Intraday Analysis and Prediction of Volume Distribution on the Stockholm Stock Exchange

AN EXPLORATORY STUDY OF VOLUME DISTRIBUTION AND AUTOMATED TRADING

HENRIK RIBOM, MATHIAS SJÖBERG
Intraday Analysis and Prediction of Volume Distribution on the Stockholm Stock Exchange

An exploratory study of volume distribution and automated trading

HENRIK RIBOM
MATHIAS SJÖBERG

Degree Project in Applied Mathematics and Industrial Economics (15 credits)
Degree Progr. in Industrial Engineering and Management (300 credits)
Royal Institute of Technology year 2015
Supervisors at KTH: Henrik Hult and Anna Jerbrant
Examiner: Boualem Djehiche

TRITA-MAT-K 2015:05
ISRN-KTH/MAT/K--15/05--SE

Royal Institute of Technology
School of Engineering Sciences

KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci
Abstract

The purpose of this study is to create a model of prediction for the volume distribution. Due to the lack of previous studies on the subject, an exploratory approach is used, with the purpose of serving as a proof of concept for further research. By looking at all market data from the Stockholm stock exchange the volume distribution of individual order books are matched with a mixed beta distribution and scaled by a prediction based on a linear regression. The model provided in this study outperforms the floating mean by quite a good margin. Some days are, almost by definition, impossible to get an accurate prediction on. Intraday news with a big impact have a tendency to skew the results away from the predicted value. To remedy this the initial model is expanded by using intraday data to catch up on trends.

Keywords:

Trade Volume, Regression, Prediction & Mixed Beta Distribution
Referat

Analys av volymfördelning på Stockholmsbörsen


Nyckelord:
Handels volym, Regression, Prediktion & Mized beta fördelning
Acknowledgements

Pantor Engineering AB

Pantor builds software for the financial industry with a focus on algo-rithmic trading of electronically traded securities. Their focus lies on low latency and at being tolerant to faults. The company was founded in Sweden back in 1999 and has about 25 employees.

Two people at Pantor deserve a special mention in this study. They are Rolf Andersson (CEO) and Lukas Magnusson. At first when we contacted Rolf, we had nothing but big dreams of what could be done. No real substance. But with the help of these lovely people we managed to find a subject that we could dive in to. Working at the office of Pantor at Luntmakargatan 26 in Stockholm was always a pleasant experience and we are grateful for the continuous support that we got during our long days there. Even when we asked questions that, to them, must have been rather trivial they always helped us and gave us food for thought on a daily basis.

Thanks a thousand times for an awesome experience.
/Mathias & Henrik

NASDAQ and SHoF

We would also like to say thanks to both NASDAQ and Swedish House of Finance for providing us with the immense amount of data. Without them we would not have been able to write about this subject. Big thanks!
## Contents

I Analysis and Prediction of Volume Distribution

1 Introduction

1.1 Background .............................................................................. 3
1.2 Problem Statement ................................................................. 6
1.3 Purpose .................................................................................. 7
1.4 Limitations ............................................................................ 7

2 Literature

2.1 Books ..................................................................................... 9
2.2 Other Literature ..................................................................... 9

3 Theoretical frame of reference

3.1 Gauss-Newton ........................................................................ 11
3.2 Linear Regression

3.2.1 Perfect model ...................................................................... 12
3.2.2 Imperfect Model .................................................................. 13
3.2.3 Prediction ........................................................................... 14
3.2.4 P-value and F-test ................................................................. 14
3.2.5 $R^2$, Adjusted $R^2$ and $\eta^2$ ........................................... 15
3.2.6 BIC, AIC and VIF ................................................................. 15
3.3 Q-Q Plots ............................................................................... 16
3.4 Beta Distribution .................................................................... 16
3.5 Standard Deviation ................................................................. 18

4 Methodology

4.1 Literary Study ........................................................................ 19
4.2 Data ....................................................................................... 19
4.3 Removing Data ....................................................................... 20
  4.3.1 Manual Removal of Non Fitting Data ................................... 20
4.4 Prediction Method ................................................................. 21
  4.4.1 Modeling the trade volume ................................................ 21
  4.4.2 Initial Linear Regression ..................................................... 22
  4.4.3 Assessment of Covariates ................................................... 23
9.1.2 Mean volume traded over time . . . . . . . . . . . . . . . . . . 57
9.1.3 Floating Mean vs Linear Regression on Intraday Data . . . . . 58
9.1.4 Results from Order Book Two and Tree . . . . . . . . . . . . . 58

Bibliography 61
Nomenclature

\( \beta \) \hspace{1cm} \text{Vector with scalars for x}

\( \beta_j^0 \) \hspace{1cm} \text{Assumed value of } \hat{\beta}_j \text{ in F-test}

\( \eta^2 \) \hspace{1cm} \text{Measurement of effect size}

\( \Gamma \) \hspace{1cm} \text{Gamma function}

\( \hat{\beta} \) \hspace{1cm} \text{Optimal } \beta

\( \hat{e} \) \hspace{1cm} \text{Optimal (minimum) residual vector}

\( \min_x \) \hspace{1cm} \text{Minimum of a vector with respect to x}

\( \nabla \) \hspace{1cm} \text{Symbol for the gradient of a vector}

\( \nabla^2 \) \hspace{1cm} \text{Second order gradient}

\( R^2 \) \hspace{1cm} \text{Coefficient of determination adjusted to sample size}

\( \partial \) \hspace{1cm} \text{Partial derivative}

\( \partial^2 \) \hspace{1cm} \text{Second order partial derivative}

\( \sigma \) \hspace{1cm} \text{Homoscedastic standard deviation}

\( \nu \) \hspace{1cm} \text{Volume distribution}

\( \varepsilon \) \hspace{1cm} \text{Scalar used in mixture beta distribution}

\( \varphi(x) \) \hspace{1cm} \text{Function of x}

\( \varphi_j(x) \) \hspace{1cm} \text{J:th element of the function } \varphi(x)

\( e_i \) \hspace{1cm} \text{Element i in the residual vector}

\( f(x) \) \hspace{1cm} \text{Function of x}

\( f_i \) \hspace{1cm} \text{The i:th element of the vector f}

\( J^i(x) \) \hspace{1cm} \text{The Jacobian of vector x}
\( R^2 \)  Coefficient of determination

\( s^2 \)  In the context of linear regression, an unbiased estimator of variance

\( s_k \)  In this study, a measurement on the effectiveness of predictions

\( s_k \)  Step \( k \) in the Gauss Newton algorithm

\( t_i \)  The \( i \)-th element of vector \( t \)

\( x_0 \)  Value at iteration 0

\( x_i \)  In gauss newton context, The \( i \)-th element of vector \( x \)

\( x_k \)  Value at iteration \( k \)

\( y_i \)  Element \( i \) in the response variable

\( y_i \)  The \( i \)-th element of vector \( y \)

AIC  Akaike Information Criterion

Algo  Short version of “algorithm”

AT  Algorithmic Trading

Backtesting  The process of testing a trading strategy on prior time periods. Instead of applying a strategy for the time period forward, which could take years, a trader can do a simulation of his or her trading strategy on relevant past data in order to gauge the its effectiveness. (Investopedia)

Basis points  Common unit of measure in finance. One basis point equals 1/100 of 1% i.e. 0.01%

BIC  Bayesian Information Criterion

Black-box  A black box is a device, system or object which can be viewed in terms of its inputs and outputs, without any knowledge of its internal workings (wikipedia)

C  Constant used in the derivative of mixture beta

CDF  Cumulative Distribution Function

D  Vector used in prediction with linear model

Dummy variable  A numerical value used to indicate a true/false statement, implying subgroups in the regression
e  Residual vector
EA  Execution Algorithm
F  F-value used in F-test
F(1,n-k-1)  F distributed with degrees of freedom 1 and n-k-1
f(x;\alpha,\beta)  Beta function of x with shapes \alpha,\beta
HFT  High Frequency Trading
L  Maximum value from the likelihood function
MD  Market Data
Order Book  An order book is the list of orders (manual or electronic) that a trading venue (in particular stock exchanges) uses to record the interest of buyers and sellers in a particular financial instrument. (wikipedia)
Prediction power  The power of the prediction according to the test that was setup in 4.4.2
q  Quantile
SD(x)  Standard Deviation of x
Var(x)  Variance of x
VWAP  Volume Weighted Average Price
X  Covariates matrix
Y  Response variable
Part I

Analysis and Prediction of Volume Distribution
Chapter 1

Introduction

Chapter one will address the background, formulation of problem, purpose and limitations of this thesis.

1.1 Background

"It is over. The trading that existed down the centuries has died. We have an electronic market today. It is the present. It is the future."  
Robert Greifeld, NASDAQ CEO, april 2011

Technology has paved the way for a revolution in the way we trade. The computerization of the stock exchanges began in the early 1970s with the introduction of the DOT system (designated order turnaround). In unison with the exponential increase in available computing power, Algorithmic Trading (AT) has blossomed both in volume and complexity. In order to understand AT one must first understand the difference between its two main subgroups.

1. Execution Algorithm (EA) - Seeking not to "lose" basis points using different trading algorithms to lower impact and hiding market entry.

2. High Frequency Trading (HFT) - Seeking to "win" basis points by market making, statistical arbitrage and other methods.

In what can be described as an arms race between institutions with large amount of capital using EA, and HFT firms hunting for basis points there seems to be a gap between the actual knowledge base and what is presented in academia and to the public eye.

There is no denying the fact that AT is a big part of the trading landscape. Even though some would argue that the initial "hype" is over, and that the industry now is
CHAPTER 1. INTRODUCTION

becoming more mature\[2\] it is still a topic that gains much attention in both the media and the finance industry. In 2009, estimates of HFT activity were as high as 61% - 73% of the traded volume on the American markets, only to fall to about 51% three years later.\[2, 3\] Even so, the days of easy arbitrage between markets are gone and spreads have never been lower.\[27\]

One obvious advantage of this development is that investors are able to control more parameters of the trading process. A good example of when this is favourable is when dealing with big orders. Let's define "big orders" as orders sizable enough to have a real impact on the current market equilibrium. This is bad for the investor for two main reasons.

1. **Market impact.** The market is a system governed by two parameters; supply and demand. This causes a rather paradoxical situation for users of the market. If you increase supply (sell stocks) you will get a lower selling price, and if you increase the demand (buy stocks) you will push for a higher price. For the average person these effects are so small that they have no real effect on the individual trades.\[8\] But as volume increases, so does the impact and the negative sides it brings with it.

2. **HFT** firms might identify that a big order is coming. They can then go in with an automated momentum algorithm, designed to profit (i.e. win basis points) from the increased activity. Given high enough frequency these basis points really add up, as the HFT industry had over a billion dollars in revenues 2014, in the US alone.\[4\]

Before the norm became AT, financial powerhouses like pension fund managers and insurance companies would simply hire a broker to carry out their trades. This broker would then try to find a suitable counterparty for the entire deal in one block, or alternatively go to the nearest stock exchange to "work the order". Market impact was obviously still a concern, but for which the remedy was the individual traders judgement and and general 'know how'.\[6\] This trader was also bound to a physical location (market). And communication between markets was, at least by today’s standards, very slow.

However, from advances in technology an opportunity for change presented itself. Market digitization removed the requirement of a physical presence. Now the marked data came through the Internet instead. At first the traders had DMA systems (Direct Market Access) that communicated with the different markets. Through the DMA system investors were able to manually put out orders on the different markets. Some orders were however executed at a less then optimal price, i.e not at the best quoted price across all markets. This due to human error an incapability to process information fast enough. So quite naturally the DMA quickly evolved in to what is now known as a SOR (Smart Order Router). What the SOR systems added to DMA was, as the name implies, the smart side. As a investor you generally do not care were and how a specific order is filled out. The investor just want the best price available. The SOR can then for example slice up the order in
1.1. BACKGROUND

to smaller parts and to out to different markets simultaneously in order to profit from minor price discrepancies between the markets.

