Venue liquidity and price prediction in the US equity market

Zhenyu Chen, Jia Xu, Yilun Qian, Kuangdi Zheng, and Jiongjia Fang

December, 2013

Abstract and Findings

In today's fragmented US equity market, multiple trading venues have proliferated following the Regulation Alternative Trading Systems (ATS), which allows exchanges to become for-profit entities and Regulation National Market System (NMS), which stipulates top of book order protection. Due to the proliferation of trading venues, understanding stocks' trading distribution among these venues can potentially offer valuable insights for market makers about what their competitors and other market players are doing. In this paper, we use a robust optimization model to produce a good estimate of stocks' distributions among trading venues by minimizing the misclassification rates. We also show the construction of a short-term price prediction model using tick level trades and quotation data. Finally, we demonstrate the model's predictive power.

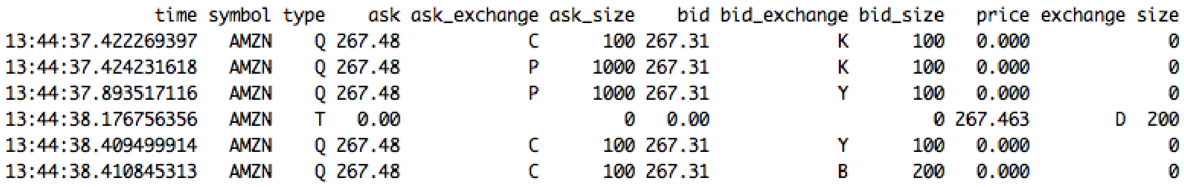
This report aims to provide market makers with an edge. Understanding stocks' trading venue distribution and short-term price momentum allow market makers to minimize their risk by optimizing their order executions.

The report is separated into six sections. The first section describes the raw market data that are used in this project. The second section is optional reading that shows the factors shaping market makers' behavior; it includes some crucial statistics for all top market makers in the US equity market report. The third section explores venue liquidity and contains detailed descriptions of the aforementioned optimization model. The fourth section shows the construction and analysis of a two-factor short-term price prediction model. The fifth section shows an application of the price prediction model. The sixth section will discuss further studies we would like to undertake in the future.

1. Market Data:

1.1 Raw Market Data Format Explanation

The raw market data were National Best Bid and Offer (NBBO) quotes and consolidated trades for US equities in April and May 2013. Several entries (after parsing) provided in the following snapshot for demonstration



```
                time symbol type    ask ask_exchange ask_size    bid bid_exchange bid_size   price exchange size
13:44:37.422269397   AMZN    Q 267.48            C      100 267.31            K      100   0.000            0
13:44:37.424231618   AMZN    Q 267.48            P     1000 267.31            K      100   0.000            0
13:44:37.893517116   AMZN    Q 267.48            P     1000 267.31            Y      100   0.000            0
13:44:38.176756356   AMZN    T   0.00                     0   0.00                     0 267.463        D  200
13:44:38.409499914   AMZN    Q 267.48            C      100 267.31            Y      100   0.000            0
13:44:38.410845313   AMZN    Q 267.48            C      100 267.31            B      200   0.000            0
```

Figure 1.1: Raw Data

While trades and quotes represent different types of information, for convenience, we serialize the data together and process them sequentially. All data are sorted according to the recorded time, and all time formats are in UTC time. The column “Type” indicates whether the entry is a trade or a quote. The key attributes for quote entries are 1) Ask, 2) Ask Size, 3) Bid, 4) Bid Size. The key attributes for trade entries are 1) Price, 2) Exchange and 3) Size.

1.2 Raw Data Clean up

Quotes and Trades obtained from the SIP (Securities Information Processor) feed originate from multiple geographically-separated sources. The sequence of messages in the raw data is as they were received by our market data provider[1]; but different market participants might receive the messages in a slightly different order. When we see a trade report mere milliseconds after a quote message, it is possible that the latest quote had not yet been observed by the decision process that caused the trade. We shall match every trade with a corresponding quote that regulated the execution price. In order to do so, we introduce a time delay for all quotes and re-sort them with all trades data.

According to Regulation National Market System (Reg. NMS), venues may not execute a trade at a price inferior to a protected quote posted on another venue. We shall find the delay that minimizes the un-weighted fraction of regular, non-trade-thru-exempt trades executed at prices inferior to the most recent quote. The example listed here is for April 23, 2013, and the same pattern is discovered for the entire month of April.

After the optimization, it shows that the optimized delay tends to behave in two different ways (the behaviors is consistent through all trading days):

Market makers and dark pools report trades using the FINRA display facility (trading venue D); and the optimized delay to match the quote is 30 milliseconds:

[1] Cantor Fitzgerald and Exegy.

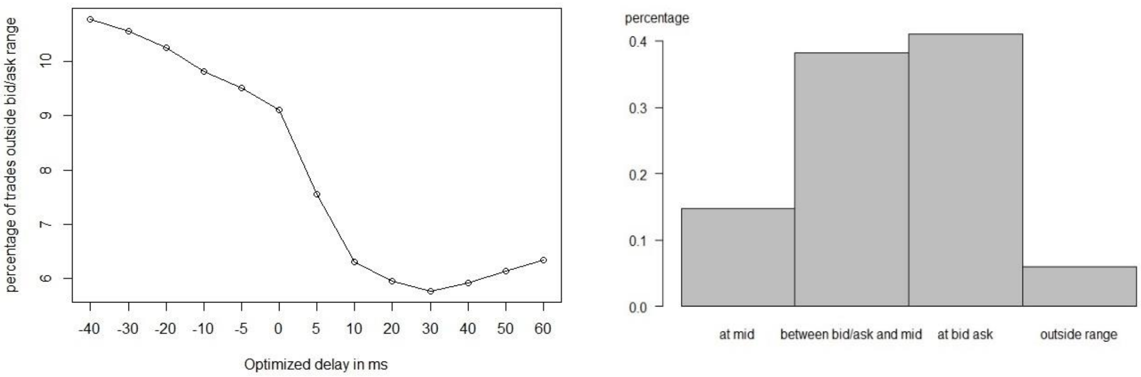


Figure 1.2 and 1.3: Optimal Delay for Trading Venue D

The distribution of trades falling within the NBBO is about 95%, therefore it is consistent with Reg. NMS. For the other major exchanges[2] (NYSE Arca, NYSE, Nasdaq, etc), the optimized delay to match the quote is exactly 0 milliseconds, and nearly 100% of trades fall within the NBBO, further validating our approach:

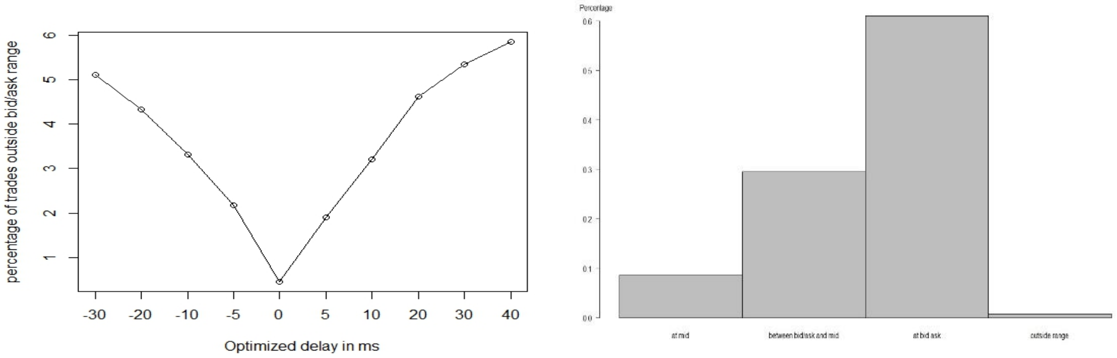


Figure 1.4 and 1.5: Optimal Delay for other major exchanges

X-Axis for Figure 1.5 (at mid/between bid & ask/at bid ask/outside range)

Y-Axis for Figure 1.5 (0.6/0.5/0.4/0.3/0.2/0.1)

1.3 Further Data Cleanup for venue liquidity estimation

After matching the quote, we have enriched the trade data with respective bid and ask price. Thus we are able to calculate the effective spread:

$$|\text{Trade Price} - \text{Midpoint of NBBO}| \times 2$$

[2] For National Stock Exchange, the optimal delay is 30ms similar to trading venue D

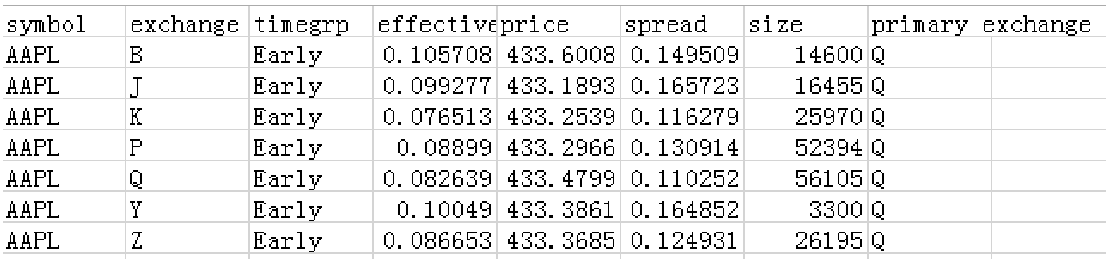
Effective spread is a good measure of exchange quality and liquidity. Lower effective spreads suggest lower trading costs; hence we want to test whether this is a significant factor in determining where the stock trade across trading venues.

Also, we noticed that traders tend to behave differently during a trading day, thus we further divided a trading day into three groups:

1. First 10 trading minutes (opening auction affects trading behavior)
2. Last 10 trading minutes (market closing affects trading behavior)
3. Rest of the day

Also we are interested in investigating market maker trading behavior. Since both market maker trades and dark pool transactions are reported in the FINRA display facility, we treat trade occurs near Midpoint of NBBO as subgroup D1 (which has 0 effective spread, more likely to be dark pool) and the rest as market maker activities D2 for a rough estimation.

The following figure is an illustration of processed data:



symbol	exchange	timegrp	effective	price	spread	size	primary exchange
AAPL	B	Early	0.105708	433.6008	0.149509	14600	Q
AAPL	J	Early	0.099277	433.1893	0.165723	16455	Q
AAPL	K	Early	0.076513	433.2539	0.116279	25970	Q
AAPL	P	Early	0.08899	433.2966	0.130914	52394	Q
AAPL	Q	Early	0.082639	433.4799	0.110252	56105	Q
AAPL	Y	Early	0.10049	433.3861	0.164852	3300	Q
AAPL	Z	Early	0.086653	433.3685	0.124931	26195	Q

Figure 1.6: Processed Data

The data is aggregated per ticker and per exchange. Effective spread, price and spread are volume weighted. These processed data are used as inputs for model development and optimization in the later part of this project.

1.4 SEC Market Center Monthly Data

Section 2 of the project also uses trading data from Market Maker Data on the FINRA Display Facility. Each Market Maker data is provided on a monthly basis per market maker, we aggregate the data by taking the top eight market makers. Part of section 2 reconciles publicly available provided by VistaOne Regulatory Services (then called Thomson Transaction Analytics) with market data provided by Exegy, and examines if the patterns uncovered are similar. The time period of this study is between April, 2013 and May, 2013.

