BI Norwegian Business School - campus Oslo

GRA 19502

Master Thesis

Component of continuous assessment: Thesis Master of Science
Final Master thesis – Counts 80% of total grade

What Is the Empirical Relationship Between Trading Volume and Stock Returns on Oslo Stock Exchange?

Navn:	Jan Petter Iversen, Astri Skjesol

Start:	02.03.2018 09.00
Finish:	03.09.2018 12.00

Master Thesis
BI Norwegian Business School

What Is the Empirical Relationship Between Trading Volume and Stock Returns on Oslo Stock Exchange?

Supervisor:
Associate Professor Costas Xiouros

Examination Code and Name:
GRA 19502 – Master Thesis

Programme:
Master of Science in Business – QTEM with Major in Finance

Jan Petter Iversen
Jan.P.Iversen@Gmail.Com

Astri Skjesol
Astri.Skjesol@Gmail.Com

September 2, 2018

Acknowledgements

First and foremost, we would like to thank our supervisor, Associate Professor Costas Xiouros, for his support and helpful comments in writing this thesis. Further, we would like to thank PricewaterhouseCoopers AS for granting us their master thesis scholarship. We would also like to thank the QTEM Masters Network, the faculty of finance and the library at BI for their support. Lastly we want to express our gratitude towards *Oslo Børs Informasjon* where we obtained all our data.

Oslo, September 2018

Jan Petter Iversen | Astri Skjesol

Jan Petter Iversen | **Astri Skjesol**

Abstract

In this thesis we have investigated the relationship between stock return and trading volume at the Oslo Stock exchange. Our research question was *"What is the empirical relationship between trading volume and stock returns on Oslo Stock Exchange*".

Our sample consist of daily stock return and turnover data from 1980 to 2017 for 505 stocks on Oslo Stock Exchange. Using cross-correlation analysis, multivariate regressions, GARCH and EGARCH models, and a Granger causality test we found evidence of both contemporaneous and causal relationships. Our findings lend support to the sequential information arrival hypothesis.

Keywords: Volume, turnover, return, volatility, Oslo Stock Exchange

Contents

List of Figures

List of Tables

Abbreviations

2SLS	**Two**-**S**tage **L**east **S**quare
ADF	**A**ugmented **D**ickey-**F**uller
AIC	**A**kaike **I**nformation **C**riterion
AR	**A**uto**r**egressive
ARCH	**A**uto**r**egressive **C**onditional **H**eteroskedicity
ADEX	**A**thens **De**rivatives **E**xchange
AMH	**A**daptive **M**arket **H**ypothesis
ASE	**A**thens **S**tock **E**xchange
BIC	**B**ayesian **I**nformation **C**riterion
BOVESPA	**BO**lsa de **V**alores do **E**stado **S**ão **Pa**ulo
CAPM	**C**apital **A**sset **P**ricing **M**odel
CRAN	the **C**omprehensive **R** **A**rchive **N**etwork
EGARCH	**E**xponential **GARCH**
EMH	**E**fficient **M**arket **H**ypothesis
GARCH	**G**eneral **ARCH**
GJR-GARCH	**G**losten-**J**agannathan-**R**unkle **GARCH**
GMM	**G**eneralized **M**hetod of **M**oments
HAM	**H**eterogeneous **A**gents **M**odels
HF	**H**igh **F**requency
HFT	**H**igh **F**requency **T**rading
IGARCH	**I**ntegrated **GARCH**
IPSA	**Í**ndice de **P**recio **S**electivo de **A**cciones
KOSPI	**Ko**rean composite **S**tock **P**rice **I**ndex
MDH	**M**ixed **D**istribution **H**ypothesis
NYSE	**N**ew **Y**ork **S**tock **E**xchange
OBI	**O**slo **B**ørs **I**nformasjon
OSE	**O**slo **S**tock **E**xchange
OSEBX	**O**slo **S**tock **E**xchange **B**enchmark Inde**x**
OLS	**O**rdinary **L**east **S**quares
P-P	**P**hillips-**P**erron
REH	**R**ational **E**xpectations **H**ypothesis
SARV	**S**tochastic **A**uto**r**egressive **V**olatility
SC	**S**chwarz **Criterion**
SIAH	**S**equential **I**nformation **A**rrival **H**ypothesis
SSE	**S**hanghai **S**tock **E**xchange
SZSE	**S**hen**z**hen **S**tock **E**xchange
VA	**V**olume **A**ugmented
VAR	**V**ector **A**uto**r**egressive
VPS	**V**erdi**p**apir**s**entralen
WBAG	**W**iener **B**örse **AG**

1 Introduction

Stock trading and returns has been studied for over a century and has been a central part of financial research since the late 50s. The relationship between stock return, return volatility, and trading volume specifically has been studied extensively. However, to our knowledge, there has not been conducted any recent studies regarding this on Oslo Stock Exchange (OSE). Thus, our aim is to add to the current literature on the volume-return relationship by studying the Norwegian stock exchange.