In unison with the development of handling orders, there also came regulations. The Markets in Financial Instruments Directive (MiFID) is a law in the European Union that implements a unified set of rules and regulations to the 28 EU states, Iceland, Norway and Liechtenstein. The stated intentions of MiFID are to protect the consumers and to promote competition. The scope and impact of the law is very broad, but there are two very important parts in the context of this study.

Firstly what is called "best execution". In order to protect the costumers rights against the financial firms, MiFID requires that (within reason) the best execution of orders is taken for the clients. The "best execution" is a combination of the parameters; execution price, cost, speed, likelihood of execution and other relevant factors. Secondly MiFID brought with it a greater fragmentation of the markets. And where institutions previously only had to look at a single market they now had the possibility (and in a way, the obligation) to look at all of the markets. As these changes made the nature of trading more complex the emergence of the SOR was not only natural, it was a necessity.

<table>
<thead>
<tr>
<th>Users</th>
<th>Provider</th>
<th>Trading Venue</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investor</td>
<td>Intermediation</td>
<td>Primary Exchange</td>
<td>Closed Deal</td>
</tr>
<tr>
<td>Investor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investor</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![SOR diagram](image)

**Figure 1.1.** Classical trading value chain, with the impact of SOR visualized by the dashed lines

*Figure 1.1* demonstrates an important development for modern investors; the ability to seamlessly look at all the relevant markets simultaneously. This development implies a consolidation of order books between markets which results in their deepening. Mathematically, the deepening of order books is increasing the size of orders able to affect the current market equilibrium between buy and the sell side of the book due to a larger volume on both sides

Even though there are clear advantages that comes with AT, the industry has faced meticulous scrutiny by the media. This scrutiny is however not completely without justification. Speed has the unfortunate effect of amplifying faults in a system. Examples of when some faults got out of hand is the famous *Flash Crash*

5
and what happened to *Knight Capital* in the first of August 2012.

The 2010 *Flash Crash* a.k.a. "The Crash of 2:45" is the name of a famous event on the US markets. In only a few minutes the Dow Jones Industrial Average nose-dived 998.5 basis points, only to recover most of it within the next few minutes. As of April 2015 a London-based high-frequency trader was arrested by the police for allegedly playing a part in the flash crash. He stands accused of using a 'spoofing' program to create a large amount of sell orders to push down the prices, only to cancel them and buy the stocks at the lower price.\footnote{Knight Capital lost about 460 millions dollars in the time span of minutes due to a faulty line of code. Their computer put out orders for millions of stocks that it was not meant to and when it stopped sending the orders, Knight had assumed a net long position in 80 stocks of approximately $3.5 billion and a net short position in 74 stocks of approximately $3.15 billion.}\footnote{Knight Capital lost about 460 millions dollars in the time span of minutes due to a faulty line of code. Their computer put out orders for millions of stocks that it was not meant to and when it stopped sending the orders, Knight had assumed a net long position in 80 stocks of approximately $3.5 billion and a net short position in 74 stocks of approximately $3.15 billion.}

### 1.2 Problem Statement

To prevent the problems with big orders mentioned in 1.1, a common strategy is breaking them down into multiple smaller orders and placing them in the market(s) over a period of time. For stocks that are considered highly liquid a common strategy is matching some percentage of the entire traded volume. If an investor wants to buy 100'000 Ericsson B stocks and only want to partake in a maximum of 10\% of the trades during some time interval, he/she has to be sure that at least 1'000'000 Ericsson B are traded in that same interval. If only 500'000 stocks were traded in that time space you are facing unwanted risk and maybe would have liked to reconsidered the situation. This implies that a general understanding of how much volume that will be traded has a big value. Knowing the amount of trades that will occur during a given time period allows investors to count backwards to what percentage of the trades they have to enter during the intended time period (or for how long they will have to trade when only entering x\% of the trades). The current lack och publication on this subject calls for a "proof of concept" and an assessment of whether further research has potential.

### 1.3 Purpose

The purpose of Part I of this study is to explore and determine a mathematical foundation for intraday volume distribution on highly liquid order books (that in the context of this study is defined as companies included in the OMXS30 index as of writing the study). This is done in order to predict the volume better than an 'educated guess'. Evaluating prediction credibility and variance is at the very center of realizing the purpose. As of now there is a mismatch between the sophistication of the trading algorithm and trade volume predictions. This study aim to lesser the distance between the two and to act as a proof of concept for further research on the subject.
1.4. LIMITATIONS

Predicting the weather for tomorrow can be seen a relatively easy task, maybe even the weather for three days from now. But the ability to predict the weather quickly deteriorates as you look further in to the future. It is impossible to have a significant prediction of what the weather will be like one year from now. However one could postulate that it might be a bit warmer on average (due to global warming). The long term prediction is in this case based on a general understanding of the surrounding environment of the problem. In analogy with this, the purpose of Part II is to give a general understanding for how AT emerged and what the present has to offer. This is explored with the goal of gaining an understanding of the present and to act as a knowledge base for theories of what might happen in the future.

1.4 Limitations

The first limitation of this study is that only order books on the Stockholm stock exchange will be considered due to data and scope limitations. The data interval is also limited from November 2012 to data that is current as of writing the study [Spring 2015].

As mentioned in the purpose, this study is also limited to highly liquid order books. In the context of the study, 'highly liquid stocks' are defined as stocks in the OMXS30 index.

<table>
<thead>
<tr>
<th>OMXS30</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABB Ltd</td>
</tr>
<tr>
<td>Alfa Laval</td>
</tr>
<tr>
<td>ASSA ABLOY B</td>
</tr>
<tr>
<td>Atlas copco A</td>
</tr>
<tr>
<td>Atlas Copco B</td>
</tr>
<tr>
<td>AstraZeneca</td>
</tr>
<tr>
<td>Bolden</td>
</tr>
<tr>
<td>Electrolux B</td>
</tr>
<tr>
<td>Ericsson B</td>
</tr>
<tr>
<td>Getinge B</td>
</tr>
<tr>
<td>Hennes &amp; Mauritz B</td>
</tr>
<tr>
<td>Investor B</td>
</tr>
<tr>
<td>Kinnevik B</td>
</tr>
<tr>
<td>Lundin Petroleum</td>
</tr>
<tr>
<td>Modern Times Group B</td>
</tr>
<tr>
<td>Nordea Bank</td>
</tr>
<tr>
<td>Nokia Oyj</td>
</tr>
<tr>
<td>Sandvik</td>
</tr>
<tr>
<td>SCA B</td>
</tr>
<tr>
<td>SEB B</td>
</tr>
<tr>
<td>Securitas B</td>
</tr>
<tr>
<td>Sv. Handelsbanken A</td>
</tr>
<tr>
<td>Skanska B</td>
</tr>
<tr>
<td>SKF B</td>
</tr>
<tr>
<td>SSAB A</td>
</tr>
<tr>
<td>Swedbank A</td>
</tr>
<tr>
<td>Swedish Match</td>
</tr>
<tr>
<td>Tele2 B</td>
</tr>
<tr>
<td>TeliaSonera</td>
</tr>
<tr>
<td>Volvo B</td>
</tr>
</tbody>
</table>
Chapter 2

Literature

This chapter will present the literature used in this study.

2.1 Books

Harald Lang, Elements of Regression Analysis, 2014
Provided knowledge of regressions and common practice and mistakes when using it.

Used for reading up on the Gauss Newton Algorithm

Bruce E. Hansen, Econometrics, 2000
Famous work in the field of econometric, listing common practice and mistakes

2.2 Other Literature

http://www.r-bloggers.com/
Used to develop the our code in R

http://stackoverflow.com/
Used to develop the our code in R

Used to construct the interviews in part II
Chapter 3

Theoretical frame of reference

This chapter will clarify and in detail explain formulas and assumptions used in the study.

3.1 Gauss-Newton

The Gauss-Newton algorithm is one of many methods to solve non-linear least squares problems. The algorithm minimizes the sum of squares function values, given that it converges. With each iteration the sum of squares decreases until a minimum is reached. The method requires an initial guess of the function values. If this guess is not adequate the algorithm does not converge. Naturally a good guess leads to faster convergence.

The goal of the algorithm is to minimize a function on the form

\[ \min_x \varphi(x) = \frac{1}{2} ||f(x)||_2^2 \]  \hspace{1cm} (3.1)

Where \( f \) is a \( m \)-dimensional vector function of \( x \) where the \( i \):th element of \( f \) is defined as

\[ f_i = S(x_i, t_i) - y_i \]  \hspace{1cm} (3.2)

\( y_i \) and \( t_i \) are observed values, hence \( f_i \) is the residual. To use the algorithm the hessian matrix and second order hessian matrix need to be calculated, these are defined as:

\[ \nabla \varphi(x) = J'(x)f(x), \quad J(x) = f'(x) \]  \hspace{1cm} (3.3)

\[ \nabla^2 \varphi(x) = J(x)^t J(x) + \sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x) \]  \hspace{1cm} (3.4)
CHAPTER 3. THEORETICAL FRAME OF REFERENCE

Given this information the min_x \( \varphi(x) \) is calculated by:

1. Guess an appropriate \( x_0 \)

2. \( s_k = \left( J^T(x_k)J(x_k) \right)^{-1} J^T(x_k)f(x_k) \), \( x_{k+1} = x_k + s_k \) (3.5)

Repeat step 2 until it converges

Deriving this formula can be done from the one dimensional Newton’s method, which states that:

\[ x_{k+1} = x_k + \nabla^2 \varphi(x)^{-1} \nabla \varphi(x) \] (3.6)

From (3.1) we show that (3.3) by:

\[ \nabla \varphi_j(x) = \sum_{i=1}^{m} \frac{\partial \varphi_i(x)}{\partial x_j} = \sum_{i=1}^{m} f_i(x) \frac{\partial \varphi_i(x)}{\partial x_j} = J^T(x)f(x) \] (3.7)

(3.4) is calculated by the following equation, ignoring the second derivative term:

\[ \nabla^2 \varphi_{jk}(x) = \sum_{i=1}^{m} \frac{\partial f_i(x)}{\partial x_k} \left( f_i(x) \frac{\partial f_i(x)}{\partial x_j} \right) = \sum_{i=1}^{m} \left( \frac{\partial f_i(x)}{\partial x_k} \frac{\partial f_i(x)}{\partial x_j} + \frac{\partial^2 f_i(x)}{\partial x_k \partial x_j} \right) \approx J^T(x)J(x) \] (3.8)

Hence what is needed for (3.6) is known. Note that the function \( \varphi(x) \) does not need to have second order derivatives [11].