2 Market Maker Landscape (Optional)

2.1 Looking at Market Maker Data

Before delve into the comparison of the data, we first examine the market maker to get a general picture against our intuition. We examine the distribution of Market Order Fraction, Limited Order Fraction and Other Limit Order Fraction. Since different tickers have different trading behavior with varying market order/limit order fraction, we look at them in aggregate:

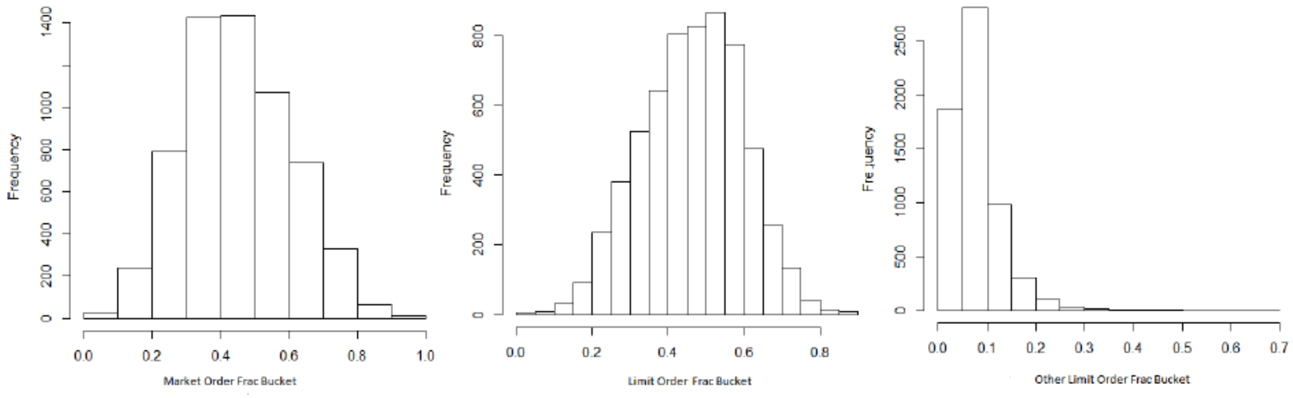


Figure 2.1, 2.2, and 2.3: Market Order Fraction/ Limited Order Fraction/ Other Limited Order Fraction

As observed in the distribution chart, Market orders consist of 44% of all executed volume, limit order consists of 48% and other limit order consists of 6%. Market orders and limit orders account for half of the majority of executed volumes as expected.

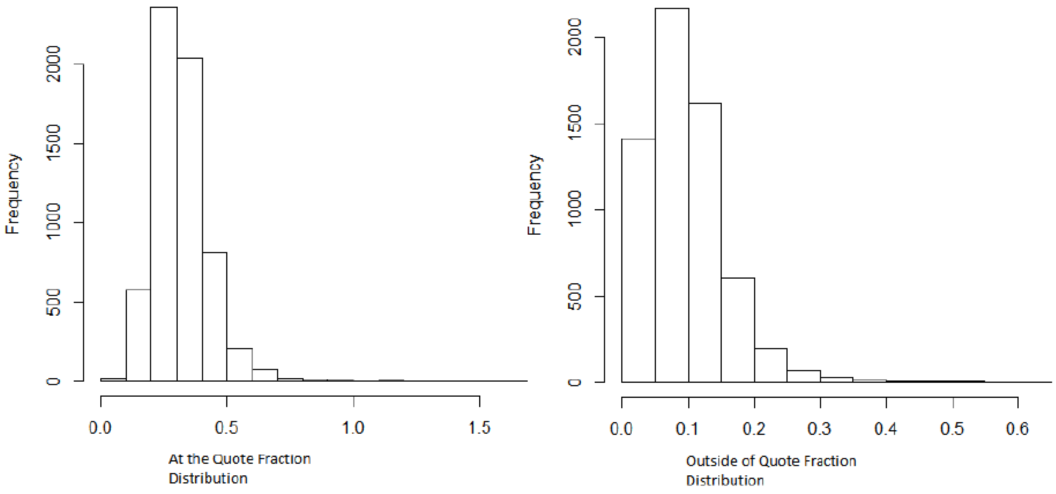


Figure 2.4 and 2.5: At the Quote / Outside of Quote Fraction

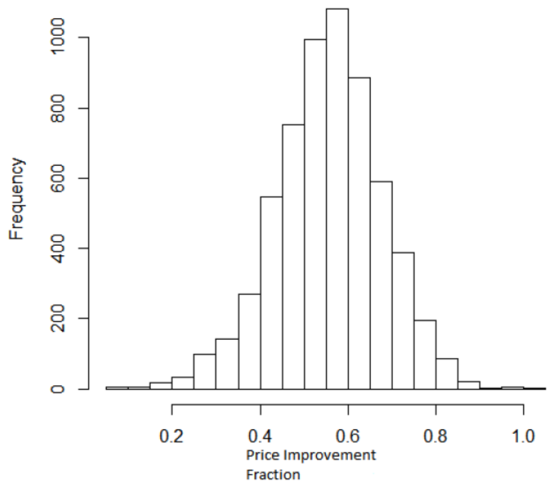


Figure 2.6: Price Improvement Fraction

Next we look at the fraction of volume that falls at the quote and outside of the quote. As observed in the figure 2.3 and 2.4, on average, the at the quote trade makes up 34% of all trade volumes, outside of the quote consists of only 8% of trade volumes. The low probability of a trade occurring outside of quote and high probability of trade occurring at the quote matches our intuition and is consistent with order protection rule stipulated by Reg NMS.

Lastly, we look at the percentage of trade volume with price improvement. As observed in figure 2.6, on average, trade improvement volume makes up roughly 55% of all trade volumes; this is evidence that market makers increase market liquidity.

2.2 Is effective spread consistent between Market Maker (VistaOne Regulatory Services data) and D Venue (Market maker & Dark pool data provided by Exegy)?

Now we have a general picture of the Market Maker Volume, we can compare the Market Maker Effective Spread vs. the D Venue Effective Spread. Effective Spread tells us the trade price point relates to the mid-price. Before we compare the effective spread, we scale the effective spread by the price. Because the volume can vary greatly in magnitude and outliers can affect the analysis significantly, we chose to compute the rank for each market maker and Venue D and make comparison on the rank. The R Square of the weighted (Effective Spread)/Price for market maker versus D Trade is 0.048. The R Square of the rank linear regression is 0.4. Given the low R Square value, we conclude that the effective spread/price of the two are not consistent.

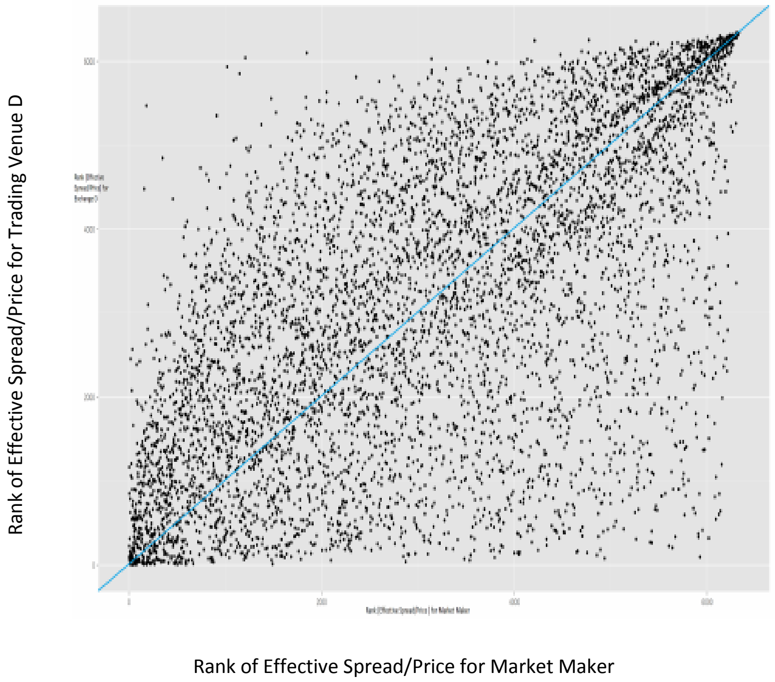


Figure 2.7: Rank of Effective Spread/Price for Market Maker versus Trading Venue D

2.2 Is Trading Volume consistent between Market Maker and Venue D?

To investigate whether the trading volumes are consistent across the market maker and venue D, we look at that from two perspectives:

1. We examine the trading volume by looking at the percentage of trade occurred relative to the quote.
2. We examine the trading volume by looking at the consistency of their trading pattern.
3. We examine the trading volume by looking the consistency of the trading price pattern.
4. We examine the relative magnitude of the trading volume.

Trade Volume relative to quote

We computed the percentage of the trades occur relative to the quote for the Venue D and compare with our result from prior section. The results are summarized in the table below. We can see that Market Maker have more trades occurring outside of quote in trade of the trades occurring at the quote. But the overall trade volume relative to quote for the two are consistent and matches our intuition.

Exchange	Market Maker	D Venue
Outside Quote	8%	3.5%
At the Quote	34%	22.5%
Between the Quote	58%	74%

Figure 2.8: Trades occur relative to the quote for the Venue D

2.3 Looking at Trade Volume Pattern

We computed the rank for Executed Volume for each ticker from the market maker data. Since we need to perform an "apple to apple" comparison, and Executed Volume is denominated on a per month basis, we aggregate the D Trade Volume for each ticker over one month. The result is shown in Figure 2.9. The R Square of the linear regression of the volumes (without rank) is 0.6. The R square of the ranks of the volumes is 0.8. Given the high R Square, we conclude that the two are correlated and they are consistent.

To delve deeper into the comparison, we color coded the tickers by their respective volumes. The tickers are places into ten different volume buckets based on their volume percentile. The lowest 10% traded stocks are in bucket 1, the second lowest 10% traded stock are in bucket 2 (Percentile 10%-20%). We observe that the least traded stock located at the upper right of the chart and the most popular tickers are on the bottom left, which matches with our intuition. If we start from the origin, and draw a 45 degree line to the upper right corner, we will have a line that divides which venue have higher relative trading activity. Tickers locate above line will have higher relative D trading activity and tickers locate below the line will have higher relative Market Maker Activity.

For the trading volume, there is no obvious distinction on how different volume skew to one side of the 45 degree line or the other. To aid our understanding on the volume skew, we sort data by the executed volume in increasing order, then computed a moving average using a window of 255 points. The moving average will help us see the volume skew when compare to the 45 degree line. The red line in Figure 6 is the result of the 255 point moving average smoothed out. We observe the shape of the moving average to be concave down, then into concave up. This tells us that at high volume, there is a skew towards the D Venue, and at low volume, there is a skew towards the market maker. One possible interpretation is that Dark Pool participant on the D Venue tends to trade more volume. And for low volume/less popular stock, there is a bias to turn to the market maker for trades.

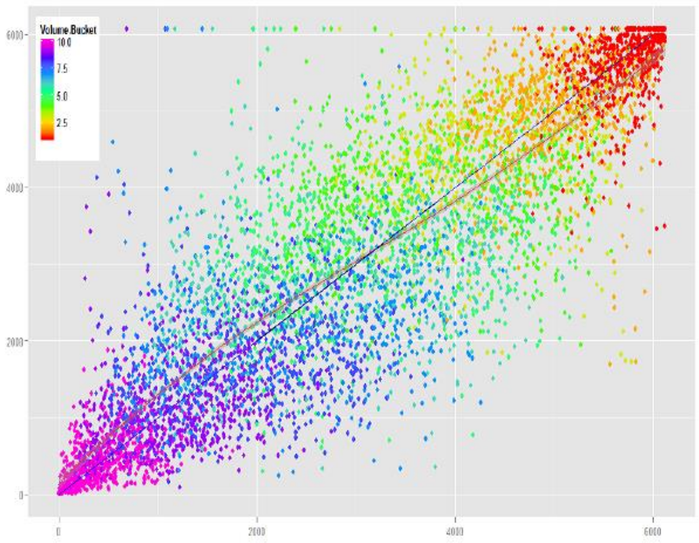


Table 2.9: Rank of Executed Volume for Market Maker vs. D Venue Trading Volume Colored by volume

Red line: the 255 point moving average, Blue line is the 45 degree line

Horizontal Axis: Rank of Market Maker Executed Volume

Vertical Axis: Rank of Trading Venue D

Looking at Trade Volume Price Pattern

Another perspective to look at the volume is by breaking it down for each price. We divide the entire stock universe by the average price over the month and place them into different buckets. Then we aggregate the percentage of trade occurring within those price buckets, we came up with the distribution as shown in Figure 2.10. Note that the price buckets are in log scale, and we see the most trades occurring between 10 and 20 USD. We conclude that the Trade Volume Price Pattern are consistent across Venue D and Market Maker.

Looking at Relative Magnitude of the Trading Volume

Though the relative trading volume behavior are similar, but we performed the analysis based on the relative size of the volume. Now we examine the absolute magnitude to wrap this section up and put things in perspective. From Figure 2.11, we observe that on average, the trading volume on the Venue D is 3X to 5X the market maker, illustrating the vast amount of volume traded on Venue D. Notce that there are some instance of the distribution in which the market maker executed volume exceeds the D Volume (To the right tail of the distribution). For those ticker, the occurrence can be explained by extreme low volume, ticker name change and/or a result of record error.

