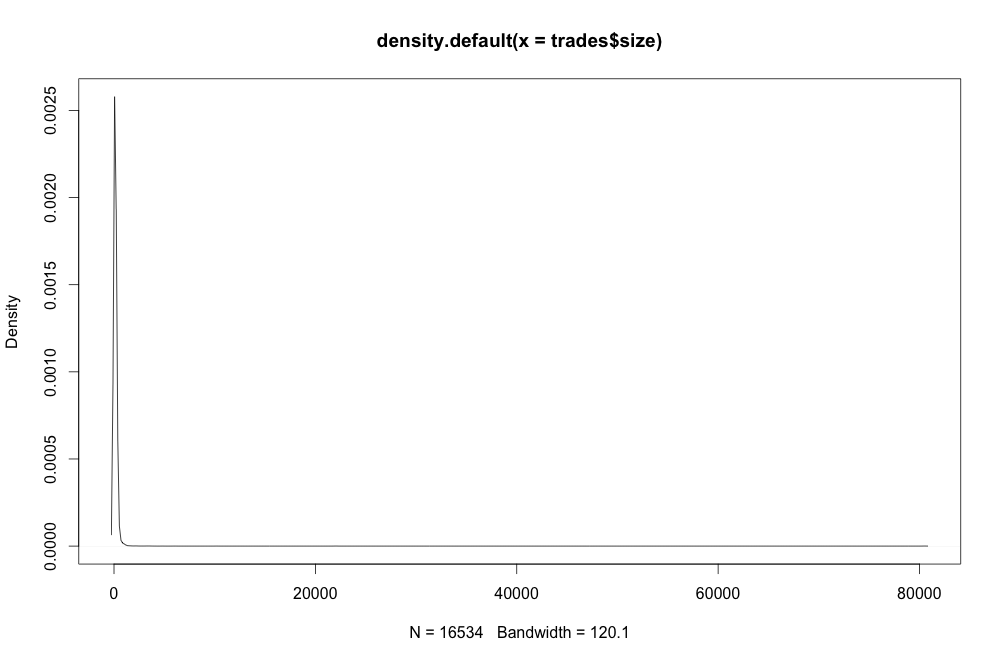


Table 1: Best results:

|  |  |
| --- | --- |
| **BucketSize\_TimeBin\_Threshold** | **R^2** |
| 10000 \_ 135 \_ 1000 | 41.30% |
| 10000 \_ 150 \_ 1000 | 41.30% |
| 10000 \_ 120 \_ 1000 | 37.40% |
| 8000 \_ 135 \_ 1000 | 37.10% |
| 9000 \_ 120 \_ 1000 | 36.60% |
| 10000 \_ 105 \_ 4000 | 36.50% |
| 8000 \_ 150 \_ 1000 | 36.20% |
| 10000 \_ 105 \_ 1000 | 36.20% |
| 8000 \_ 120 \_ 1000 | 36.10% |
| 10000 \_ 90 \_ 1000 | 35.60% |
| 8000 \_ 105 \_ 1000 | 35.10% |
| 10000 \_ 105 \_ 3000 | 35.01% |
| 10000 \_ 75 \_ 1000 | 34.95% |
| 9000 \_ 135 \_ 1000 | 34.94% |
| 10000 \_ 105 \_ 5000 | 34.13% |
| 6000 \_ 150 \_ 1000 | 33.52% |
| 7000 \_ 120 \_ 1000 | 33.14% |
| 8000 \_ 150 \_ 5000 | 33.12% |
| 6000 \_ 135 \_ 1000 | 32.86% |
| 10000 \_ 45 \_ 1000 | 32.77% |
| 10000 \_ 90 \_ 4000 | 32.72% |
| 10000 \_ 120 \_ 4000 | 32.69% |
| 9000 \_ 150 \_ 2000 | 32.63% |
| 8000 \_ 45 \_ 1000 | 32.60% |
| 8000 \_ 105 \_ 3000 | 32.56% |
| 7000 \_ 75 \_ 5000 | 32.41% |
| 9000 \_ 90 \_ 1000 | 32.39% |
| 10000 \_ 90 \_ 3000 | 32.37% |
| 8000 \_ 120 \_ 3000 | 32.20% |
| 10000 \_ 75 \_ 4000 | 32.12% |
| 9000 \_ 75 \_ 1000 | 31.98% |
| 9000 \_ 150 \_ 1000 | 31.88% |
| 10000 \_ 135 \_ 3000 | 31.82% |
| 6000 \_ 135 \_ 2000 | 31.79% |
| 7000 \_ 105 \_ 1000 | 31.54% |
| 10000 \_ 60 \_ 1000 | 31.52% |
| 6000 \_ 120 \_ 1000 | 31.44% |
| 8000 \_ 60 \_ 1000 | 31.19% |
| 10000 \_ 60 \_ 4000 | 31.01% |
| 8000 \_ 135 \_ 5000 | 30.96% |
| 9000 \_ 120 \_ 3000 | 30.80% |
| 7000 \_ 135 \_ 5000 | 30.76% |
| 9000 \_ 105 \_ 1000 | 30.76% |
| 7000 \_ 90 \_ 1000 | 30.64% |
| 8000 \_ 105 \_ 5000 | 30.60% |
| 10000 \_ 45 \_ 4000 | 30.60% |
| 8000 \_ 75 \_ 1000 | 30.59% |
| 8000 \_ 90 \_ 1000 | 30.39% |
| 8000 \_ 120 \_ 5000 | 30.38% |
| 9000 \_ 90 \_ 5000 | 30.33% |
| 10000 \_ 75 \_ 3000 | 30.30% |
| 9000 \_ 90 \_ 3000 | 30.16% |
| 10000 \_ 150 \_ 5000 | 30.10% |
| 9000 \_ 150 \_ 5000 | 30.09% |

**EXCHANGE LOCATION TEAM:**

Trade Composition aggregated by Primary Exchange:

A

N

P

Q

Z

With the exception of the Z primary listed stock, we can observe D dominating midday and early and late period exhibits similar behavior. Second most traded exchange tends to be the primary exchange.

**Analysis of the D trade flows:**

MSFT: Size(Absolute, not scaled)

Red: Total D Trade size;

Blue: D1 Trade size(trades occur close to the mid of the previous quotes). We take the most prevailing bid and most prevailing ask to compute the mid.

YHOO

We noticed that there is spike in the D volume in the midday with the D2 traffic dominating. This shows that the majority of the D traffic are on the “outside” of the mid.

Results not included:

1. There is a more detail number on the trade flows per exchange per time frame, but it is available under the /RScript
2. There is also a heat map of the vector of stock vs exchange. It is too long to include it here (~3000 columns), but it is available under /RScript

Issues:

1. Some strange error with reading ticker “QQQ”, will look into it. The RStudio just display a picture of a bomb, claim that it encounters an fatal error with telling what error it is. It does not seem like a memory error as it only occurs at the same ticker and removing it fixes the problem
2. Speed up some of the code. It takes about 1-2 sec to read, parse, aggregate a ticker, 2700 tickers will take about 40 minutes to run. May not be a big issue if we dump the aggregate data and run from it.

Next Week:

Look into the change in spread and how they relate to the trade venue

Work with the vector data and study its clusters.

Look into the NASDAQ exchange system status log and see if there is delay data to correlate that venue. The theory is that trade traffic tends to move to venue with the least delay.