3.2 Linear Regression

3.2.1 Perfect model

A linear regression is a method to show relations between covariates and a response variable. The response variable is modeled as a dependent random variable which depends on deterministic covariates. Examples of covariates could be the price of a stock in \( \mathbb{S} \) or a "dummy" for genders with a 1 for female and 0 for male. An example of a response variable used in this study is total volume traded on a day. On vector form the relation is as follows [12]

\[ Y = X\beta + e, \text{ or } y_i = x_i \beta + e_i \text{ or } i = 0, 1, 2 \ldots n \] (3.9)

where the residuals \( (e_i) \) are assumed independent normally distributed with \( \mathcal{N}(0, \sigma) \) and thus \( \mathbb{E}(e_i) = \sigma \) and \( \mathbb{E}(e_i e_j) = 0, j \neq i \), these are known as the normal equations. With the normal equations an estimation of \( \beta \) is constructed with Ordinary Least Squares (OLS). Some calculations give the optimal \( \beta \):
3.2. LINEAR REGRESSION

\[ \hat{\beta} = (X'X)^{-1}X'Y \]  
(3.10)

Using \( \hat{\beta} \) from (3.10) it can be proven that covariance matrix of it is:

\[ \text{Cov}(\hat{\beta}) = (X'X)^{-1}\sigma^2 \]  
(3.11)

In calculation with this method \( \sigma^2 \) is approximated by the unbiased estimator \( s^2 \)

\[ s^2 = \frac{1}{n - k - 1}\|\hat{e}\|^2 \]  
(3.12)

3.2.2 Imperfect Model

According to the econometric framework, residuals with the same standard deviation (known as homoscedasticity) rarely exist and thus the calculations are often done under the assumption that they are not (known as heteroskedasticity) [13]. This has some implications for the regression results, most significant for this study is the change structure of the covariance matrix, called White’s consistent variance estimator [12]:

\[ \text{Cov}(\hat{\beta}) = (X'X)^{-1}X'D(\hat{\varepsilon}^2)X(X'X)^{-1} \]  
(3.13)

\[ D(\hat{\varepsilon}^2) = \begin{pmatrix} \hat{\varepsilon}_1^2 & 0 & 0 & 0 \\ 0 & \hat{\varepsilon}_2^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \hat{\varepsilon}_n^2 \end{pmatrix} \]  
(3.14)

There are some common problems when dealing with regression that out of necessity is dealt with in this study. Multicollinearity is a problem that can arise and skew the results of the regression, due to two or more covariates being to some extent linearly dependent. For example, if a regression was to determine the wage at a certain age depending on both experience and education, then there should exist some multicollinearity since your years of experience is dependent on how many years you spent in school. Solving this can be done by increasing the number of observations or removing and/or changing the covariates. It is commonly found by using the VIF test described in 3.2.6. Additionally, if dummy covariates that represent the same fact, i.e. male and female, are used in a regression the multicollinearity is removed by letting one of them be represented by the intercept. So if you want to test something on males and females, a dummy with 1 for male and 0 for female (or vise verse) should be used, not one covariate each for male and one for female. In this example the value for female would be the intercept, \( \beta_0 \).

Another problem is endogeneity. This occur when the residual is dependent on at least one covariate. Solving this can be done with 2SLS regression which will not be expanded further since it will not be used in this study due to a lack of...
replacement covariates needed. Identifying endogeneity is however still of use for an analytic purpose.

There are several types of endogeneity but only one discussed in depth in this study: “Missing relevant covariates”. As the name states this problem arise when some of the covariates that explain the dependent variable are missing. This leads to the residual containing information from these missing covariates. The remedy for this problem is in theory to insert the missing covariates, which proves to be a difficult task in this study, as discussed in 6.2.

### 3.2.3 Prediction

Let $Y_i$ define the volume traded at day $i$ and $X_i$ define the observed data from day $i-1$. To make a prediction on the value of tomorrows volume i.e $Y_{n+1}$ using today’s observed data i.e $X_{n+1}$ a linear regression is used. In this method $Y_{n+1}$ is set to 0 and observed data for time $n+1$ added to the dataset. Then a new covariate $D$ is introduced which is 0 for 1,2...n and -1 for $n+1$. Running the regression will give $D$ the estimated value of $y_{n+1}$ and the standard deviation of it corresponds to the standard deviation of $Y_{n+1}$. To make several predictions on known data, all data after the prediction date are removed temporarily and the method above is used. Then the data previously removed is attached again and a new prediction date (presumably the next day) could be used.

### 3.2.4 P-value and F-test

The p-value can be described as the probability to obtain the observed result (or 'more extreme' results), given that the null hypothesis is true. The null hypothesis is in the case of regression that $\beta_i$ is equal to zero. This implies that a low p-value is preferable in order to gain a high significance. The p-value take a range of values from 0 to 1 where 0 represents a 0 % probability of the $\beta_j$ true value being zero. Setting up the test in this manner is referred to as testing the null hypothesis, i.e assuming a specific $\beta_i$ is zero and testing the probability of this.

To derive this probability a F-test is used since $X$ is assumed F distributed under the null [12]. In the heteroscedastic model only one beta at a time can be tested contrary to the homoscedastic where testing more than one is possible, this is one of the implications mentioned in 3.2.2 [12]. The F-test is done in the way described below.

Calculate the F-value:

$$ F = \left( \frac{\hat{\beta}_j - \beta_j^0}{SD(\hat{\beta}_j)} \right)^2, \text{ where } \beta_j^0 \text{ is set to } 0 \quad (3.15) $$

Calculate the p-value:

$$ p = P(X > F) \text{ where } X \in F(1, n - k - 1) \quad (3.16) $$
3.2. LINEAR REGRESSION

where \( n \) is the number of observations and \( k \) is the number of covariates. Alternatively a confidence interval with confidence level of \( \alpha \) % for a chosen beta is constructed by:

\[
\hat{\beta}_j \pm \sqrt{F_{\alpha}(1, n-k-1)SD(\hat{\beta}_j)}
\]  

(3.17)

The probability of the true \( \beta_j \)-value being outside the interval is \( \alpha \)%.

3.2.5 \( R^2 \), Adjusted \( R^2 \) and \( \eta^2 \)

\( R^2 \) is a measurement from 0 to 1 explaining "goodness of fit". Where 1 means that all of the variance is explained by the model. The same is true for adjusted \( R^2 \) with the key difference being \( R^2 \) is not adjusted for sample size. A clear guideline for what level is tolerated or viewed as "good" is not set in this study, \( R^2 \) and adjusted \( R^2 \) are used to portray the efficiency of the model, not to evaluate \[4.4.2\].

The definition of \( R^2 \) is:

\[
R^2 = 1 - \frac{Var(\hat{\epsilon})}{Var(y)}
\]  

(3.18)

The definition of adjusted \( R^2 \) is:

\[
\bar{R}^2 = 1 - \frac{Var(\hat{\epsilon})(n-1)}{Var(y)(n-k-1)}
\]  

(3.19)

Additional \( \eta^2 \) is used in linear regression to determine the effect size of each covariate. It can be constructed by running the regression without the covariate you want to analyse and construct the \( R^2 \) for that model, for convenience called \( R^2_* \) and using the following equation:

\[
\eta^2 = \frac{R^2 - R^2_*}{1 - R^2}
\]  

(3.20)

What size of \( \eta^2 \) that is sufficiently large is not exact. A rule of thumb is 0.1 < small, 0.1 < medium < 0.5, 0.5 < large [14].

3.2.6 BIC, AIC and VIF

Bayesian information criterion or BIC is used to evaluate what set of covariates is optimal in a linear model since too many covariates might lead to a less accurate result. The smallest BIC value indicates what variation of the model that best fit the reality. Both BIC and AIC uses the maximum value from the log likelihood function \( \ln(L) \) which is defined as:

\[
\ln(L) = \sum_{i=1}^{n} \ln[(2d_i - 1)P(x_i\hat{\beta}) + 1 - d_i]
\]  

(3.21)
where \( d_i = 1 \) if event under study occurred and \( d_i = 0 \) otherwise. BIC is defined as \[15\]:

\[
BIC = -2 \ln(L) + k \times \ln(n) \tag{3.22}
\]

Akaike information criterion or AIC is similar to BIC. It is a function used for selecting the best model compared to the alternatives. In contrast to BIC the model generating the largest number is considered the best with AIC. AIC is defined as \[16\]:

\[
AIC = 2k - 2 \ln(L) \tag{3.23}
\]

VIF or Variance inflation factor is a measurement for multicollinearity among co-variats. To test a covariate the equation (3.9) is used with \( Y \) removed and replaced by the covariate that is to be tested. Then \( R^2 \) is calculated from equation (3.18) and finally:

\[
VIF = \frac{1}{1 - R^2} \tag{3.24}
\]

If the VIF value exceeds 10 then the covariate is a cause of multicollinearity \[17\].

### 3.3 Q-Q Plots

To understand and use Q-Q plots one must first understand quantiles. Quantiles (q) are obtained from the inverse of a CDF of a random variable. All quantiles have the same interval length and the i:th quantile represent the probability of the random variable being less than \( x \) is at most \( i/q \).

\[
P(X < x) = \frac{i}{q} \tag{3.25}
\]

Q-Q plots is an investigating method for determining whether two distributions are similarly distributed and thus linearly related using their quantiles. Quantiles obtained from one distribution is chosen for y-values and x-values are chosen from the other. Comparing these points to the line \( y = x \) gives an interpretation as to the similarity of the two distributions. If the points close to the line the implication is that they are similar and vice versa.

### 3.4 Beta Distribution

The beta distribution is used frequently in statistics and in various fields. It is defined on the interval \([0,1]\) and by two shapes parameters commonly named \( \alpha \) and \( \beta \). The density function of the beta distribution is as follows:
3.4. BETA DISTRIBUTION

\[ f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \]  

(3.26)

The volume traded can be modeled with a beta distribution. Each trade is then viewed as if it was a generated random beta variable. This variable is originally defined on an interval \([0,1]\) where 0 represent the start of the daily trade and 1 the end. It is common practise to change this interval by removing \(x = 0\) and \(x = 1\) and replacing them with for example \(x = 0.001\) and \(x = 0.999\) (example taken from this study) [18]. This is done due better reflect the reality, since the edges of the PDF does not reach infinity with this change.

Generating a large number of values, for example: \(x_1, x_2, \ldots , x_{10000}\) with some approximated parameters the frequency of \(x_i\) in each part of the day represents the frequency of trades in that part of the day (this has previously been observed [19]). Note that this modeling gives the ratio between the volume traded not the actual value. The actual value could be retrieved by generating as many \(x\) as there are trades during a day or more easily both models can be scaled as percentages and compared, this is shown in 3.1.

**Figure 3.1.** Random beta distribution, chosen to use the similar shapes to the ones found in this study.

Along with the beta distribution a mixture beta distribution is used in this study. A mixture beta distribution is a linear combination of two or more beta distributions. The relevance of this distribution is explained in 4.4.1. For calculations made further on it is important to maintain probability properties of the original distributions. These are kept by using the following constraint when constructing the distribution [20]:

\[ 17 \]
CHAPTER 3. THEORETICAL FRAME OF REFERENCE

\[ f(x) = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n, \text{ where } \alpha_i \geq 0 \text{ and } \alpha_1 + \alpha_2 + \ldots + \alpha_n = 1 \] (3.27)

The mixture beta distribution used in this study is defined as:

\[ v = \left(1 - \frac{x}{\varepsilon}\right) \frac{\Gamma (\alpha + \beta)}{\Gamma (\alpha) \Gamma (\beta)} x^{\alpha - 1} (1 - x)^{\beta - 1} + \frac{x}{\varepsilon} \frac{\Gamma (\gamma + \delta)}{\Gamma (\gamma) \Gamma (\delta)} x^{\gamma - 1} (1 - x)^{\delta - 1} \] (3.28)

The variable \( \varepsilon \) determines the weight of each beta distribution at the current time \( x \) while keeping the total distribution continuous. As time approaches 1 the weight shifts from the first beta distribution to the second one.