In conclusion, the overall volume of the Market Maker and Venue D are consistent. However, we still see a bias on the volume arises from the volume of the ticker.

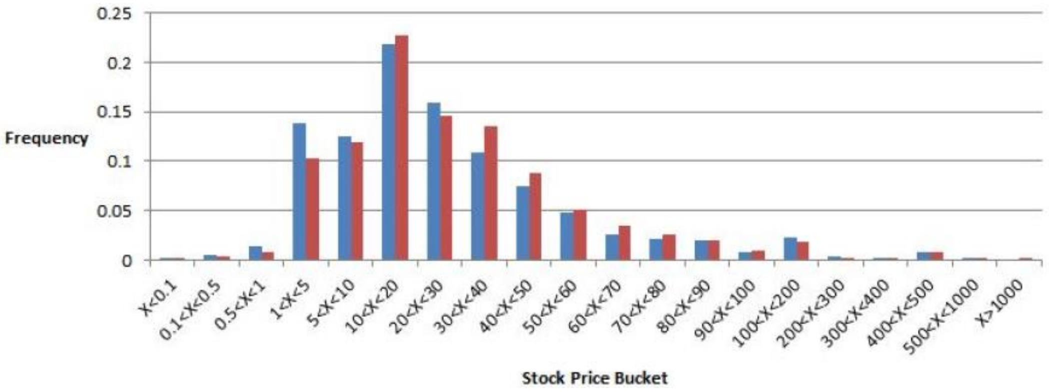


Figure 2.10: Distribution of Trade by Different Price Bucket.

Red Bar represents the Venue D, Blue Bar represents the Market Maker Executed Volume

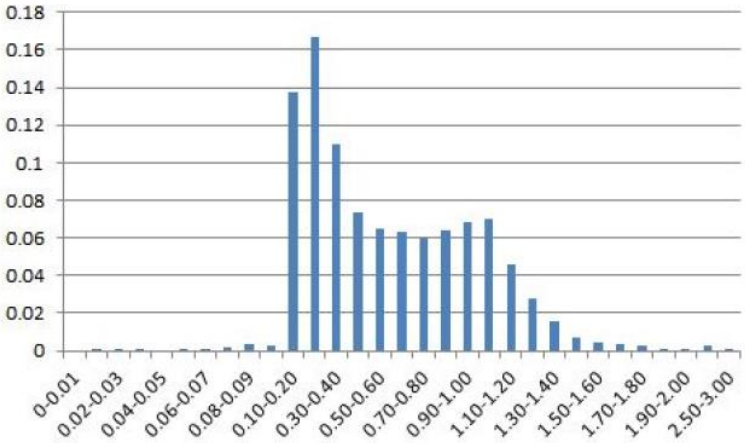


Figure 2.11: Distribution of Executed Volume/D Volume by Bucket

Horizontal Axis: Executed Volume/D Volume Bucket, Vertical Axis: Frequency

2.3 What are the movement behavior of the trading volume?

The data we observed from the prior section is based on the data from the month of April, 2013. Now we look at the movement of trade volume progresses over time. We investigate how the trading volume pattern change from April, 2013 to May, 2013. We first propose a measurement call rank distance. The definition of Rank Distance is:

$$Rank\ Difference = Rank[Exchange\ D\ Volume] - Rank[Market\ Maker\ Executed\ Volume]$$

Rank Difference tells us the relative position of the ticker to the 45 degree line. A positive number means above the 45-degree line, and a negative number means below the 45-degree line. We define outliers as the tickers with highest 10% or the lowest 10% in rank difference. Then, we track the movement of the outliers after one month, the result is shown in Figure 2.12. It is obvious that the outliers gets closer to the 45 degree line in the next month. There is a mean reversion behavior for the outliers and their position on the chart are not stable.

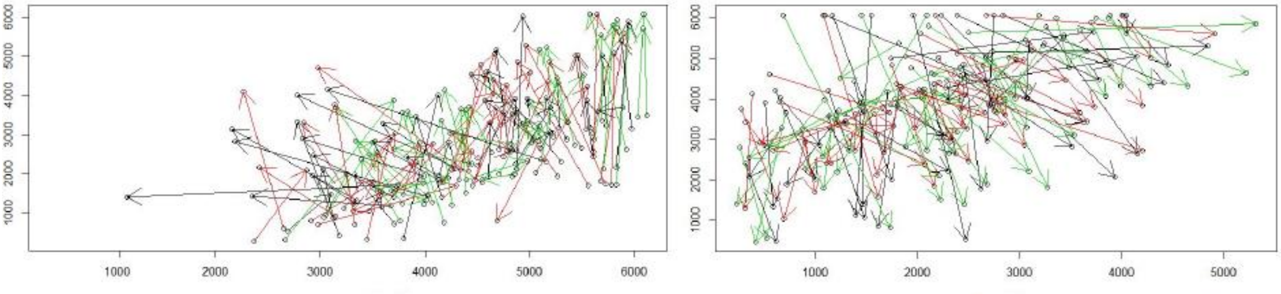
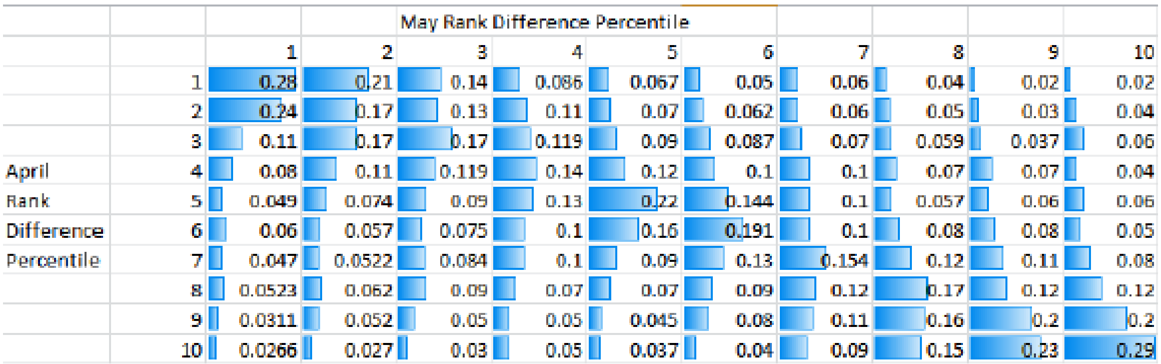


Figure 2.12: Movement of Outliers on Executed Volume Rank vs. D volume Rank from April 2013 to May 2013



		May Rank Difference Percentile									
		1	2	3	4	5	6	7	8	9	10
	1	0.28	0.21	0.14	0.086	0.067	0.05	0.06	0.04	0.02	0.02
	2	0.24	0.17	0.13	0.11	0.07	0.062	0.06	0.05	0.03	0.04
	3	0.11	0.17	0.17	0.119	0.09	0.087	0.07	0.059	0.037	0.06
April	4	0.08	0.11	0.119	0.14	0.12	0.1	0.1	0.07	0.07	0.04
Rank	5	0.049	0.074	0.09	0.13	0.22	0.144	0.1	0.057	0.06	0.06
Difference	6	0.06	0.057	0.075	0.1	0.16	0.191	0.1	0.08	0.08	0.05
Percentile	7	0.047	0.0522	0.084	0.1	0.09	0.13	0.154	0.12	0.11	0.08
	8	0.0523	0.062	0.09	0.07	0.07	0.09	0.12	0.17	0.12	0.12
	9	0.0311	0.052	0.05	0.05	0.045	0.08	0.11	0.16	0.2	0.2
	10	0.0266	0.027	0.03	0.05	0.037	0.04	0.09	0.15	0.23	0.29

Figure 2.13: Rank Difference Transition Matrix from April 2013 to May 2013 by Volume

Vertical Axis: Starting Rank Difference Percentile in April, 1 represents lowest 10% percentile, 2 represents 10% to 20%, etc.

Horizontal Axis: Ending Rank Difference Percentile in May

We further expand our analysis for tickers with other Rank Difference. We compute the conditional transition matrix. The transition matrix paints a different picture and suggests that tickers are more likely to return its original rank Difference percentile. This seems to contradict our prior observation of mean reversion. It turns out that the distribution of the rank Difference is highly concentrated between -200 and 200. As a result, even with big rank Difference movement, the outliers could still retain its original rank Difference percentile.

Factors affecting the Market Maker Volume Rank vs. Venue D Volume Rank

We conclude the section by looking into the factors affecting discrepancies between Market Maker Volume Rank and Venue D Volume Rank. We investigate different factors and observe their effect on the ticker's placement relative to the middle 45 degree line.

Price

One of the factors is Price. Since stocks have prices spanning at least three orders of magnitude, we chose to look at the logarithm of price. As we can see in figure 2.14, there is a "price effect" that biases higher price stock towards market maker (Below the 45 degree line).

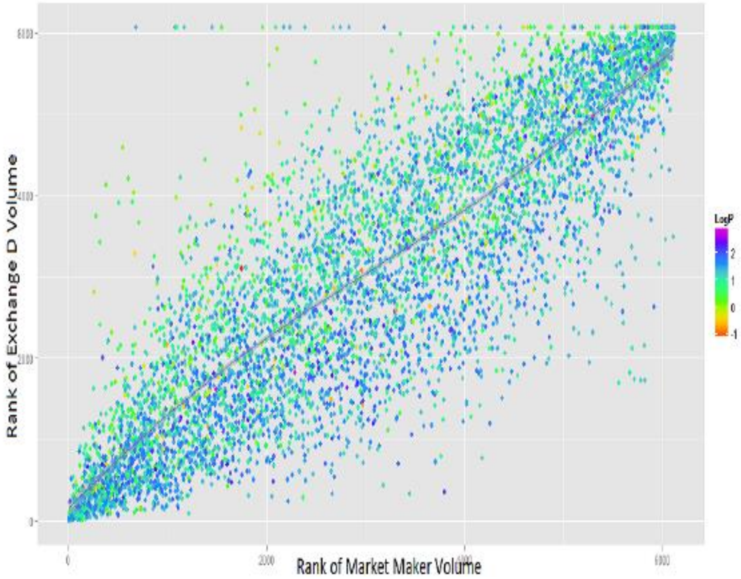


Figure 2.14: Rank of Executed Market Maker Volume vs. D Trade Volume filled by log price

Volume

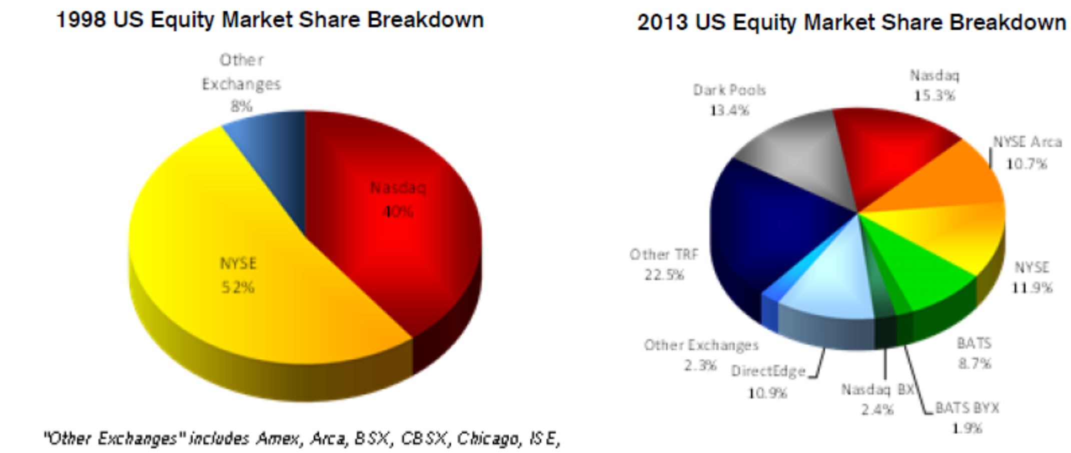
Another factor we investigated is volume. As we already noted in the prior section, there is a bias for Market Maker to trade at low volume and a bias for Venue D at high volume. At individual ticker level, there is a small skew towards Market Maker.