Both the mixture beta distribution and the regular beta distribution are first order differentiable, which is a requirement in order for equation 3.5 to work. Simplified the derivation of beta can be written:

\[ \frac{\partial}{\partial x} \left( C x^{\alpha - 1} (1 - x)^{\beta - 1} \right) = -C x^{\alpha - 2} (1 - x)^{\beta - 2} (1 + \alpha (x - 1) + (\beta - 2)x) \] (3.29)

The simplified derivation of mixture beta is:

\[ \frac{\partial}{\partial x} (\text{Mixture Beta}) = -C x^{\alpha - 2} (1 - x)^{\beta - 2} (1 + \alpha (x - 1) + (\beta - 2)x) + C x^{\alpha - 1} (1 - x)^{\beta - 2} (\alpha (x - 1) + (\beta - 1)x) + D x^{\gamma - 1} (1 - x)^{\delta - 2} (\gamma (x - 1) + (\delta - 1)x) \] (3.30)

### 3.5 Standard Deviation

In this study the corrected sample standard deviation is used and referred to as 'standard deviation'. This formula is used instead of the uncorrected one since it corrects the bias in the estimation of the population variance. It is defined as:

\[ SD(x)^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \] (3.31)

where \( n - 1 \) corresponds to the degrees of freedom
Chapter 4

Methodology

This chapter will go over the method as well as why it was chosen.

4.1 Literary Study

Collecting data for the analysis was done by reading scientific publications as well as the literature in 2.1 and 2.2. The supervisor at KTH, Henrik Hult, gave advice on common practice and methods used and helped us narrow the search for relevant information to the analysis.

All used publications were found by searching with Google Schoolar and the KTH library. Search words used can be found in the table below.

<table>
<thead>
<tr>
<th>Literature study</th>
<th>Search words</th>
</tr>
</thead>
</table>

4.2 Data

As NASDAQ would face monetary losses as a result of the data becoming public, the data storage has been secretive and secure by nature. Non disclosure agreements had to be signed. To ensure security the raw-data (hundreds of gigabytes of text files) was stored by Pantor and only be accessible through in to their servers.

The huge amount of information is given in ITCH-feed witch is the direct data-feed protocol that NASDAQ uses. This is saved in a format not readable to humans and consist of every action on the Stockholm stock exchange. This is obviously a huge amount of information and the handling of said information is important to be able to process the problem in a timely fashion and making calculations more efficient.

The first step is to filter out only the relevant information from the raw data and then save it in a suitable format (tab-separated in our case, gives easy handling in
CHAPTER 4. METHODOLOGY

R). The relevant information for this study is the information surrounding executed orders on individual order books. The resolution of the raw data is 1/1000 seconds, and is lowered to 1 minute in the final format. This entails that all trades during one minute are accumulated and assigned to the minute. The price for the minute in question is the VWAP that same minute. Continuous trading last for 505 minutes during a day, i.e starts at 09.00 and ends at 17.25. Followed by a Pre-close period with no automatching. The Pre-close period lasts approximately for 5 minutes and ends with the closing call uncross that randomly among Order Books takes place between 17:29:30 and 17:30, according to NASDAQ OMX Nordic Market Model 2.17, section 4.2.3

As a frame of reference the size of the sorted data from the 600 days is about 15MB. This is a reduction of size by a factor of about 10’000 and is simplified the handling and computation time by quite a bit.

4.3 Removing Data

As stated in the purpose the aim of this study is to predict the intraday volume distribution. During the roughly three year period observed there are some days where trading is not transpiring from 9.00 to 17.30. Those days are mainly public holidays where trading ends at 13.00. They were removed from the data set since the method relies on the volume trade per full day.

4.3.1 Manual Removal of Non Fitting Data

This topic is crucial for this analysis and at the core of it is the unanswered question "exactly what can be predicted in a mathematical model?". Or perhaps it is one step further, "exactly what can be predicted?". The difficulty in predicting stock price is widely acknowledged and one reason behind this is the human psychology that is included in trades [21]. Hence, the price of even the most stable stocks are volatile at times. While this does not necessarily imply that the volume traded is as volatile, a correlation between the two is expected, see 4.4.3.

With the volatility examined the next step is its predictability. Is it, for example, feasible to predict a doubling of volume traded from one day to another? Using the method in this study the answer is: yes and no. As explained in 6.2 a covariate for "important news" would be needed in order to explain a large deviation. Furthermore data that does not fit the model will change the outcome to the linear regression. With this in mind the value of keeping the data in the model needs to be examined. The predictions should improve on the remaining days when non fitting data is removed, this does not however automatically mean that it should be removed.

There is a fine line but an important difference between recognizing inadequate data and cherry picking. The purpose of removing data would be to recognize what is not predictable with any conceivable change to the model. i.e a company just signed a big client leading to an increase in the volume traded. Events like this
4.4 Prediction Method

This section will outline and motivate the method used when predicting the trade volume. For a summary of the method see 4.5

4.4.1 Modeling the trade volume

A floating mean distribution for a fitting period of days is calculated since the observed volume distribution differs between the days. The "mean volume traded distribution" is retrieved by taking the mean value of volume traded at each point in time during the day and collecting them in a cumulative distribution. The question of how many days should be included in the "mean volume traded distribution" is relevant for increasing the correlation between each predicted day and the model used. Since trading patterns presumably change over time and irrelevant data could skew the results and thus reduce the correlation. This is reviewed by sampling and empirically determining the optimal number of days.

In this study three methods of modeling the volume distribution are analyzed, and with the Gauss Newton algorithm (explained in section 3.1) the optimal value of the parameters in each model is obtained. The appropriateness of each distribution is then measured with correlation between the floating mean volume traded distribution and the model and illustrated in Q-Q plots. The distributions analyzed are as follows:

1. Normal distribution
2. Beta distribution
3. Mixed beta distribution

Normal distribution was used as a benchmark and as an initial guess. The beta distribution was tested due to its similarity with the trading pattern. Finally this was expanded on with the mixture beta distribution (3.28). The motivation for use of a mixture is the time difference in the opening between the American and Swedish stock markets. When news of the trading at the US market reaches the Swedish one volume traded is expected to increase. With this occurring at the same time every day the effect is expected to significantly influence the result and has the characteristics of a smaller day (with opening and closing). Hence the second beta distribution.
4.4.2 Initial Linear Regression

A heteroscedastic linear regression with covariates specified in 4.4.3 is used to make a prediction on the total volume traded per day. The heteroscedastic model is used due to the expected difference in error terms. It is reasonable to assume that some periods are less volatile then others thus creating sub groups of variance. If no major developments occur in a particular company for a period of time, the variance in the corresponding order book should differ from a period with major developments, presumably the variance would be larger in the latter one.

To assess the prediction power of the linear regression the accumulated volume per day will be predicted on historical data for every day of a three month period, chosen to be the period leading up the latest available data (January-April 2015). These results are then compared with the floating mean of the volume from a period. The period for the floating mean is derived empirically to give the most accurate result. The floating mean is meant to reflect and to roughly quantify an *educated guess* and can therefore be used as a benchmark for the prediction.

The prediction power will be tested on three different order books to ensure that results are not limited to one of them. Prediction power of the model is measured in two ways:(1) A test of errors is constructed in similarity with (3.12) but with the key difference that the factor is \( \frac{1}{n-1} \) and \( |\hat{e}|^2 \) is defined as \( |\text{predicted value - true value}|^2 \). This is used to measure the errors in a way that is comparable for different lengths of time periods. For simplicity this is henceforth called \( s^2 \) and presented in the results as \( s \) (for the convenience of smaller numbers). Note that since several linear regression are used the usual measurement, i.e (3.12), is not applicable.(2) The correlation between predicted value and true value during the three month period is measured.

Using \( s^2 \) provides a comparison to the floating mean while the correlation evaluates the usability of the model. Since it is outside the scope of this study to examine what level of inaccuracy is tolerated the actual value of \( s^2 \) is not of interest, only the difference between the \( s^2 \) from the linear regression prediction and the floating mean prediction.

Additionally, several combinations of covariates are tested with BIC, \( R^2 \) and \( \overline{R}^2 \) to find the best linear model describing the data. The motivation for excluding BIC is that it is preferred to AIC since the data set is determined to be sufficiently large \( [22] \). This information is used in combination with the \( s^2 \) and correlation to determine the optimal set of covariates and thus the optimal model under the limitations of this study. The final model is then tested with VIF to verify that no multicollinearity exists.

Finally, to generate the prediction of the cumulative volume distribution the modeled distribution from 4.4.1 is scaled with the total volume of the day and thus a prediction for all possible intervals during the day is retrieved. In this context scaled denotes multiplying the distribution with values spanning from \([0,1]\) with the total volume making \([0,\text{total volume}]\) (y-axis).

It should be noted that the aim of this study is not to achieve the most significant
4.4. PREDICTION METHOD

method of predicting but rather to achieve the best possible prediction. With this mathematical elegance is sacrificed for efficiency. Thus the decision of whether to keep a covariate is based on the results from the prediction not $R^2$, adjusted $R^2$ or BIC, the test is hence used as an indicator not a deciding factor. And while these test may yield similar conclusions usability in reality is the deciding factor.

It should be noted that the aim of this study is not to achieve the most significant method of predicting but rather to achieve the best possible prediction. This approach lacks a certain mathematical elegance. This is a compromise that is in line with the "applied" nature of this study. Thus the decision of whether to keep a covariate is based on the results from the prediction, and not strictly on $R^2$, adjusted $R^2$ or BIC. These test are hence used as an indicator and not as a deciding factor. And while these test may yield similar conclusions, usability in reality is the deciding factor.

4.4.3 Assessment of Covariates

This section will address and motivate the use of each covariate presented in the table below.

<table>
<thead>
<tr>
<th>Response Variable</th>
<th>Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Volume Today</td>
<td>SD of Price Yesterday</td>
</tr>
<tr>
<td></td>
<td>Change in Volume Yesterday</td>
</tr>
<tr>
<td></td>
<td>Change in Price Yesterday</td>
</tr>
<tr>
<td></td>
<td>OMXS30 Precentual Change Yesterday</td>
</tr>
<tr>
<td></td>
<td>Dummy for Monday</td>
</tr>
<tr>
<td></td>
<td>Dummy for Friday</td>
</tr>
<tr>
<td></td>
<td>Dummies for Time Interval</td>
</tr>
</tbody>
</table>

The standard deviation of yesterday’s price should be positively correlated with volume traded if there is a "unexpected" change in volume and negatively correlated if there is "expected open interest" [23]. Assuming that the day to day business experience little to no change in "expected open interest" most change in volume should be attributed to the "unexpected" change. In the context of this study this implies that the standard deviation of yesterday’s price should ultimately be positively correlated with the today’s volume. Whether this assumption is justified or not is verified by observing the sign of the covariate in the linear regression, given that it is in fact positive and significant, then the assumption should be correct. This relies on the trend change in "unexpected" volume to last during the day and effect the next. Stating that the total volatility of one day linearly affects the total volume the next is a questionable assumption. However, in the context of this study, it is seen as necessary assumption, which is a recurring argument in motivation of covariates. Since the prediction is based on historical data with one
day’s resolution all covariates are motivated by assumption of yesterday significantly impacting today.

Under microeconomic assumptions a price change would lead to a change in demand and since the market is imperfect the shifted equilibrium would be restored by an increase in volume. With this argument a covariate for price change between the day before yesterday and yesterday, i.e. "change in volume yesterday", will be used in the prediction of today. Accordingly a "volume change" covariate with the same structure is tested with the argument that with instruments such as VWAP used by traders the change in volume yesterday could affect the volume today. Both these covariates are tested as: dummies (up or down), % change and actual change. If price do indeed effect the volume tomorrow it would presumably effect the volume of yesterday as well. With this in mind, whether or not multicollinearity between the covariates exist or not is evaluated.

The Stockholm Stock Exchange is only open on weekdays. This means that information received on Friday 17:30 can be applied earliest Monday 09:00, which adds an extra uncertainty to Mondays in particular. To account for this the dummy "Monday" is introduced. Since trades that would theoretically been done on Saturdays or Sundays are pushed to Mondays it is assumed that Mondays will be positively correlated with the volume traded. With this in mind a dummy for Fridays is introduced as well to test if the last trading day have an effect on traders.

"Time covariates" are also introduced. For example these could be a dummy variable for the years 2012, 2013, 2014, 2015. These are evaluated since the order books could have different mean traded volume for different periods. Since the same argument could be used for introducing six month covariates or three month for that matter, what time is most representative time period to use is evaluated by testing with dummies for every three month period, six month period and year. When predicting the volume for January 1st 2015 and using for example a year covariate the data at this time would be insufficient. Time covariates are thus not used this way but rather they are "counting backwards" from the current date to the time specified. A six month covariate used to predict from January to April 2015 would then represent the dates October 2014 - April 2015. Thus eliminating the problem with insufficient data for the current period. This would shift the problem of insufficient data to the first period of data were predictions are not made.

The last covariate used is the percentual change in the OMXS30 index the previous day. This is based on the assumption that there is a correlation between the OMXS30 index yesterday and the volume traded of a individual order book today. As all the order book applied in this study are a part of OMXS30, this is not seen as a unjust assumption.

4.4.4 Linear Regression using Intraday Data

If some time \( t_n \) has passed of the day, if falls in line with intuition that the information obtained during during this interval \([0,t)\) could be used to improve the initial prediction. In order to utilize that information, a second linear regression is
4.4. PREDICTION METHOD

done. It is based on the data set of the predicted values from the daily model is used. The regression has the response variable "difference between the true value and predicted value at the end of the day" and the covariate "difference between the true value and predicted value at time \( t_n \)". Using the same method as in 4.4.2 a prediction of the response variable is obtained and it is then used to scale (i.e. correct) the initial prediction. In other words, if an error of x% is found at time \( t_n \), the regression predicts the error at the end of the day based on this by scaling x with the corresponding \( \beta \). To improve the prediction power additional covariates are used, namely, "difference at time \( t_n \)" and "SD in price at time \( t_n \)". The regression it thus:

\[
\text{Difference at closing} = \beta_0 + \beta_1 \times \text{(difference at time } t_n) + \beta_2 \times \text{(intraday SD at time } t_n) + \beta_3 \times \text{(Price difference between closing yesterday and at time } t_n) + e
\]

The covariates "intraday SD at time \( t_n \)" and "Price difference between closing yesterday and at time \( t_n \)" are chosen with similar reasoning as described in 4.4.3. Due to this, the reasoning behind the choice will not explained once more, however, they are tested the same way described. Note that, since "intraday SD at time \( t_n \)" is not defined for one value (3.31) this linear regression cannot be used at time one \( (t_1) \), which in this study is minute one. This could be solved by removing the covariate in the regression for \( t_1 \). But in order to maintain consistency this change is not made and \( t_1 \) is left out from the calculation. Since use of predictions made at the first minute are not easily determined and a change is easily made it will not be investigated. Another simplification made is if no trades are made the first minute, the price used at minute one is the first price at the next minute with a trade. Once again whether this simplification adds or subtracts value is difficult to assess. Arguably, it is most likely have no effect at all, except for the algorithm working properly since in the data set this occurred only three times. Referring to the argument made in 4.4.2 this simplification is viewed just if the covariate yields significant results.

To make predictions on a three month period with this expansion the initial prediction needs to be expanded in order to give significance to the first linear regressions made. If not, the first linear regression would have no historical predictions to base the regression on. A large number of initial predictions could decrease the initial prediction power and thus affecting the expansion. A small number would keep roughly the same prediction power but as previously stated decrease the significance of the intraday prediction. This potential trade-off is examined and the optimal number of predictions needed is then decided. The optimum found is not necessarily the same for finding the optimal amount of historical data to scale the prediction with and to use as initial data in the linear model, which is thus decided separately. In other words: the number of relevant days when constructing a prediction on the trading patterns (distribution) is not necessarily the same as when constructing a prediction on the total volume traded (linear regression).

To make calculation more efficient the actual mean value at time \( t \) from the
selected data is used instead of a mean value calculated with Gauss Newton. With 504 minutes (505 minus the first one) and 60 predictions that makes 60*504 = 30240 linear regressions if a prediction with each new data point is made. If Gauss Newton was used then one for each linear regression would be needed. Along with this comes the risk a Gauss Newton calculation not converging. One ill-fitted period would result in the model defaulting. It was thus decided to neglect the Gauss Newton algorithm in this step.

4.5 Summary of method

This section provides a summary of the prediction method used in the study, followed by a visualization in Figure 4.1.

1. Sort the data to minute resolution on a chosen order book

2. Choose the best distribution for the volume trading pattern, with Gauss Newton algorithm.

3. Choose the linear regression model with the best prediction power when predicting the total volume traded during a day (not necessarily best by other tests such as $R^2$ and BIC).

4. Scale the model from 2 with the volume from 3 to obtain the prediction for the entire day.

5. Use data gathered during the day up to time $t$ to run an additional regression. This regression predicts the outcome of the trade volume in relation to the first prediction.
4.5. SUMMARY OF METHOD

Figure 4.1. Visualization of method
Chapter 5

Results

This chapter will present some of the results gathered during the process of this study. In order to archive consistency without presenting excessive results the first one tested of the three order books examined will be presented and henceforth referred to as order book one. This is due to order books yielding similar results, or rather non-contradicting results. There is one exception to this which is the result presented in 5.3. Here result from all three order books are presented to show how results deviate between order books, which in this case yields different conclusions as to the effectiveness of the model. In other words: when the same conclusions are made from slightly different results of the three order books, only one is presented to spare the reader excessive results. Results for the remaining order books can be found in the appendix.
5.1 Evaluating the Parameterization

With the Q-Q plot framework a perfect model would have all the plotted points on the y=x line indicating they are virtually the same. The difference of each point and the line represent the difference in value at that particular time. Thus, the first dot (counting left to right) is minute one, second minute two and so on until minute 505. These plots are made with random generated distributions plotted against the mean value distribution with data from all days available. For theory of Q-Q plots see 3.3.

Observing the graphs plotted one can see that the normal distribution deviates from the values in the beginning of the day and at the end, while representing the middle of the day well. The beta distribution on the other hand, manages to account for trades during the entire day, with the exception of a small miss when about 75% of the day has gone by. As explained in section 4.4.1 this is due to the American market opening at that time. Finally the mixture beta distribution has no apparent misses with the y=x line and seem to be a almost perfect fit. From these results it is clear that the mixture beta best represent the mean volume traded, and in sample testing with random days the mixture beta outperformed the others as well.

To show the implication of the Q-Q plots the cumulative plot of mixture beta distribution as well as the mean volume traded distribution is presented in the bottom right corner. The mean volume traded distribution is shown in percent to demonstrate the similarities in the distributions. The red line is the mixture beta distribution and the black line is the mean volume traded distribution. The correlation between the mixture beta and the mean volume traded is 99.98%.
5.2 DERIVING THE FINAL MODEL

5.2 Deriving the Final Model

<table>
<thead>
<tr>
<th>Tests</th>
<th>Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>R^2</td>
</tr>
<tr>
<td>20563.538</td>
<td>0.28508269</td>
</tr>
<tr>
<td>20557.894</td>
<td>0.28407038</td>
</tr>
<tr>
<td>20552.357</td>
<td>0.28324181</td>
</tr>
<tr>
<td>20546.736</td>
<td>0.28224644</td>
</tr>
<tr>
<td>20541.737</td>
<td>0.28056270</td>
</tr>
<tr>
<td>20544.405</td>
<td>0.14080829</td>
</tr>
</tbody>
</table>

Table 5.1. BIC, R^2 and Adjusted R^2 on a number different model selections for the linear regression, made on all available data (2012-2015) as well as VIF test on the final model. 1 means the covariate was used and 0 that it was not.

The final model is derived by testing different sets of covariates described in 4.4.2. In Table 5.1, the covariate with least impact is removed in each step and so long as the BIC is decreasing with the current model it is accepted as the new model which is done until the iteration where BIC is no longer decreasing. This method is used from row one to the second last one (row five) and the optimal model is found on row four. The last row of the table (ignoring the VIF test line) is not derived by this method, it is the final model chosen and does not have the lowest BIC value. It does however have the smallest \( \sqrt{s^2} \) when tested which was the criteria setup to evaluate the choice of covariates 4.4.2. One indication that the BIC iteration would not produce the best model is the fact that both R^2 and Adjusted R^2 increases when the volume covariate is kept, suggesting it should stay, while BIC point in the other direction. Nevertheless, some covariates could indisputably removed such as: Price change, Friday and OMXS30. Consequently no VIF test was made for these, only covaraites making it to the final model are tested. As can be seen in the last row of the table no values are above the limit of 10 3.2.6 and hence multicollinearity is not a large factor. Why there are only two time covariates used in the final model is explained in the discussion.

Not all of the variation of covariates that was to be tested by 4.4.3 are shown in the table. For example only "volume change as percent" is shown and not "volume change as a dummy". By measuring the p-value of the variations the best fitted "representation" was chosen. The result from this was the covariates that are shown in the table.

5.2.1 Evaluation of the Linear Regression

Table 5.2 shows a summary of the results from the daily based linear regression in its final form. The Estimates are the \( \beta \)'s calculated in the regression. The std. error is the standard error from the hetroscedastic model. Eta.sq is the \( \eta^2 \) were the covariate in question is analysed and finally p.value is the p-value for the same covariate.

None of the covariates are classified as having a large or even medium impact by the \( \eta^2 \). The p-values do however indicated that all covariates except for volume...
Estimate | Std.Error | Eta.sq | P.value
--- | --- | --- | ---
Intercept | 2732606.28 | 210284.65 | 0.30022 | 0.0000
SD in price | 3262187.56 | 623854.01 | 0.09038 | 0.0000
Volume change % | 181071.72 | 180484.54 | 0.00227 | 0.3162
If Monday | -757416.30 | 159170.99 | 0.03435 | 0.0000
Oct 2014 - April 2014 | -444243.44 | 158330.58 | 0.01085 | 0.0052
April 2015 - Oct 2014 | 638142.86 | 171444.00 | 0.02341 | 0.0002

Table 5.2. Summary of linear regression of the final model.

up in percent have a significant impact on the total volume traded. The reasoning
behind keeping non-significant covariates in the model is explained in 4.4.2. And
the choice of only two time covariates is explained in 6.1.

### 5.2.2 Distribution of residuals

The residuals of a hetsoscedastic linear regression with all data and the final list
of covariates are Q-Q plotted against the normal distribution. This is to test the
assumption described in 3.2. As can be seen, the end quantiles fit less than optimal
while the middle ones do. This indicates that the residuals are not perfectly nor-
manly distributed. For the context of this study, and with regards to the 'results
over elegance' argument, it is viewed as a sufficiently good fit.

![Normal Q–Q Plot](image)

Figure 5.2. Residuals Q–Q plotted against the normal distributions
5.3 Predictions for a Three Month Period

The following results in Figure 5.3 are gathered by using the method described in 4.4.2. It was found that using five days as a base for the floating mean was roughly optimal for order book 1. While this differ between order books it was used on all in order to achieve consistency. Additionally the difference with ± a day was so small it was decided that it would not be investigated further. The number of days tested was: all, 200, 100, 50, 20, and 10 to 1. The "mean error in percent" is defined as the mean of the absolute values of the difference of the true values and the predicted values divided by the true value. Since there was a problem with collecting data for the same dates with order book 3 the 60 days interval closest to the other two was used.