3. Venue Liquidity Prediction

Now readers have gained an understanding of market makers' trading characteristics, we show a prediction model on venue liquidity and explain why such a model is crucial to gaining a competitive advantage.

3.1 Introduction and Motivation

Ever since Regulation NMS, US equity market has been increasingly fragmented. A study by Credit Suisse demonstrated the following market shares among exchanges:



"Other Exchanges" includes Amex, Arca, BSX, CBSX, Chicago, ISE, National and PHLX

Source: Credit Suisse Trading Strategy

"Other TRF" includes broker capital commitments and internalizations
"Other Exchanges" includes NYSE Amex, CBSX, Chicago, National and PSX

Figure 3.1 and 3.2: US Equity Market Share Breakdown in 2009 and 2013

"Note that 10 of the exchanges are the product of just 4 exchange groups. Each exchange group has created multiple venues with different rules and cost structures to suit the needs of different investors" (Credit Suisse – The Dark Pool Debate). Due to the proliferation of exchanges, and the different maker/taker structures, market players with different motives will post their trades on different exchanges. Most exchanges offer a rebate fee for limit orders that add liquidity to the market, and charges remove liquidity fee for market and marketable limit orders; whereas few exchanges, such as Direct Edge (EDGA) and Boston Exchange (BX) offer a rebate for trades that take away liquidity, and charges a fee for trades that add liquidity.

Understanding the distribution of stocks' trading venues can potentially offer valuable insights for market makers about what their competitors and other market players are doing. In addition, knowing the distribution of equities' trading volume across trading venues can potentially elicit hidden patterns, which might have been overlooked, that have caused a stock to behave in a certain way. HFT algorithms can also incorporating stocks' trading venue distribution when deciphering trading patterns used by competitors.

On top of that, understanding market liquidity and market depth of each trading venue has a great influence on order execution quality; this study has important implications for order placement. The target of this section is to construct a robust model to predict the optimal number of shares to be placed across each trading venue with a potential input of security symbol.

Figure 3.3 is the list for current available trading venues in the US.

Symbol	**Trading Venue**	**Symbol**	**Trading Venue**

A	NYSE MKT	N	NYSE Euronext
B	NASDAQ OMX BX	P	NYSE Arca Exchange
C	National Stock Exchange	Q	NASDAQ OMX
D1/D2	FINRA Trade Reporting Facility (D1 = Midpoint trades, D2 = All other trades)	W	Chicago Board Options Exchange
J	EDGA Exchange	X	NASDAQ OMX PHLX
K	EDGX Exchange	Y	BATS Y-Exchange
M	Chicago Stock Exchange	Z	BATS Exchange

Figure 3.3: Symbols for Trading Venues[3]

3.2 Trade Distributions across Venues

For market center prediction analysis, we include all the securities listed on A, N, P, Q, and Z. Each row represents one security with a unique ticker, and each column represents a trading venue. The cell value is the trading volume of the certain security took place in the certain trading venue from April 2nd to 30th 2013, and we treat the first ten minutes (early), last ten minutes (late), and other time period (midday) separately.

We divide the trading venue into three categories: Dark Pool and Market Maker (Both needed to be reported to FINRA Trade Reporting Facility), PE (primary exchanges of a stock), all the other trading venues.

	Early		**Midday**		**Late**	
Dark Pool (D1)	11,053,568	8%	294,233,016	10%	25,106,646	9%
Market Maker (D2)	60,328,627	44%	1,242,967,045	42%	112,183,091	39%
Primary Exchange	21,584,663	16%	384,945,903	13%	55,678,520	19%
All Other Exchange	44,294,515	32%	1,002,826,249	34%	97,855,479	34%

Figure 3.4: Contingency Table for Number/Percentage of Shares

Figure 3.4 displays the distribution of shares of all the stocks in one day and *t*he percentage of number of shares took place in each trading venue for early, midday and late time period.

From *Figure 3.4,* we can see that the proportion of trades in other trading venues remains consistent. The percentage of Dark Pool trades that take place in the last ten minutes of the days are lower than that of the beginning 10 minutes and that of the rest of the days.

[3] Please refer to the Consolidated Tape System (CTS) specification and the UTDF specifications.

Generally around 50% of the trades took place in Dark Pool and 17% of them took place in primary exchanges through the whole day.

On top of that, we used ternary plots, barycentric plots on three variables which sum to a constant, to show the trading venue distributions and trading migration across venues for different times of the day. The ternary plots graphically depict the ratios of the three variables as positions in an equilateral triangle. In *Figure 3.5*, we use rainbow color to label each stock with their total trading volume. The lighter points represent stocks with smaller trading volume and the darker points represent stocks with larger trading volume. Different point sizes represent different levels of average trading prices of stocks (cex = log (price)/6). In *Figure 3.5*, we color each stock by their primary exchanges ternary plot for midday has clearer pattern. We have several observations in these ternary plots.

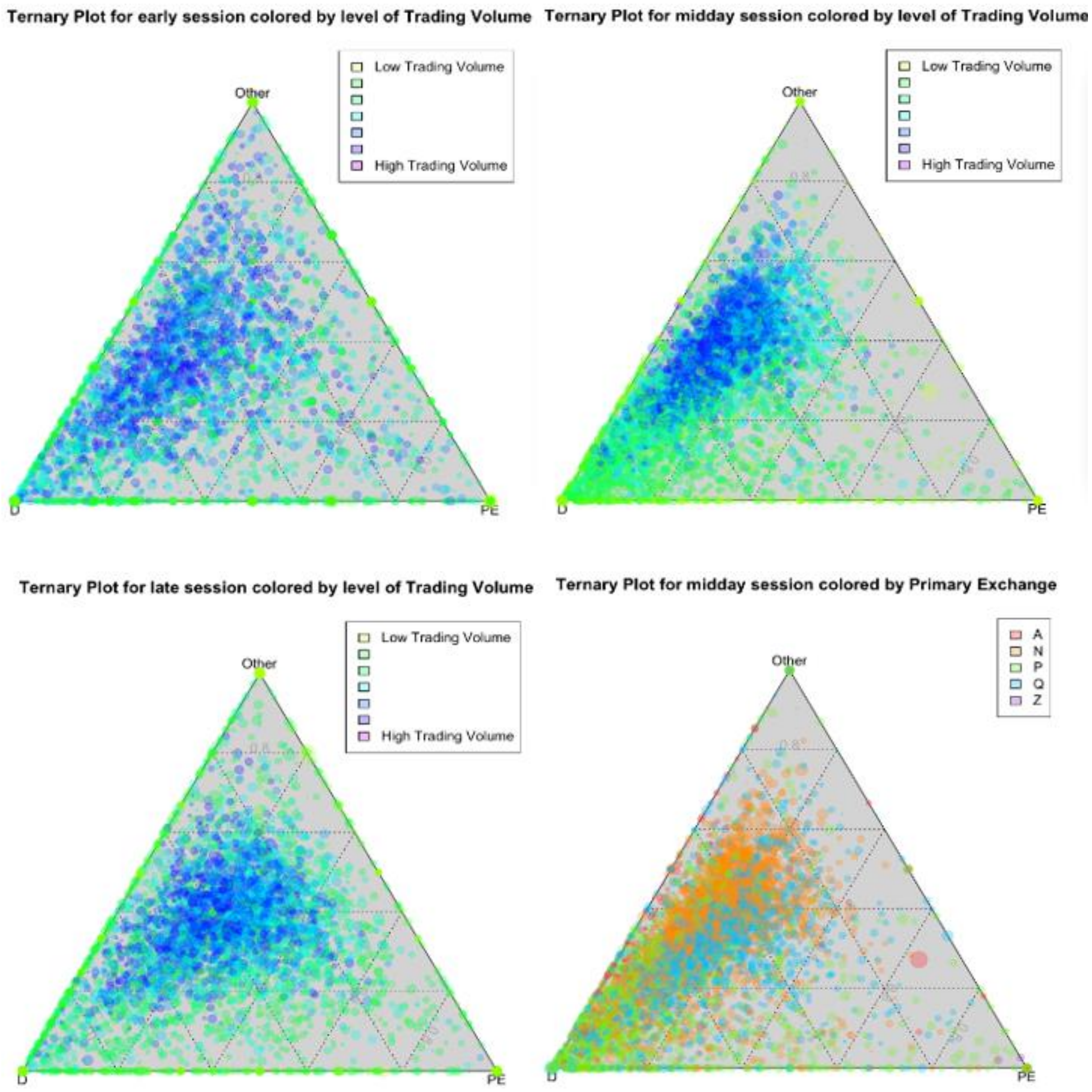


Figure 3.5 The lower left corner represents dark pool and market makers volume; the lower right corner represents volumes in the primary exchange; and the top corner represents volumes in all other venues.

To highlight some important aspects of the plots: in the first ten minutes, stocks distribute evenly among dark pool/market maker, primary exchange, and other exchanges; but in the midday, fewer stocks were traded in their primary exchanges. Late day pattern falls in between. Also in general, stocks are more likely to be traded through market makers and dark pool combined than through their respective primary exchanges.

The pattern of trading volume in the first ten minutes of the day is more random compared with midday and closing times. A possible explanation of this empirical finding can be due to the increased liquidity and market depth during the intense first ten minutes of trading, therefore, investors resort less on dark pools and market makers for transactions.

Another important finding is that stocks on the edge or the corner of the plot are always with low trading volumes. The empirical finding is consistent with common sense. Stocks with low trading volumes are stocks with less liquidity by definition, therefore, in order to minimize this liquidity shortfall risk, investors are more likely to seek market maker's help, or utilize dark liquidities in order to reduce market impact when exchanges are not as liquid (i.e. mid days).

In the last 10 minutes, most of the stocks with high trading volumes occupy the middle area of the triangle; while in the midday, the blue and purple dots are more concentrated and close to the left edge of the plot, this suggests that stocks with large trading volume are mostly traded in Dark Pools and with market makers, but not as much on the primary exchange.

The pattern for average price is similar with total trading volume but less obvious.

In *Figure 3.5*. Stocks listed in Exchange A are primarily traded in D Venue or other exchanges; stocks listed in P are more likely to be traded with market makers and dark pools, compared with stocks listed in Q. stocks listed in N are the least reliant on their primary exchange.

3.3 Separating key market venues and peripheral market venues

The proliferation of trading venues introduces problems when analyzing across venue equity distribution. The main issue we are facing is the number of dimensions modelers must deal with. In this section, we aggregate market centers with similar characteristics and thus reduce the number of dimensions we have to work with. The resulting market center groups will be used as trade distribution fractions instead of all market centers.

The method of choice is PCA (Principal Component Analysis)

	PC1	**PC2**	**PC3**	**PC4**
A	-0.016	0.036	-0.069	-0.166
B	-0.139	-0.105	0.431	-0.287
C	-0.011	0.008	0.002	-0.105
D1	-0.086	-0.085	0.126	0.238
D2	0.773	-0.064	0.021	-0.06
J	-0.111	-0.16	0.245	-0.417
K	-0.137	-0.156	0.088	0.07
M	-0.007	-0.074	-0.014	-0.038
N	-0.246	0.712	0.331	0.285
P	-0.305	0.225	-0.648	-0.414
Q	-0.295	-0.545	-0.132	0.45
W	-0.074	-0.065	0.208	0.068
X	-0.048	0.094	0.067	0.1
Y	-0.129	-0.125	0.36	-0.404
Z	-0.282	-0.193	-0.009	-0.073

Figure 3.6: First Four Principal Component Directions for Early Data

Using PCA, we are able to reduce the data set to 4 principal components. *Figure 3.6* display the first 4 principal component directions.

In total, the first 4 PCs can explain 34% of the overall variance. In the first PC, D2 is dominant and it is negatively correlated to all the other variables. In the second PC, N and Q become significant, and they are negatively correlated with each other. B, J, P and Y become significant in the third and fourth directions. Though using the first four PCs only explains 34% of the data, they are successful in reducing the dimension by 73%.

The principal component analysis result for late trading is similar to early trading. As for midday day trading, trading venue B, J and Y are no longer significant in the first four principal directions. This indicates that the exchange B, J and Y are less active in the midday trading compared with early and late periods.