<table>
<thead>
<tr>
<th>Dates</th>
<th>Order Book 1</th>
<th>Order Book 2</th>
<th>Order Book 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015/01/02 - 2015/04/01</td>
<td>0.25430404</td>
<td>0.27730529</td>
<td>0.26995142</td>
</tr>
<tr>
<td>60</td>
<td>0.2252525.8</td>
<td>1961891.8</td>
<td>2307946.2</td>
</tr>
<tr>
<td>Number of days</td>
<td>2049609.5</td>
<td>2147682.1</td>
<td>2121131.5</td>
</tr>
</tbody>
</table>

Table 5.3. Results from predicting the volume with order book 1 to 3 for 60 days

Note that in order book one and three the prediction outperforms the floating mean. Observing the $s (\sqrt{s^2})$ of the prediction on all three they seem to be stable across in opposite to the mean prediction which is substantially lower on order book two. The prediction itself is thus not less accurate on the second one, the floating mean is more accurate.

5.4 Intraday Prediction

Results in this section are all gathered using the method described in 4.4.4. The following table (5.4) is a snapshot of the heteroscedastic linear regression at 60 minutes intraday data and at the last day of prediction. For reference the $R^2$ of this regression is 0.57 which is substantially better than the 0.22 in the initial regression.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>Eta sq</th>
<th>P.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.18308312</td>
<td>0.062398148</td>
<td>0.09478</td>
</tr>
<tr>
<td>Diff start</td>
<td>2.50100187</td>
<td>0.287767986</td>
<td>0.50360</td>
</tr>
<tr>
<td>SD so far</td>
<td>-0.30556463</td>
<td>0.223767116</td>
<td>0.01930</td>
</tr>
<tr>
<td>Price diff</td>
<td>-4.05482660</td>
<td>2.486505298</td>
<td>0.02942</td>
</tr>
</tbody>
</table>

Table 5.4. Summary of linear regression results for 60 min of data
Presented in Figure 5.5 are results retrieved with three different quantities of intraday minutes from a period of 60 days. As mentioned in 4.4.4 the initial prediction needed to be expanded. The values of \( s \) of the first prediction varied greatly when different number of days were tested. This is due to the model preforming better at certain times, in particular the last months of 2014. This makes it difficult to determine whether a certain number of days are optimal or if it is due to initial predictions improving. The final choice was 50 days since it gave roughly optimal results, and for explained reasons, fine tuning this number was neglected. Tested numbers include (100, 50, 30, 20, 10).

<table>
<thead>
<tr>
<th>Minutes of Intraday Data</th>
<th>2 minutes of data</th>
<th>60 minutes of data</th>
<th>120 minutes of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s ) without new information</td>
<td>2127121.5</td>
<td>2127121.5</td>
<td>2127121.5</td>
</tr>
<tr>
<td>( s ) with new information</td>
<td>1610030.3</td>
<td>1189298.7</td>
<td>1026366.6</td>
</tr>
<tr>
<td>Correlation without new information</td>
<td>0.44578619</td>
<td>0.44578619</td>
<td>0.44578619</td>
</tr>
<tr>
<td>Correlation with new information</td>
<td>0.74624139</td>
<td>0.84936415</td>
<td>0.89157567</td>
</tr>
</tbody>
</table>

Table 5.5. Comparison when predicting with different intraday information

The correlation presented in the table is the correlation between the observed value and the predicted one and \( s \) is described in 4.4.2. Both the correlation and the \( s \) improves as time increases. Notably the effect of 2 minutes improves the model proportionally more than 60 or 120 minutes, this is further discussed in 6.3.

For frame of reference, here is a table presenting the number of times (in percent) the prediction gave a value that was closer to the true value then the value produced by the floating mean.

<table>
<thead>
<tr>
<th>Minutes of Intraday Data</th>
<th>0</th>
<th>2</th>
<th>60</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of prediction better than floating mean</td>
<td>0.60</td>
<td>0.65</td>
<td>0.72</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 5.6. Showing the "win rate" with different amounts of intraday data
5.5 Illustrations and Examples

In figure Figure 5.3 the black bars are the true total volume, red is the prediction with initial linear regression, blue the prediction using the five day floating mean and green the prediction with 60 minutes of intraday data.

![Figure 5.3. Comparison between observed value and predictions, blue = mean, red = initial prediction, green = prediction with 60 minutes of data](image)

Figure 5.3 only demonstrates the prediction on the total volume and does not take into account the intraday prediction. The two graphs Figure 5.4 and Figure 5.5 illustrates the trades during a day and the prediction of it using the initial prediction (blue) and the intraday one (black), the red line is the actual volume traded traded. Figure 5.4 is taken from a typical day. Figure 5.5 is taken from a day where some news hit the market the day before (see the deviation in the graph). The correlation presented in the graphs are the correlation between the intraday values and the true values after the point in time used in the intraday prediction. Consequently, while the graph illustrates the entire day only information after the time specified is of use.
At 60 minutes (0.1 on the x-axis) there is a difference between the prediction and the true value. This is rectified by the intraday prediction.

While neither prediction matches the true value the intraday matches it better. The slope after 60 minutes are similar for the intraday prediction and the true value, indicating a good prediction. The cumulative distribution differs due to the trades made in the beginning of the day.
5.5. ILLUSTRATIONS AND EXAMPLES

*Figure 5.6* is an example of a weakness in the model, only information at time $t$ is used and not the information leading up to it. Using 100 minutes of intraday data the method is not able to detect the two contradicting patterns. Namely, after the initial increase there was a steady decrease which can be observed with all the data up to 100 minutes, however if only the data at the 100 minutes is observed only an increase is observed. This is discussed in further in 7.2.

*Figure 5.6.* illustrating the problems with not using the data up to 100 minutes and just using data at 100 minutes

*Figure 5.7.* shows the marginal gain of intraday information. The trade-off that comes with the amount of intraday data used and the volume that remains will be discussed further in 6.3.

*Figure 5.7.* x-axis is minutes of the trade day, red = % of volume to trade left, blue = correlation between prediction and outcome, orange = $s$ from prediction (divided by the largest value for scale), black = mean error in percent. Note that the value at minute one is the same as minute 0 (see 4.4.4)
Chapter 6

Discussion

This chapter will discuss the results found and their implications. The focus of discussion will be the value of the final model and its practical use.

6.1 Initial Regression

The chosen regression to model the volume trade per day is:

\[
\text{Volume traded} = \beta_0 + \beta_1 (\text{SD of price yesterday}) + \beta_2 (\text{Dummy for Monday}) + \\
+ \beta_3 (\text{Percentage change in volume yesterday}) + \beta_4 (\text{April 2015 - Oct 2014}) \\
+ \beta_5 (\text{Oct 2014 - April 2014})
\]

Here the intercept, \(\beta_0\), is all dates and days in the data set that is not included in a dummy.

Based on the results, some of the assumptions in 4.4.3 needed to be withdrawn and rethought while others were to some extent confirmed. The "price change" covariate was indeed an inferior choice to the "volume change", although the assumed relation between the two was unwarranted since the VIF test showed no multicollinearity (9.1.1). Additionally it was shown that "volume change in percent" better reflected the mindset of an investor than a dummy covariate for volume change. This is also true for the OMXS30 covariate, but even so, all variations of the covariate were omitted from the final model.

The covariate that showed the most promising results was the "standard deviation of price". And as shown in 5.2 the assumption of positive correlation was indeed true. By the framework established in 4.4.3 this indicates that "unexpected" change in volume is the main effect on the volume. However, with the small impact made by the covariate this is not to be taken as a truth.

Contrary to the assumption the Monday covariate yields a significant negative value. One explanation of this could be that the accumulation of knowledge during the weekend actually stabilizes more than it makes the market volatile. Either way the p-value indicates it is not a random pattern and thus it was kept in the model.
CHAPTER 6. DISCUSSION

The Friday covariate on the other hand did not yield any significant result, it would seem that Fridays are treated by investors the same way "any other day" of the week is.

A difference in volume traded between different periods is observed in 9.1.2 which at first glance would indicate that using some form of time covariate is warranted. Yet, if the change in mean volume is already explained by the model, time covariates would not be necessary. With the large difference between the time periods this was assumed not to cause problems, and the results in 5.1 support the usage of time covariates. Interestingly, while the model by the linear regression measurements used in this study profit from the use of time covariates the prediction power decreases when all time covariates (except for one as an intercept) are present. The final model only contains two "time covariates" which represents "April 2015 - October 2014" and "October 2014 - April 2014" thus "April 2014 - November 2012" is the intercept. Why would this model be preferred to the model with six month time covariates consistently? One explanation could be that the prediction power for the entire data set increases with the use of consistent time covariates while the prediction for 2015 does not gain from it. The reason for this can only be speculated about and random chance cannot be excluded.

6.2 Prediction Power

The results of order book two in table 5.3 indicates that when volatility is low (relative to the other order books) the prediction power stay the same while the floating mean prediction naturally improves. This suggest that the model is best fitted for volatile stocks. This is to be expected since an increase in volatility by definition lower the level of accuracy of the floating mean prediction but not necessarily the prediction with linear regression, given that the covariates are able to predict some of the volatility.

It is clear from the results that no combination of covariates used in this study, explains the volume traded efficiently. Still, the data set of about 600 days should be sufficient in rejecting any random patterns, which is reflected by the low p-values, 5.2. All covariates are below any standard limit of p-value limitation, except for the volume one which is kept by the "results over elegance" argument. It is thus concluded that covariates are not randomly increasing prediction power. Although, with none of the covariates above "small impact" the actual prediction power of them needs to be discussed. One possible explanation of low p-values and low impact is that the covariates do effect the trade, just not primarily in the day to day trading. In other words, price volatility of market may effect the trader substantially during the same day it is measured, or perhaps the last few hours of yesterday effect the first few of today. The assumption in 4.4.3 of trends lasting during at least one day seem hard to justify in the result. Although, there is no guarantee of another way to model the effect would increase the impact. It is not unlikely that the impact of the quantitative covariates are simply not causing as substantial impact. By this
argument either some relevant quantitative covariates are missing or qualitative covariates are. While the first cannot be excluded we argue that the latter is most likely since much of the trade is based on news and events that occur the same day, which is, at least in its "true" form, qualitative. The question is then if these qualitative covariates are quantifiable in any way? In the model used in this study this would be a difficult task for two main reasons. Firstly quantifying news are being done but is still in the early stages (See part II). Secondly, the 'news of today' needs to be predicted the day before, since the model is based on historical data with one day resolution. Trying to predict news or events is in itself paradoxical, because by doing so it would be common knowledge beforehand, and therefore not by definition not news. With this in mind, the prediction power of the initial model is not necessarily low, rather the unexplainable part of the volume is high.