In conclusion, D2 (market maker) is always the most significant trading venue; N, P and Q- the primary exchanges are also very important regarding venue liquidity; exchange B. J and Y are of interest as well – they are different from other exchanges because they have a maker/taker fee structure that rebates takers and charges a fee to providers.

3.4 Exploratory analysis for determining significant variables

Regression Analysis

Regression analysis can help us to predict one variable based on our knowledge of other variables. According to the available data, we are going to discover the linear and non-linear relationship between trading volume fractions and other explanatory variables, including trading price, trading venue, trading time, and effective spread. After determining the optimal latency for each exchange, we are going to compare the effective spread and trading volume fractions through all trading venues. For effective spread, we standardize the spreads by dividing them by average trading price. In *Figure 3.7*, scatter plots for different trading venues have different patterns. For non-D trading venues, effective spreads and trading volume fractions are negatively correlated. Narrower the effective spread is correlated with higher trading volumes. The direction of scatter plot for D2 (market maker) is in the opposite way.

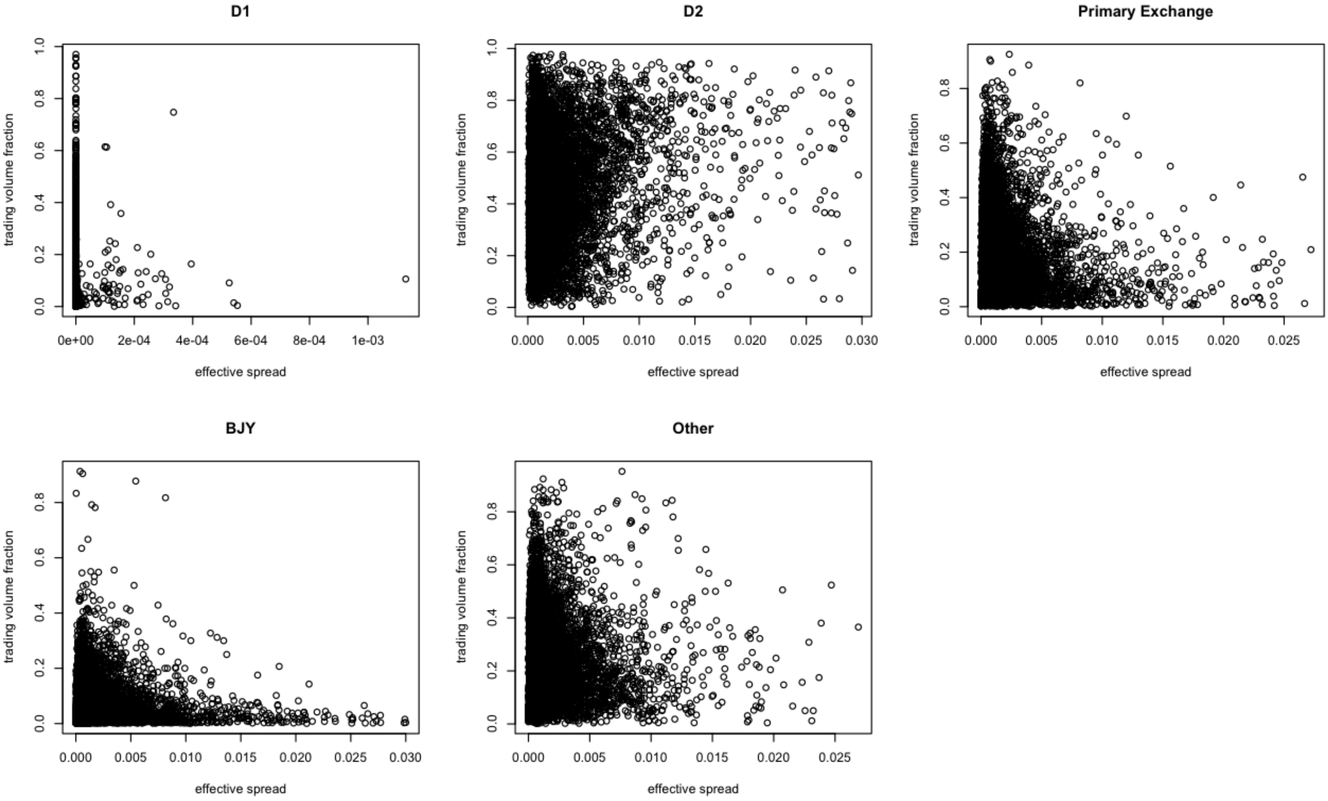


Figure 3.7

We develop the following regression model for different data set and compare the coefficient for effective spread.

$$\text{Trading volume fraction} = \beta \times\text{normalized effective spread} + \varepsilon$$

Model		β (coefficient of effective spread)	P-value
Whole data		10.59	<2e-16 ***
Split data by trading period	Early trading	12.33	<2e-16 ***
	Midday trading	9.92	<2e-16 ***
	Late trading	12.79	<2e-16 ***
Split data by trading venue	D1	9.04e+02	3.47e-10 ***
	D2	15.58	<2e-16 ***
	Primary Exchange	-8.08	5.66e-08 ***
	BJY	-2.927	1.84e-05 ***
	Other	-3.63	0.00263 **

Figure 3.8: results of trading volume fraction ~ effective spread

In *Figure 3.8*, the coefficient of effective spread for the whole data set is 10.59, which indicates the wider the effective spread, the higher the trading volume. When splitting the data by different trading venues, β become negative for primary exchange, BJY, and other exchanges. These observations are consistent with results we have found from the scatter plots. For all the non-D2 venues, an exchange with narrower effective spread means a more liquid market. It will attract more investors, and finally result in higher trading volume fraction. However, this theory doesn't stand in trading venue D2, because they are all market-maker-trades. Some other reasons may cause the popularity of D2, like desirable commission fee, etc.

Therefore, when developing the regression model, the interaction term of trading venue and effective spread cannot be neglected.

Figure 3.9 summarizes the final regression model we develop.

	Coefficients	P value

Intercept	0.08	< 2e-16 ***
Spread	-3.22	1.80e-11 ***
Trading venue: D1	0.00	0.09
Trading venue: D2	0.29	< 2e-16 ***
Trading venue: Other exchange	0.18	< 2e-16 ***
Trading venue: Primary Exchange	0.12	< 2e-16 ***
Spread×D1	64.07	0.30
Spread×D2	18.80	< 2e-16 ***
Spread×Other exchange	-2.36	0.00133 **
Spread×Primary Exchange	-4.12	9.13e-09 ***

Figure 3.9: results of trading volume fraction ~ effective spread×trading venues

In this model, D1, D2, Other exchange and Primary exchange are all indicator variables that are either 0 or 1. This model can be reorganized as:

Trading volume fraction = $\beta_0 + \beta_{D1} \times \mathbb{1}_{D1} + \beta_{D2} \times \mathbb{1}_{D2} + \beta_{Other} \times \mathbb{1}_{Other} + \beta_{PE} \times \mathbb{1}_{PE}$ + effective spread× $(\beta_{spread} + \beta_{spread\times D1} \times \mathbb{1}_{D1} + \beta_{spread\times D2} \times \mathbb{1}_{D2} + \beta_{spread\times Other} \times \mathbb{1}_{Other} + \beta_{spread\times PE} \times \mathbb{1}_{PE})$ + ε

R^2 for this model is 0.4724 and AIC is -62912.6. Trading venue, effective spread and their interaction terms are significant in the model.

One main drawback of the regression model is that the predicted trading volume fractions are not guaranteed to be in the range from 0 to 1, and sum of trading fractions through all trading venues might be far from 1. Comparatively, optimization model is more open-formed.

3.5 Predicting Stocks’ execution venue distribution - An optimization approach

In the optimization model, we will use 20 business days of market data to minimize the misclassification rate. We will have an initial guess for trading volume fractions for each group of trading venues ($G_1 = Dark\ Pool\ Volume, G_2 = Market\ Maker\ Volume, G_3 = Primary\ Exchange, G_4 = Exchanges\ B, J, and\ Y, G_5 = All\ other\ exchanges$), notice these are the groups produced via the principal component analysis discussed in section 2.3. The define error contributed by each symbol to be:

$$\text{Error}_{\text{i}} = \sum_{market\ center} \left\| F_{ij} - G_{ij} \right\|$$

Then we compute the weighted error by taking into account the trading volumes, and the final misclassification rate is defined as:

$$\text{Misclassification rate} = \frac{\sum_{days} \sum_{symbols} error_i \times Volume_i}{2 \times \sum_{days} \sum_{symbols} \times Volume_i}$$

And the goal of our optimization algorithm is to minimize the misclassification rate by producing the optimal Market Center Distribution Vector (MCDV) ($H_1\ to\ H_5$).

Based on the optimal MCDV, we add the effect of normalized effective spread of each stock/trading venue. For each symbol i, the normalized effective spread for D1 is 0, so we compute the weighted mean for the remaining four trading venues j.

$$mean_i = \frac{\sum_{trading\ venues} normalized\ effective\ spread\ \times\ trading\ volume}{\sum_{trading\ venues}\ trading\ volume}$$

$$adjusted\ fraction_j = MCDV + n\ \times\ (mean_i -\ effective\ spread_{ij})$$

Then we can compute the adjusted misclassification rate based on adjusted trading volume fractions. We optimize the adjusted trading volume fractions by optimizing the value of n that minimizes the misclassification rate.

		A	**N**	**P**	**Q**
Early	Rate without spread	21.77%	19.85%	28.05%	20.55%
	Rate with spread	21.07%	19.79%	27.83%	20.27%
Midday	Rate without spread	15.99%	13.44%	20.01%	14.18%
	Rate with spread	15.9%	13.14%	20.01%	13.79%
Late	Rate without spread	16.05%	14.00%	20.02%	15.36%
	Rate with spread	15.95%	13.67%	20.02%	14.87%

Table 2: misclassification rate using 21st day as test data

The baseline rates based on optimization model can better improve the misclassification rate than any other empirical method. For example, using 20 days' data and computing the volume-weighted mean of trading fractions as baseline, the 21st day's misclassification rate is around 2-5% higher than that of optimization model. The results are summarized in Table 2. We can conclude that optimization model can better improve the error rate we define.

That's the merit of optimization model: we can define the error rate, and the model will return the optimal baseline to minimize the error rate.

Since using one training may lead to over fitting. Cross-validation is a simple and intuitive way to estimate the expected prediction error. K-fold cross validation considers training on all but the kth part, and then validating on the kth part, iterating over k = 1,...,K.

- Divide the set {1…n} into K subsets (i.e., folds) of roughly equal size, G1,...,GK.
- Consider all other data set as training data set, and kth fold as test data set.
- For each test data, compute the estimate on the training data set,   and record the cumulative prediction error on the validation data set.
- See whether the misclassification rate is stable through one month's data

Generally, the misclassification rate remains stable through the whole month. Take stocks listed on exchange A as example, the mean misclassification rates of early, midday, and late after cross validation are 22.91%, 16.19%, and 18.64%. The cross-validated results are similar to one day's misclassification rate.

		SP100	**SP500**
Early	Rate without spread	13.56%	16.33%
	Rate with spread	13.40%	16.28%
Midday	Rate without spread	10.41%	11.90%
	Rate with spread	9.8%	11.76%
Late	Rate without spread	13.57%	12.82%
	Rate with spread	12.78%	12.57%

Figure 3.10: mean misclassification rate using cross validation for SP100 & SP500

The misclassification rate for overall analysis is higher than that of S&P100 and S&P500 stocks. This is because a large portion of stocks only traded in only 1 or 2 trading venues, which will definitely increase the error.

In all the cases above, the optimal values of n are negative, which means the wider the normalized effective spread, the higher trading volume will be. In regression analysis, we have got the similar result, and it is mainly due to effective spread's different effect on D2 and non-D2 venues. In the future, we can develop new optimization models, and treat D2 and non-D2 venues separately.

4. Price Prediction:

In this section, we construct a two factors regression prediction model that aims to improve one period ahead price prediction using order imbalance information from the current time period. We first introduce the notions of signed order imbalance, signed quote imbalance, and the intuition behind their predictive powers. We will also show analysis of the model performance and introduce some sample applications using this type of model.