6.3 Adjusting With Intradays Data

Arguably the most important result in this study is the ability of intraday data to adjust the initial prediction, and in doing so increasing the prediction power of the entire model by a huge margin. A perfect illustration of this is given in Table 5.4. With this expansion of the model there comes a clear trade-off between how good a prediction you get and the available volume left within the day. Under the assumption that a person using this model is willing to trade at any time during the day when should they start? While it is true that more information increases the accuracy of the prediction, more information also leaves less volume to be traded. Theoretically there should be a point where, depending on the volume you want to trade, you have the optimal amount of information to improve the prediction while the shares left to trade are sufficiently many. In Figure 5.7 this trade-off is illustrated. As the red line (volume) decreases the blue (correlation) increases and orange (s) decreases. When comparing the results from the expansion of the model to the original one the usefulness of the first regression at first glance look a bit questionable. In 4.4.2 the argument was made that if the floating mean outperformed the linear regression it should be used as the prediction in the intraday model. This proved to be an unjust modification since using it actually decrease the prediction power of the intraday model, even when it was substantial better in the initial prediction. As can be seen in Figure 9.1 the first few minutes are more accurate using the floating mean but after 7 minutes the linear regression actually preforms better. The reason for this could be that the linear regression model makes systematic errors. The intraday prediction "corrects" the initial prediction and thus systematical errors would be picked up by the model and rectified. Since floating mean do not make any systematical errors finding a "standard correction" is then more difficult. Moreover this suggest that there are at least one missing covariate (which was already known). However, with the small difference between the floating mean and linear regression prediction the impact of this covariate, if used, should also be small.
Chapter 7

Conclusion

This chapter will determine the practical use of the model along with the most important findings of the study and their implications. Additionally thoughts on suggestions for further research will be discussed.

7.1 Proof of Concept

The purpose of this study was to explore and determine a mathematical foundation for intraday volume distribution on highly liquid order books in order to predict the volume on specific order books. The study set out to provide a proof of concept due to the limited amount on publications on the subject. A potential result could have been that there was no concept to prove, due to data not showing patterns and predictions being unreliable. However that was not the case. Thanks to the realization that just a small amount of intraday news were able to increase the prediction power by a large margin, the results in this study has apparent real world applications for AT in general.

As illustrated in (5.4.2) the volume traded during a day is volatile and predicting with yesterday’s data gives inaccurate results. However, when assessing the usability of the model exactly what level of inaccuracy is tolerated is a decision conditioned on the usage. It is outside the scope of this study to examine what level is tolerated. Yet, under the assumption that there are currently no other way to predict the volume, this model would per definition be usable since it outperforms using the floating mean. Even when no data during the day is known (which is true for two out of three order book examined).

On the other hand it would not be wise to rely solely on this model, if news that was not known the day before are conveyed then the model should be adjusted accordingly. While this could technically be done by machine (see evaluation of paradigm) it would realistically be done by hand. In other words: if one obtains knowledge that potentially changes the volume an estimation of that change should be inserted in the current prediction. This could arguably also be done without knowing the cause of a deviation, but merely observing it, as made in the intraday
prediction. Consequently the original prediction would to some extent be used as a 
benchmark, if the volume traded approximates the prediction keep to the model, if 
not, adjust accordingly.

7.2 Suggestions for further research

The findings in this study leaves many questions unanswered. Mainly, only one 
way of predicting was examined, linear regression, are there others which could 
be effective? In the start-up process of this study other ways of predicting was 
evaluated as well. For example time series forecasting could be used to predict.

Further research within the current model is also a possibility. This would in-
clude searching and changing covariates to increase the predictability. For example, 
using more qualitative covariates such as 'number of news article published about 
company A' or 'number of google for company A' could improve predictions.

Additionally using the standard Gauss Newton algorithm on mixture distribu-
tion is not recommended [25]. This proved problematic when trying to match the 
mixture beta distributions to a subset of data from a random day. This was tested 
since it could increase the correlation between the model distribution and the ac-
tual one by using only intraday data in the modeled volume distribution. Replacing 
the standard Gauss Newton with a hybrid Expectation maximizing/Gauss Newton 
algorithm could thus allow the subset to be smaller by converging faster.

Figure 5.7 illustrates a problem with the intraday model. The current model 
does not use the information leading up to the current time, only the information at 
the specific time. Improving the model could be done by savoring the information 
gathered up to the current time. One way of doing this is using the previous 
response variables as covariates at time $t$. In other words, using a time series. 
The calculations required would increase and given that minute resolution is not a 
requirement, this model would require thought and careful consideration on what 
level of resolution should be chosen. The calculations need to be fast enough so 
that none (or in reality very little) new information is received whilst calculating 
the prediction. Exploring this trade-off is important when improving the model by 
a time series analysis.

7.2.1 Next Step

Seeing as a stock can be traded on many different markets, a natural next step 
would be to take the same analysis done in this study and expand it to all the 
relevant markets. This expansion would result in a view of how the volume is 
distributed for the entire stock, and thus more relevant for real world applications.
Part II

Algorithmic Trading as a Technological Paradigm
Chapter 8
Algorithmic Trading Paradigm

8.1 Introduction

Technology has revolutionized the way we trade financial instruments. The stereotypical picture of a loud trading floor, influenced by the 1987 movie "Wall Street" and others like it, is now a thing of the past [26]. Times change, what was once done by hand is now digital. Buy pads, sell pads and cancellation pads are now replaced with iPads.

Albert Menkveld, a professor of finance, mentions people interrupting his lectures and asking him “but don’t they [HFT firms] just steal our money?”[8]. Inspired by this we set out to answer the questions of how and why did we got here, despite the negative views? What effect does AT have? and what drove this change? In this chapter these questions will be addressed and the technological paradigm of automated trading will be analyzed and evaluated.

8.2 Methodology

8.2.1 Literary Study

Collecting data for the analysis was in three ways, reading scientific publications, reading news articles and semi-structured interviews with experts on the subject. All used publications were found by searching on Google, Google Scholar and the KTH library. The search words that were used are presented in the table below

<table>
<thead>
<tr>
<th>Literature study</th>
<th>Search words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithmic trading paradigm</td>
<td>&quot;Algorithmic trading (+paradigm)&quot;, &quot;Automated trading (+paradigm)&quot;, &quot;AT impact&quot;, &quot;HFT&quot;, &quot;HFT discussion&quot;, &quot;AT media&quot;, &quot;Industrial dynamics (+trading)&quot;, &quot;Technological advances&quot;, &quot;Cumulative-ness&quot;, &quot;HFT volatility&quot;, &quot;Patents security&quot;</td>
</tr>
</tbody>
</table>
CHAPTER 8. ALGORITHMIC TRADING PARADIGM

To filter between the scientific publication found by the literature study, abstracts of publication with relevant titles were read and from there a decision on their relevance i.e should we keep reading, was decided. Publications representing different views on AT, namely, "good" or "bad" were chosen to represent opinions while publications presenting AT as a concept and not its implications were used to understand the subject.

Furthermore knowledge obtained from the following courses are used ME1313 Industrial and Technical Transformation and ME1312 Understanding the Interface of Technology and Business.

8.2.2 Interviews

For each interview an interviewer and a secretary is decided. The interviewer’s role is to ask the questions and maintain the flow of the interview. The secretary takes notes and keeps track of questions asked and answered so that none are excluded. In addition to this the sound from the interviews is recorded.

Following the guide presented in "Other Literature" (part I) semi structured interviews were conducted. In comparison with structured interviews that has strict rules for the questions, semi-structured interviews are more open. The interviewer has general themes and questions prepared but is ready to adapt the follow-up questions to the interviewee. Semi structured was preferred to structured since the intention of the interviews was not to derive results empirically, rather understanding opinions and views within the industry. The semi structured approach works well for comparing different views, while not losing information by not allowing “sidetracks”. Scope creep is a potential problem with semi structured interviews and to actively maintain set topics the questions used was constructed to be precise yet not excluding. This is a difficult task and some questions are bound to be less than optimal, to remedy this the interviewer will be tasked with maintaining what is said relevant to the subject.

Interviewees

- **Henrik Talborn**: Studied Physics and Economics & Management at Lund University. He has previously worked at different investment banks with producing mathematical models for statistical arbitrage. Currently employed by Swedish House of Finance.

- **Rolf Anderson**: Studied economics at Stockholm School of Economics. Help setup one of the worlds first electronically stock exchanges at the Stockholm stock exchange. Currently CEO at Pantor Engineering.

- **Lukas Magnusson**: Studied Product Design with an Industrial Management and Engineering master at the Royal Institute of Technology. Did an internship at Credit Swiss in London after graduation and is currently employed by Pantor Engineering.
8.3 Dosi’s 4 Dimensions of a Technological Paradigm

From our literature studies Dosi’s model for describing a technological paradigm was chosen as a base for our analysis (ME1312). The following section describes AT through the eyes of this model, with its four dimension. The interpretation of each dimension in the context of this study is described in each subsection.

8.3.1 First Dimension, Technological Opportunities

What advantages are there in the identified technological paradigm compared to the paradigm it is replacing?

An important concept in quantitative methods is the ability to backtest them on historical data. These tests allow the practitioner to assess the usability of a specific algo/trading plan. The nature of computers calls for strict rules of action. A computer cannot postulate its own theories and draw its own conclusions and must therefore be told exactly what to do given any specific situation. Backtesting allows for testing that there are no loopholes in these strict rules. The computational speed is now at a point that it to human eyes seems instantaneous. This means that a faulty line of code quickly can become costly. Two examples of this is the famous flash crash and what happened to Knight Capital in back in 2012, as mentioned in 1.1.

Modern trading boasts an exponential increase in speed that allows for huge amounts of decisions every second, being able to accept any order almost at the instant some criteria are met and running many different trading planes simultaneously. But even as speed increases, the consistency is rising in unison. Consistency is in general a good indicator of a skillful trader. But even for skillful traders, the losses sometimes stacks up. Prospect theory tells that a loss generally has a greater impact on people then a win, this might cause the trader to doubt his plan of action (ME1312 Understanding the Interface of Technology and Business, Prospect Theory). So regardless of how good the plan was, it is easy to overcompensate for a temporary losing streak because of the fear of more losses. Losing money is generally, and quite obviously, a traumatizing experience. An automated system would then help in realizing the plans full potential and thus increasing the consistency. A good example of how AT increases consistency is the HFT firm Virtu Financial. During its five years in the business, the firm was profitable 1’277 out of 1’278 days.

[5]

Lowering the impact of emotions is a natural effect of the automated trading systems. The "hands of" approach, and post plan evaluation of methods encourage a rational approach to lost and won trades. As trades are executed automatically,
hesitation is removed from the equation and replaced with computing speed. During volatile market conditions the automatic nature also preserve trading discipline. Fear and general greed poses a constant threat and is hard to find the balance between.

8.3.2 Second Dimension, Appropriability of Innovations

*How can innovation within the technological paradigm be monetized?*

The evaluation of appropriability in the industry will be done for two identified key parts: users and providers. The user category includes AT firms as well as potential clients using the technology provided by these. Providers include companies and markets like NASDAQ, Burgundy and the London Stock Exchange. Innovations regarding technology speed is selected to be evaluated for both user and providers. Since findings in the literature indicates it is viewed as the most relevant field of innovations in the industry.

Providers profit from innovations which decrease latency and by then not automatically providing all users with it. Namely, they profit from selling physical presence and shorter distance to the market, which then lowers latency. There are several layers of latency with each layer increasing in price [8]. In other words: a users with a faster algorithm does not necessarily "win" the trade if they are not provided with the fastest connection to the market. A natural consequence is then that the recent development of increased competition in the industry favors the appropriability of providers.

Users on the other hand do not stand to benefit from the current development. Eric B. Budish et.al describes it as an “arms race” between HFT firms

"...competition has not affected the size or frequency of the arbitrage opportunities, it has only raised the bar for how fast one has to be to capture them." [27]

The amount of arbitrage opportunities stay the same while the cost operating in HFT increases, thus lowering profits. The same article argues that introducing an discrete trade would rid the arms race and benefit society as a whole. Albert Menkveld argues that a difficulty in implementing a discrete trade is that no single user benefits from entering such a market while others stay. Thus even if a provider who offers discrete trading enters the industry, and even if users stand to gain if all change provider, they may not switch [8]. Consequently this game theory logic directly effects the appropability of innovations relating to the discretization of the market.