4.1 Definition of Signed Order Imbalance (SOI) and Signed Quote Imbalance (SQI):

$$SOI = \sum_{i=1}^{T} w(i) * b(i)\,, where\ b(i) \in \{0,1,-1\}, w(i) = volume(i)/\sum_{k=1}^{T} volume(k)$$

$$SQI = (Size_{bid,t} - \ Size_{ask,t})/\ (Size_{bid,t} + \ Size_{size,t}), t = \sup\{T > t > t_1\}$$

$$b(i) = \begin{cases} +1, & classfied\ as\ buy \\ 0, & not\ classified \\ -1, & classified\ as\ sell \end{cases}$$

SOI essentially measures the imbalance of trades within a pre-specified time interval. In the above formula, $\sum_{k=1}^{t} volume(k)$ defines the total volume of all trades occurred within the time interval, and w(i) represents the weight of each individual trade. The pre-specified time interval needs to be fixed throughout the day for the measure to yield consistent and meaningful results. volume(i) defines the total volume in a pre-defined time interval. b(i) is a discrete function only takes value {+1,-1,0}, which is used to classify the trade as either a buy, a sell, or neutral in case a trade fails to be accurately classified.

SQI measures the imbalance of bid and ask orders among a pre-specified time interval. Since SQI always reflects the current imbalance without needing to look back into the history, only the latest available SQI measure prior to the prediction window is needed.

To illustrate the intuition behind this strategy, imagine there is a huge imbalance between bid and ask quotes size, and there are much more bid quotes than order quotes. This is an indication there are more buyers in the market than sellers, suggesting the price is likely to go up.

4.2 Trade Classification Methodology

The simplest trade classification method is known as the tick test. In the tick test, transactional level prices for trades are used. A trade is classified as a buy trade if the transactional price at time t is higher than the price at t-1, and classified as a sell trade if the transactional price at time t is lower than the price at t-1. In case both transactional prices are identical, we continue the process until a trade becomes classified. The rationale behind this is in case of the tie, we implicitly assume the momentum will continue. The shortcoming of this approach is it ignores the quote information presented in the market data, and fails to associate a trade with its corresponding quote.

Alternatively, we can look at the latest available quote prior to the transaction. If the transacted price is closer to the bid of the quote, we classify the trade as a sell by setting b to -1 (i.e. we are taking the bid). Alternatively if the transacted price is closer to the offer of the quote, the trade is classified as a buy (or b = 1). If the transacted price is equidistant, we assigned the trade as neutral. The shortcoming of this approach is that it will result in significant non-classified trades if a large portion of the trades occurs at the mid quote point.

To remedy this problem, our algorithm employs the Lee-Ready algorithm, first introduced by Lee and Ready in their seminal 1991 paper, in which they outlined an algorithm that can be applied when both quotations and trade transactions are available. This algorithm is essentially a combination of the quote test and tick test we outlined in the prior paragraph, but it has the merits of both. Essentially, the algorithm first uses the latest quote to distinguish a trade, in case the trade price falls at the mid of the quotation, instead of not classifying the trade, it then looks back at prior transactions to determine trade classification. Lee-Ready algorithm has the advantage of incorporating the latest quotes and achieve a non-classification rate near zero percent.

In the table provided below, we show our own empirical calibration of the non-classification percentage of using the quote test for the S&P 100 stocks, and the agreement rate between the simple tick test and the Lee-Ready algorithm, which is used for all subsequent analysis.

	Average Agreement Rate		**Average % of Trades Not Classified**
S&P 100	**Tick Test & Quotes**	**Tick Test & Lee / Ready**	**Quotes Test**
Average	62.12%	**72.55%**	**10.43%**
Std.	6.78%	6.12%	3.16%
Max	75.23%	87.06%	**18.96%**
Min	46.91%	57.71%	1.38%

Figure 4.1: Average agreement rate on a per stock basis, notice in the worst case scenario, employing quotes test alone can result in an 18.96% non-classified ratio.

4.3a Price Prediction Model Constructions

For ultra-short time interval price predictions, simple models are better than complex ones. Simple models not only require less computational times, but also produce more straight forward signals. One would not be required to spend times to comprehend all the fitted parameters, but instead act on the signals immediately upon appearance. Therefore, we choose to fit a simple two factors regression model for one period ahead price prediction:

$$\Delta_{price}(k, k+1) = \alpha + \beta_{soi} \cdot SOI(k) + \beta_{sqi} \cdot SQI(k) + \varepsilon(k),$$

$$where\ \Delta_{price}(k, k+1) = (Price_{k+1} - Price_k)/Price_k$$

When constructing a regression model, two approaches are common. One common is to construct a long time series of each individual equity of interest, and calibrate N sets of coefficients associated with these equities, and the other is to use a cross-sectional snapshot of the stock population as data points. While the first approach is desirable for some names of special importance, calibrating N sets of coefficients on a daily basis is very computationally demanding, and is unnecessary unless we know there are stock specific elements that cannot modeled but needed to be included. Yet should such factors be identifiable, cross-sectional regression model can also incorporate this into the analysis (for example equity price, or equity volume).

The second method, a cross-sectional regression model, provides a more general, more powerful, and more succinct solution. It is good at capturing, in our case, momentum specific information, that are applicable to all the populations. The stock population we used here is the S&P 100.

Since we are using cross-sectional regression model, it is necessary to adjust the returns by the stock's volatility (defined as the standard deviation of the stock's returns over one month period).

$$\Delta_{price,adj}(k, k+1) = \alpha + \beta_{soi} \cdot SOI(k) + \beta_{sqi} \cdot SQI(k) + \varepsilon(k),$$

$$where\ \Delta_{price}, adj(k, k+1) = \frac{Price_{k+1} - Price_k}{Price_k} \times \sigma_{one\ month}$$

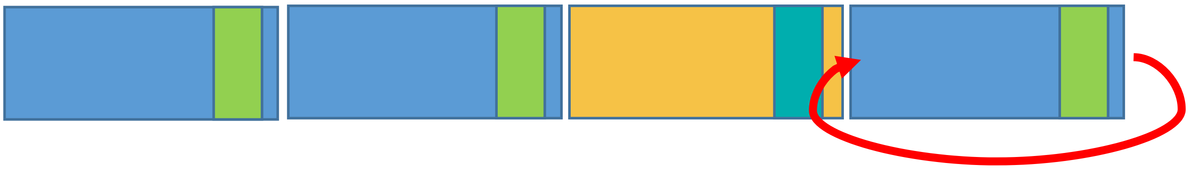


Figure 4.2

The diagram gives a better illustration. Each large rectangle represents time intervals used by SOI calculation, each smaller rectangular box represents the latest quote used for calculation. When predicting the return (represented by the red arrow), the third box contains all the SOI and SQI information we use.

The coefficients for the model calibrated for 4/23/2013 are as follow:

Time (seconds)	Quote Coefficient	t-stat	Order Coefficient	t-stat	Intercept	t-stat	Adjusted R2
t=0.01	4.36E-01	216.54	6.74E-02	62.33	2.63E-03	2.78	7.37%
t=0.5	4.03E-01	122.37	3.28E-02	18.67	6.19E-03	4.04	3.11%
t=2	4.06E-01	91.51	3.79E-02	15.95	8.68E-03	4.29	2.38%
t=5	4.40E-01	68.79	3.05E-02	8.75	1.30E-02	0.00	1.86%
t=10	4.60E-01	50.51	1.00E-02	2.68	2.00E-02	5.28	1.44%
t=60	5.10E-01	16.52	-7.00E-02	-2.61	1.00E-01	6.86	0.67%
t=120	5.20E-01	7.94	-2.20E-01	-3.44	2.40E-01	8.23	0.32%
t=180	5.50E-01	5.70	-5.40E-01	-5.11	3.20E-01	7.51	0.37%

Figure 4.3: Coefficients for regression against adjusted return

Time (seconds)	Quote Coefficient	t-stat	Order Coefficient	t-stat	Intercept	t-stat	Adjusted R2
t=0.01	6.53E-05	208.35	1.12E-05	66.63	4.16E-07	2.82	7.06%
t=0.5	6.08E-05	119.36	6.11E-06	22.52	9.28E-07	3.92	3.04%
t=2	6.18E-05	89.13	7.06E-06	19.00	1.33E-06	4.20	2.32%
t=5	6.72E-05	67.60	6.22E-06	11.48	2.01E-06	4.51	1.85%
t=10	7.04E-05	49.48	4.10E-06	5.08	3.29E-06	5.22	1.42%
t=60	8.17E-05	16.38	-8.47E-06	-1.91	1.64E-05	6.70	0.65%
t=120	8.77E-05	8.37	-3.40E-05	-3.32	3.76E-05	8.15	0.35%
t=180	9.46E-05	6.05	-8.21E-05	-4.79	5.19E-05	7.48	0.38%

Figure 4.4: Coefficients for regression against raw return

In the above charts, we vary the time interval of interest, and observe the changes in 1) strength of the predictors 2) significance of the predictors, and 3) Adjusted R^2 explained by the predictors. The patterns uncovered by the regression model are very clear. We can conclude that both SOI and SQI show significance in its predictive power of forward price change, and as the time interval for prediction becomes greater, the predictive power dissipates. The further corroborate to the validity of our model, the coefficients of both predictors are fairly consistent across time. It is also apparent that adjusting the return by the standard deviation does not vastly change the final results, but certainly scale up the coefficients.

4.3b Intuition behind the Model's Predictive Power - Concurrent Regression

While the predictive power of quote imbalance is well understood and accepted in the industry, the predictive power order imbalance is more debatable, especially after incorporating the difficulty of trade classification. To provide a general intuition on the predictive ability of SOI, we show the scatter plot of the same period returns (after trimming the returns that are larger than two mean absolute deviations) vs. SOI. We obtain an R^2 of 22.07%. The plot is as follows:

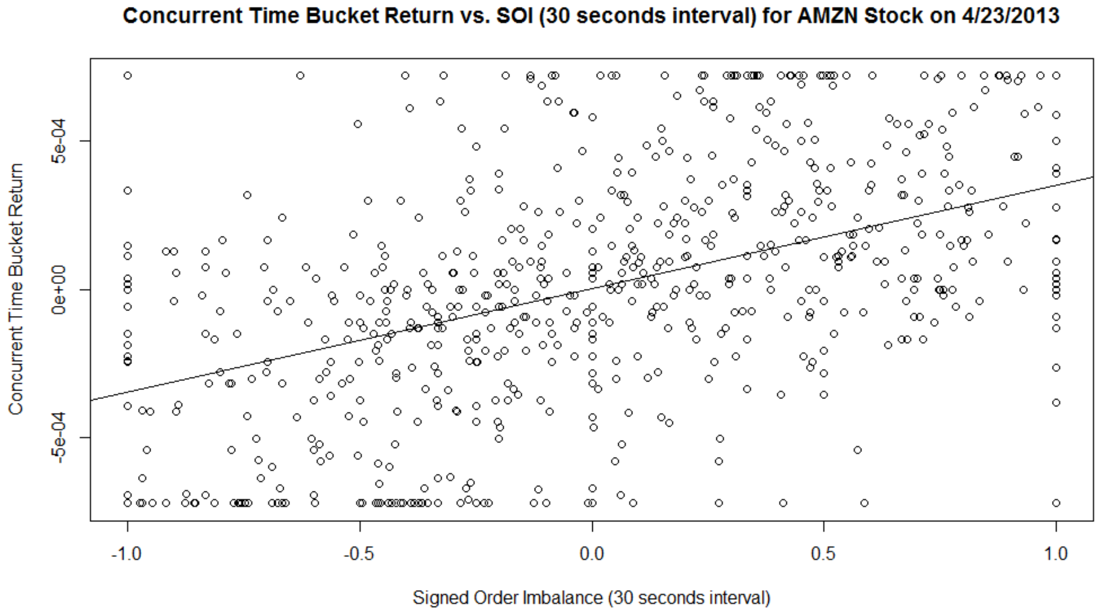


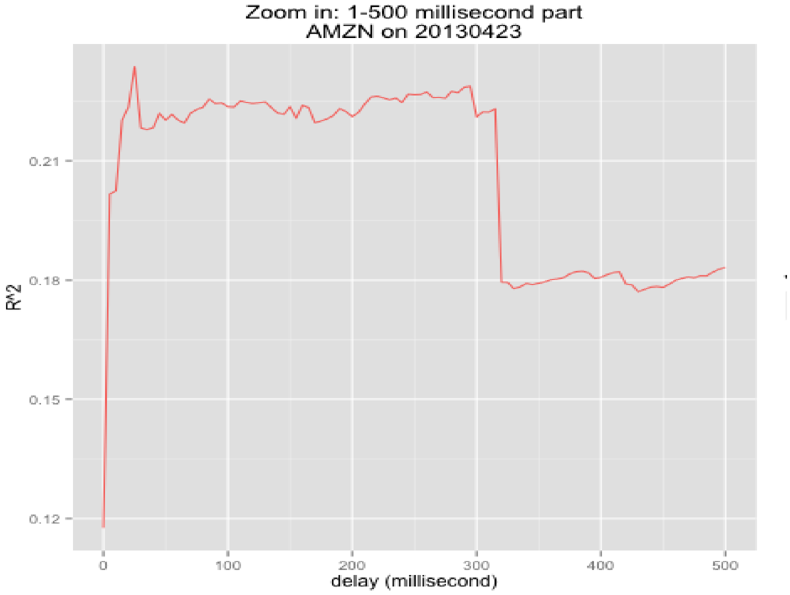
Figure 4.5: Concurrent Return vs. Signed Order Imbalance (t = 30 seconds)

	Coefficients	t-stat
Intercept	3.92E-06	0.288
SOI	3.48E-04	13.881

Figure 4.6: Concurrent SOI model parameters

The above plot shows that SOI is highly correlated with the same period return, and in section 4.3a, we had demonstrated that part of this correlation effect carries over to the next time period.

Let us also sidetrack for a minute. Recall in section one of the report, we concluded delaying all quotes by 30 milliseconds for venues D and C led to the largest fraction of trades falling in between the NBBO. By using the very raw data, and introduce various time delays for data used in concurrent regression for individual equity, we observe that delaying all quotes for 30 milliseconds has led to the highest R^2 in concurrent regression, thus corroborating to our earlier findings in section one. The R^2 plot for Amazon (AMZN) is shown below:



4.4 Price Prediction Model Analysis:

To better understand the effect of varying time window, we group all the SOI and SQI data into 20 separate bins depending on the where the SOI data point fall under. For example, all SOI data points that have values between -1.0 to -0.9 are grouped into a single bin. We then plots the average return values of all the bins against the average values of falling under each bin. The rationale behind this is to better visualize the effects and stabilities of SOI and SQI:

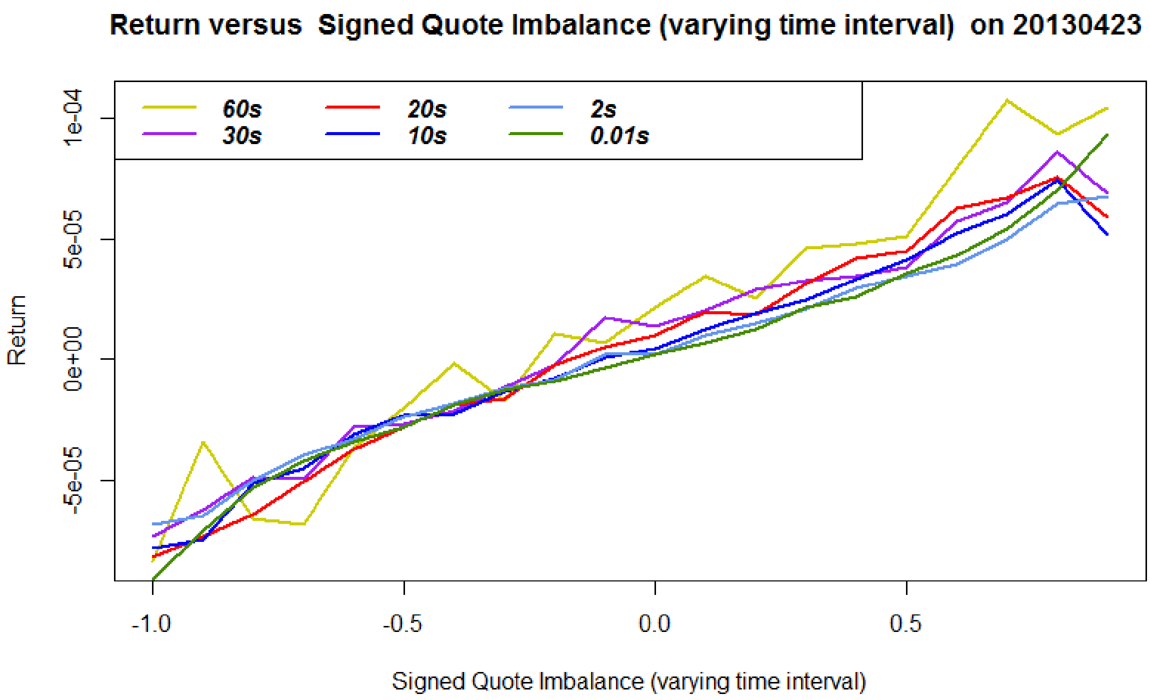


Figure 4.7: Visualization Return vs. Signed Quote Imbalance

According to the SQI, it is clear that the behind equations hold:

$$Given\ SQI(s) > 0, E[Return(s,t)] > E\left[Return\left(s,\ t-i\right)\right],\ i\ \in (0, t-s\big)$$

$$Given\ SQI\ (s) < 0, E[Return(s,t)] < E\left[Return\left(s,\ t-i\right)\right],\ i\ \in (0, t-s\big)$$

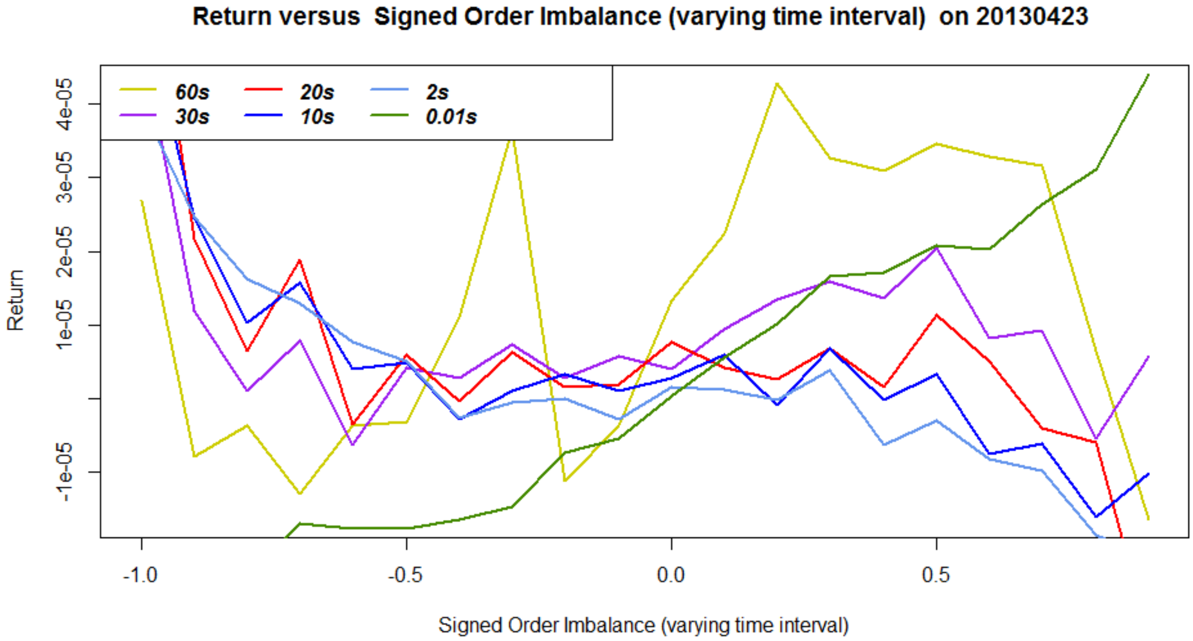


Figure 4.8: Return vs. Signed Order Imbalance Visualization

It is much harder to visualize SOI's effects, we separate the time intervals into the short end and the long end in order to better understand its effect.

Short End (t < 30 seconds):

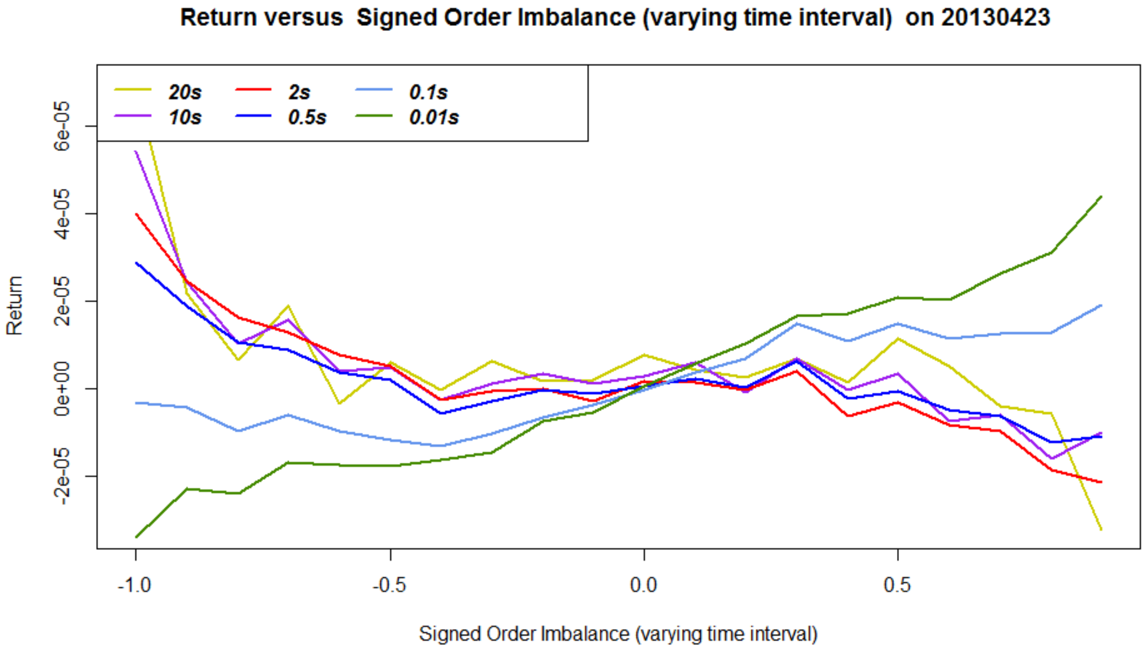


Figure 4.9: Return vs. Signed Order Imbalance Visualization (t<30 seconds)

At the short end, we believe the following set of equations always hold:

$$Given\ SQI(s) > 0, E[Return(s,t)] > E\left[Return\left(s,\ t-i\right)\right],\ i\ \in (0, t-s)$$

$$Given\ SQI\ (s) < 0, E[Return(s,t)] < E\left[Return\left(s,\ t-i\right)\right],\ i\ \in (0, X)$$

$$Given\ SQI\ (s) < 0, E[Return(s,t)] > E\left[Return\left(s,\ t-i\right)\right],\ i\ \in (\mathrm{X}, t-s)$$

X defines the turning point of the SOI coefficient. For the above plot, X ≈ t−1 seconds.