8.3.3 Third Dimension, Cumulativeness of technical advances

*To what extent is continuing the technical advances possible?*
8.3. DOSI’S 4 DIMENSIONS OF A TECHNOLOGICAL PARADIGM

From 1929 to 2009 the total market capitalization by the US stock markets have doubled each decade [29]. In Sweden the market capitalization of listed companies (as % of GDP) has gone from 49% in 1988 to 103% as of 2012. (latest data given by source) [30] In unison with the increased market capitalization we have also, since the introduction of computers, seen an exponential increase of computing power. These two developments form the soil from which algorithmic trading has grown and blossomed. The possibility of monetary gain is a force to be reckoned with, especially when in the hands of highly educated people.

During 1929 to 2009 the volume on the Dow Jones has doubled about each 7.5 years. But something has changed. The last decade, as of 2013, the phase quickened to the point that the volume is doubling every 2.9 years [29].

AT is not an exception to the cumulative nature of the technological advances within a paradigm that Dosi describes, and are closely linked to the fact that the innovations of today forms the foundation for the innovations of tomorrow. Two trajectories of innovation within the paradigm will be analysed below.

**Speed**

The speed of light can’t be changed. Every other aspect of the modern trading system is however up for improvements. Speed is an important factor for two main reasons:

1. You want all your information to be as up to date with actual data, within the physical limitations set up by distance. (e.i the “true” price in London takes a while to travel to Stockholm, when it finally gets there it might not be “true” anymore).

2. You want to be able to react as close to instant as possible to developments in the market. Examples of this is specially designed switches which were not economically viably to research and produce until the entrance of AT (Lukas Magnusson).

**Models**

Increased competition has eaten the low hanging fruits of arbitrage and “free” basis points for HFT-firms. To compensate for this, the complexity of the models naturally had to increase.

In summary it is concluded that a general “one-upping” is the driving force behind the cumulative advances within the paradigm, and a huge monetary gain stands to be gained from the one currently at the top.

**8.3.4 Forth Dimension, Properties of the knowledge base**

*How is knowledge and information managed in the technological paradigm?*
CHAPTER 8. ALGORITHMIC TRADING PARADIGM

With competition high any edge towards competitors is cherished and well guarded. Consequently these edges are not shared with the public either. Presumably the monetary incitements driving the secrecy also drive innovation and development. Yet, less secrecy would facilitate innovation by enabling combinations of technology. This section uses experience from other technological fields to answer: is secrecy in the AT industry ultimately beneficial to the innovation process? To answer this the relation between patents and secrecy will be examined [31]. Patenting is closely related to secrecy, one crucial difference being that the transparency increases with patenting. Within the industrial dynamic framework synergies between patents and secrecy are well discussed, companies often use one or both to protect innovation made. Wesley M Cohen et.al argues that protecting intellectual property is increasingly being done not by patents but secrecy. In fact, he argues that patents are used not only to prevent copying but also to hinder rivals from similar innovations (patent blocking). Additionally the historical evidents presented in Petra Moser’s study suggest secrecy is a tool used when patent laws are insufficient and that it is ultimately preferred [32]. The AT industry are thus not alone on preferring secrecy as a protection of intellectual properties.

In the literature studies made in part I finding previous studies on the subject was difficult. This lack of published work supports the claim of AT being secretive. Shortage of information might damage the public view of the industry. By not providing the tools to learn and understand transparency is lost and historically a lack of transparency is not preferable if trust is to be gained [31] [35] [33]. In alignment with game theory no single firm stand to gain from increasing their transparency, change in the close future is unlikely unless regulations are enforced.

8.4 Discussion

This sections aims to clarify the different views of both the technological trajectory of the paradigm and the paradigm itself.

8.4.1 Driving Factors of Automated Trading

What caused the markets to become digital and thus allowing the entrance of AT? Presented in this section is our understating of the interviewees views on these driving factors. One shared opinion is that fairness played an important role in the development. Rolf Andersson described it as

Those with psychical presence at the market received an edge to those who got the information second hand. Moreover even if one wanted be physically present at the market there were a limited number of spots, all in all, an unfair system.

With an electronic market the incitement to trade would increase due to it being more fair. This is profitable for markets but not for those who earlier had an
edge in trading, the stockbrokers. Not surprisingly markets started to increase their capability of electric trade, to the dislike of the brokers. Electronic markets as a technological paradigm is accepted today but it was not at the time of it’s introduction, as with most paradigm shifts it faced resistance in the early stages (ME1312 Understanding the Interface of Technology and Business). And not unlike other paradigms, the cumulativeness of electronic markets paved the way for new paradigm, the AT paradigm. And while electronic markets are widely accepted as something contribution to society, AT is not, at least not yet.

8.4.2 Implications of Speed

In 8.3.1 the lack of emotions and “speed increase” attributed to the technological change were discussed. This section aim to discuss the opinions on implications of said change.

AT is often described as a “Black box”, what is produced is known but not how or why. Events like the Flash crash [33] are used to show problem of not knowing exactly what is transpiring and problem with the speed of the trades. In the event of a computer making faulty trades a human must stop the AT system by hand (by definition, since the computer cannot recognize faulty trades and only "follows orders"). And by the time the problem is realised, significant damage might already be done. This scenario postulates AT making faulty trades and some claim that events like the flash crash cannot be attributed to AT as a whole, that AT is a tool and there is in fact no black box mentality present. Henrik Talborn at the Swedish House of Finance described it as:

“...A black box for those who do not use it perhaps, not for those building and using them.”

Another discussion relevant to the speed of AT is its fairness. While all information is distributed at the same time, it is not received at the same time by all users. This is due to the previously mentioned natural latency caused by a psychical length difference between the users and the providers. One argument for the system being fair is that with the globalisation all AT firms should be able to choose their location. Thus, any locational advantage could be recreated by a competitor and the same is true for latency. Rolf Andersson argues that as long as there are no limit on participant in each layer of latency then the system is ultimately fair. On the other hand this would further fuel the arms race described in 8.3.2. The argument of fairness is thus often tightly linked with opinions regarding the gain from the arms race.

8.4.3 Describing Reality

Human traders employ a different logic to historical data then computers do. They select what is perceived as relevant to the current situation as basis from which to act. Computers on the other hand view all data available to be a basis and use a
set of rules to determine implications of this. The models used by successful human trades, such as gut feelings or reading between the lines are not not yet integrateable in computer decision-making and while some claim the future of AT is improving computers perception of reality a gap between computer models and human traders still remains. Computers are incapable of making long term decisions. Humans still decide what to purchase or sell for long term profit while computers executes the order in an efficient way.

Why then is this not integrateable? One explanation is the human psychology in trade and its natural unpredictability, earlier discussed in \[4.3.1\]. Another is the unpredictability in nature such as the weather and earthquakes. Thus, even if all relevant information prior to the current day was known, and given that a computer could store all this information, a model predicting today perfectly could still not be made. Since the model would still be based on historical data with no guarantee of the data relevance today.

8.4.4 HFT and Volatility

By some of those advocating regulation or removal of HFT, the rise of this trade is thought to increase the volatility of the market, a thought that is disputed by the advocates of HFT. Lord Myners, former Financial Services Secretary in the United Kingdom told the guardian that:

"High-frequency trading appears so detached from the true function of capital markets, but is potentially fraught with hazard. It definitely deserves more attention than either the Financial Services Authority (FSA) or the Treasury has given it" \[39\]

The evidence is thin on both sides of the argument due to the lack of historical data to base the evidence on. Nevertheless some studies show a link in volatility and the introduction of HFT \[36\] \[37\]. While others argue the opposite, for example FIA Principal Traders Group \[38\]. One argument presented for HFT not increasing the volatility is that HFT stand to gain from an increase, and thus it would then gain from its own impact and essentially be self sustaining (which is unheard of in macroeconomics). While this argument is, when presented alone, convincing, the counterargument would be that HFT does not by default increase volatility rather an small increase can be observed the last few years and only HFT firms stand to gain from it. From our litterateur study this seem to be a frequent dispute between the experts with no real consensus as to what is true.

The European Union are planing to updated the rules for markets in financial instrument (MiFID II) To what extent and what effect this will have is to early to predict. Lukas Magnusson describes it as:

"The market seldom reacts the way that was planed when regulations were made, the outcome of a new set of rules is often that the market starts playing a different game."
8.5. CONCLUSION

What effect the update will have is thus unclear but according to Lukas one reason for this update is new constraint for HFT operations.

8.5 Conclusion

The AT paradigm was a natural next step after the introduction of electrical markets. The largest driving factor behind the paradigm shift from "floor trading" to electrical markets was fairness, which ultimately was profitable to both markets and most traders. While traders who at the time held an edge in the earlier paradigm disliked that change few people today would claim electronically markets are something that is not contributing to society as a whole.

The appropriability of HFT has surpassed its peak. Thus a more mature market is to be expected. The threshold for new entrants is steadily increasing due to a cost increase of machinery and for perks at the providers of the market. The increasing competition for the same basis point also makes the industry less attractive for new entrants while decreasing the margins of the existing firms. That said, if competitors were to leave, margins would increase. As of now, the market has a finite amount of basis points for HFT firms to share and this would only change by (1) a change in technology or (2) some regulation of the markets. While there are regulations planed, they are not intended to increase the amount of HFT, rather the opposite. A radical change in technology is always hard to predict, what can be said is that the speed of light is set and that technology is very close to this limit.

The future of AT in general is less clear. With no clear signs of radical change the model presented in part I of this study should stay relevant in the near future. As long as HFT firms are present in the market the need for smart algorithms executing pre-programed trades will be needed to minimize the damage done.
9.1 Extra Results and figures

Results and figures that were omitted from main text.

9.1.1 Extra VIF test

To test if the assumption that price change yesterday should be replaced with volume change yesterday a VIF test was constructed. With price change yesterday as the response variable and volume change yesterday as the sole covariate. The value obtained was 1.0002235. With all covariates in the final model present as covariates the VIF value was 1.0055158. Both showing no signs of multicollinearity.

9.1.2 Mean volume traded over time

With six month covariates counting backwards from April 2015 the following mean total volume traded per day was obtained:

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean Volume traded per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-2014</td>
<td>4511617</td>
</tr>
<tr>
<td>2014</td>
<td>2957051.7</td>
</tr>
<tr>
<td>2014-2013</td>
<td>4070176.4</td>
</tr>
<tr>
<td>2013</td>
<td>4254545.6</td>
</tr>
<tr>
<td>2013-2012</td>
<td>2667455.9</td>
</tr>
<tr>
<td>2012</td>
<td>2718644.7</td>
</tr>
</tbody>
</table>
9.1.3 Floating Mean vs Linear Regression on Intraday Data

![Graph showing variance over minutes for linear regression and floating mean](image)

**Figure 9.1.** Using result from order book two. Black line is the using the prediction by linear regression as the benchmark and the red one is using the floating mean.

9.1.4 Results from Order Book Two and Tree

Results and figures that were omitted form main text. The graphs corresponds to Figure 5.7 and Figure 5.9

![Graph showing marginal gain of measurements](image)

**Figure 9.2.** results from order book two
9.1. EXTRA RESULTS AND FIGURES

Figure 9.3. Results from order book two

Figure 9.4. Results from order book three

Figure 9.5. Results from order book three
Bibliography


[8] Lecture at Stockholm Business School on 15 April 2015 by Björn Hagströmer and Albert Menkveld

BIBLIOGRAPHY


63
[38] FIA Panel: *Perspectives on High-Frequency Trading and its Overall Impact on the Market* https://www.youtube.com/watch?v=0Ul7wuXrpg or https://fia.org/