Long End (t > 60 seconds):

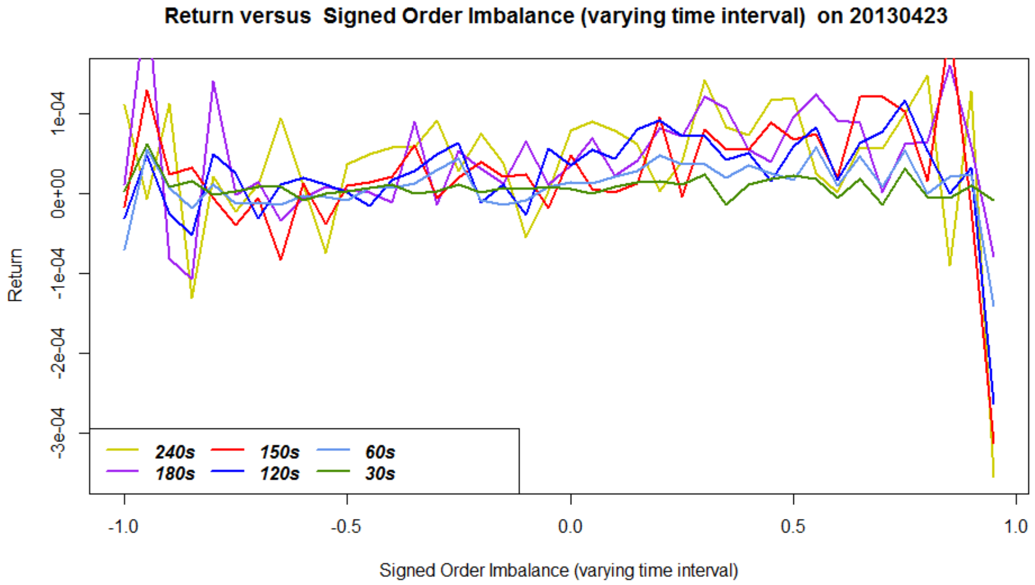


Figure 4.10: Return vs. Signed Order Imbalance Visualization (t>60 seconds)

For the long end, we further divide the bin size by half to show the increasing noise (although fitting higher degree polynomial by widening the bin size is a plausible approach, without a theoretical backup, we decide to not include that in the report). And from the comparison between SQI and SOI, the significance of the SOI coefficient decrease much faster than that of SQI; this finding is evident by the random looking shape for t = 60 seconds.

The above analysis concludes our studies of SOI and SQI as effective predictors, in the next section, we show some sample applications, which utilize these concepts.

5. Sample Applications using SQI and SOI based approaches (Optional)

To show potential applications of SQI and SOI, we hereby list two simple applications that utilize SQI and SOI measures. The first application is a direct implementation of the two factors model we just described, whereas the second application only uses the SQI as a signal generator. This section aims to inspire readers about the potential applications of the price prediction models.

5.1 Short Term Price Movement Prediction:

In this section, we use the coefficients outputted by the two factor model mentioned in the previous section, and we use the next business date (4/24/2013) as our out of sample data. The aim is test the effectiveness of price prediction by projecting the outcome of next period return as either positive or negative, and test the accuracies against 50/50 blind guess benchmark.

If the price change outputted by the two-factor model is negative, we predict downward movement, if the price change outputted by the two factor model is positive, we predict upward price movement. We exclude all trades that do not move from our samples when calculating prediction accuracies.

We did the tests for time intervals from t=5 seconds to t=60 seconds, the prediction accuracy is consistent, and hover around 58% - 61%. A consistent 10% improvement in next period price movement direction can provide a substantial edge.

5.2 Optimal Liquidation and Purchasing of shares:

In this section, we demonstrate how using the SQI to modify a naïve time-weighted average-price (TWAP) algorithm could improve the prices realized in trading.

We will also test the quote imbalance algorithm against one day of out of sample data, and evaluate the effectiveness of this algorithm against a benchmark TWAP that liquidates or purchases shares every ~7 minutes. The TWAP algorithm formula is illustrated below:

$$TWAP = \frac{1}{n\sum_{i=1}^{n} Price_{t(i)}}, \;\; t(i) - t(i-1) = c = 7min. \forall i \, \epsilon \, [2, n]$$

We assume the liquidation/purchase happens at the end of each time period of the TWAP benchmark, and we assume TWAP always uses market order for execution; this results in slightly different TWAP prices between purchase and liquidation. Our strategy peeks at the prevailing NBBO, and determines the SQI. Depending on the goal we are trying to achieve (either buying a large number of shares, or selling a large number of shares), we react differently to the SQI. Below, we distinguish between the two situations and outline the actions we take. At the end of each time period, a normal TWAP algorithm would either liquidate or purchase a pre-specified number of shares, in order to beat the TWAP benchmark and minimize risk, the algorithm also tries to liquidate or purchase a pre-specified number of shares, but only under a very strict condition specified by the SQI observed.

If the algorithm fails to liquidate a pre-specified number of shares during the time interval due to SQI signaling imminent quote change, the algorithm places a limit order that is $0.05 higher (lower) than the bid (ask) if it we intend to sell (buy) in anticipation of the imminent quote movement. If the limit order placed does not get executed, the algorithm is behind TWAP in schedule, and the shares not liquidated are put into a queue. Whenever the SQI signals favorable condition, these behind schedule shares will be liquidated right away using market orders.

What determines “favorable” versus “unfavorable” market conditions?

If the goal is to sell:

If SQI is high (i.e. SQI = 0.4 to 1), this is an indication of great buying interest. In this scenario, the dominant buying interest is likely to push the NBBO price higher. If we are at the end of the time interval, and the SQI is high, we put on a limit order, as suppose to a market order in case of TWAP.

Since we are constantly monitoring the best NBBO, once the SQI drops (i.e. SQI = -1 to -0.6), this is an indicator of predominant selling interest, in this scenario, we should liquidate our shares immediately as the predominant selling interest is likely to push the ask price lower. We do so by placing a market order seeking immediate liquidity.

If the goal is to buy:

We follow the same algorithm. When SQI is low (i.e. SQI = -1 to -0.6), it would be a signal to postpone buying orders, and when SQI is high (SQI = 0.4 to 1), it would be a signal to take the offer.

We use FFIV as a good stock candidate. It currently has a market beta of 1.62, and beginning and ending trade price are relatively at par with each other, but there is still sufficient intraday volatility for the date we are testing (4/24/2013). We show the dollar saving of the quotes imbalance algorithm over naive TWAP.

Liquidation Scenario	
Initial Shares	100,000
Shares/Trade	1,786
Ending Cash	$7,202,002.63
% Limit Order	33.93%
Compare with the TWAP benchmark	
Avg. TWAP Price (net all cost)	$71.92
Avg. SQI Price (net all cost)	$72.02
TWAP Price Std.	0.47
SQI Price Std.	0.31
Total Saving Over TWAP	$9,990.62

Purchasing Scenario	
Shares to buy	100,000
Shares/Trade	1,786
Cash Spent	$7,212,377.00
% Limit Order	46.43%
Compare with the TWAP benchmark	
Avg. TWAP Price(net all cost)	$72.24
Avg. SQI Price (net all cost)	$72.12
TWAP Price Std.	0.43
SQI Price Std.	0.43
Total Saving Over TWAP	$11,723.90

Figure 5.1

As shown in this simulation, incorporating quote imbalance information can add great value to an algorithm trading strategy that seeks to reduce market impact. The cumulative cost saving can be significant over a long time horizon.

The outperformance comes from three sources 1) Delayed execution when encountered adverse SQI signal until SQI signal is within tolerable range 2) Accelerated execution when SQI signals the trading condition will become unfavorable 3) Rebate gained from using Limit Orders when adverse SQI signals imminent NBBO change. Simple rebate assumption of $0.003/share is assumed for adding liquidity.

The Plots of $Price_{QI(n)}$ and $Price_{TWAP(n)}$ are shown below:

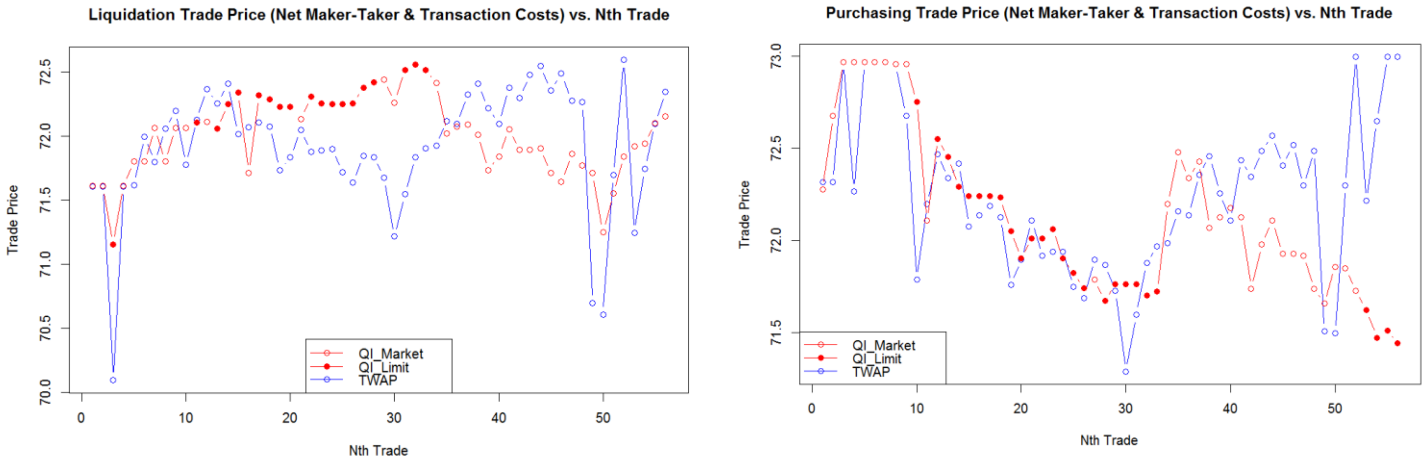


Figure 5.2 and 5.3: Liquidation/Purchasing Trade Price

In this simple demonstration, we have successfully beat the TWAP benchmark.

6. Future Studies

In the future studies, we want to improve our optimization by applying the model on all trades throughout the day, and adding a separate time variable; this way, we can better understand the time of the day effect (early, mid, late) without having to run the model three times. The impact of effective spreads are different for market maker and non-market maker, we may add a separate variable to distinguish this effect further. Our expectation is this approach will result in lower misclassification rates than the model that is currently used.

A different approach would be to construct a Naive Bayes type of model for venue liquidity prediction. Remember in section 3, we had done some exploratory analysis on stock attributes that are significant in determining stocks' trading venue distribution, there are plenty more attributes such as the volatility of a stock, the sector for which a stock belongs to, fundamental ratios of a stock, and many more characteristics. We want to test more of these attributes and see if any of them are potential keepers.

The advantage of using either an optimization model or a Naive Bayes Model is that both approaches are able to incorporate additional predictors without the burden of assuming a dependence structure among the predictors themselves. In the Naive Bayes case, conditional independence is assume; whereas in the optimization model case, the optimized parameters themselves do not need to convey any specific meanings. We would be delighted to see whether or not one model will outperform the other, and to compare the pros and cons of the two approaches.

Appendix

All codes are submitted on Github (You will need to contact author for permission to access the code base):

https://github.com/rcyeh/cfem2013/
https://github.com/rcyeh/cfem2013/MaketData_OptimalDelay
https://github.com/rcyeh/cfem2013/SubProject1_PriceImprovement
https://github.com/rcyeh/cfem2013/SubProject2_VenueLiquidity
https://github.com/rcyeh/cfem2013/SubProject3_MarketMakerCharacteristics

References

[1] Table with Rule 11Ac1-5 Sample Statistics for a Single Security "A" in May 2001. US Security
and Exchange Comission, http://www.sec.gov/interps/legal/slbim12a.htm, 28 Nov. 2013, Web.

[2] Table with Rule 11Ac1-5 Sample Statistics for a Single Security "A" in May 2001, Field List.
US Security and Exchange Comission, http://www.sec.gov/interps/legal/slbim12a.htm#q1, 28 Nov. 2013 Web.

[3] Inferring Trade Direction from Intraday Data, by CHARLES M. C. LEE, MARK J. READY, 1991.

[4] A test of the accuracy of the Lee/Ready trade classification algorithm, by Erik Theissen, February 2000.

[5] Graphical representation of particle shape using triangular diagrams: an Excel spreadsheet method. Earth Surface Processes and Landforms 25(13): 1473-1477, by Graham DJ and Midgley NG. 2000.

[6] Displayed and Effective Spreads by Market, by BLUME, M. E., AND GOLDSTEIN, M. A. 1992, University of Pennsylvania.

[7] The Dark Pool Debate, by ANA AVRAMOVIC, AND PHIL MACKINTOSH. August 22nd 2013, Credit Suisse.