There are several reasons why the return-volume relationship is interesting. First, it is important for the understanding of the microstructure of financial markets. Volume has long been linked to the flow of information – information's role in setting security prices is one of the most fundamental research topics in finance (e.g., Brailsford, 1996, p. 90). Second, knowledge about the volume-return relation might improve short term forecasting of returns, volume, or volatility. Third, because it is often applied in technical analysis as a measure of the strength of stock price movements (e.g., Gallo & Pacini, 2000, p. 167; Abbondante, 2010, p. 287). Technical analysis is, at least to some extent, used by most fund managers – especially on shorter time horizons (e.g., Taylor & Allen, 1992; Menkhoff, 2010). And last, it has implications for theoretical and empirical asset pricing, established through its effect on liquidity (see e.g., Amihud & Mendelson, 1986; Chordia, Subrahmanyam, & Anshuman, 2001). An efficient price discovery process, associated with lower volatility, makes market prices more informative and enhance the role of the market in aggregating and conveying information through price signals (Amihud, Mendelson, & Murgia, 1990, p. 439).

With the entry of algorithmic trading, and especially high frequency trading (HFT), trading volumes has increased substantially, and the low latency makes researchers question how much information each trade carry. This makes studying the return-volume relationship especially interesting, which motivates the following research question:

"*What is the empirical relationship between trading volume and stock returns on Oslo Stock Exchange?*"

As will be detailed in section 4, there is much evidence that trading volume is related to stock returns, while standard theory – outlined in section 3 – does not necessarily predict such relations. Our goal is to understand the role of trading activity in the price formation process and understand how efficient the Norwegian stock market is.

In this thesis, we examine the empirical relationship between stock return, return volatility, and trading volume. Using cross-correlation analysis, multivariate regressions, GARCH and

EGARCH models, and Granger causality tests, we found evidence of both a contemporaneous and causal relationship, suggesting informational inefficiencies at the exchange. Our results lend support to the sequential information arrival hypothesis, and favor newer market hypotheses such as the adaptive market hypothesis and the heterogeneous agent model over the efficient market hypothesis.

The rest of this thesis is organized in the following way. Section 2 is a short introduction of Oslo Stock Exchange. Section 3 explains the most relevant theories encountered in this thesis. Section 4 surveys the current literature, and will not be specific to the Norwegian stock market as most academic literature study international and in particular U.S. markets. Section 5 explains what data we have used and our data sources, with an explaination of our data preparation. Section 6 details the methodology used and presents and discusses our findings. Section 7 concludes, while the last section offers a critical view of our thesis and suggest further research.

2 Oslo Stock Exchange

In this section, we aim to provide the reader with some context. We do a short walkthrough of Oslo Stock Exchange's history, before painting a picture of today's market. As most of our literature review in Section 4 focus on the U.S. market, we will provide some findings about the Norwegian market here.

2.1 History

Kristiania Børs – the precursor to what is today Oslo Stock Exchange – was approved by King Carl Johan in 1818. This was Norway's second exchange when it opened in April 1819 (Hodne & Grytten, 1992; Mjølhus, 2010). At that time, Norway was mainly a country of farmers and fishermen, and the capital had less than 10,000 inhabitants (Kristiania børs, 1919, p. 1). According to Oslo Stock Exchange's webpage, the exchange originally functioned as an auction house for goods, ships and ship parts, and as an exchange for foreign currencies. Back then, the currency prices were updated twice a week.

Oslo Stock Exchange introduced stocks and bonds in 1881. Although trade was modest at first, the number of securities exploded between 1891 and 1900 from 40 to 165 (Hodne & Grytten, 2000, p. 170). A few daily stock quotes were introduced in 1916, and for the entire market in 1922.

Figure 1 shows some major happenings in the history of Oslo Stock Exchange.

2.2 Current market situation

Today, Oslo Stock Exchange is the only regulated marketplace for securities trading in Norway. The exchange is moderately sized by international standards (Næs & Ødegaard, 2009, p. 4), and list the shares of 189 companies[1] with a combined market capitalization of almost 324 billion USD[2]. Further, one can trade equity certificates, Exchange Traded Products (ETPs), fixed income products and derivative products at Oslo Stock Exchange. OSE offers five marketplaces: Oslo Børs, a full stock exchange listing that complies with all EU requirements; Oslo Axess, an authorised and fully regulated marketplace; and three other markets regulated to a lesser extent. OSE is a private limited company, which it has been since 2001. The exchange use the same Millenium trading system as London Stock Exchange, Borsa Italiana, and Johannesburg Stock Exchange and is organized as a continuous electronic limit order market (Ødegaard, 2017, p. 15).

Oslo Stock Exchange is dominated by a few very large companies (Jørgensen, Skjeltorp, & Ødegaard, 2017, p. 4). As can be seen in Table 1, the four largest companies make up over 50% of the total market value of the exchange.

Company	% of market value
Statoil	23.92%
Telenor	10.85%
DNB	10.18%
Norsk Hydro	5.30%
Yara International	4.23%
Orkla	3.64%
Gjensidige Forsikring	3.18%
Aker BP	2.99%
Marine Harvest	2.80%
Schibsted	2.21%
by market value 31/12-17 |	| `https://oslobors.no`

Table 1: The ten largest domestic companies at the OSE

[1] As of the 25th of March, OSE lists 192 equity instruments – including Equity Certificates and Preferred Stocks – from 189 companies. Source: `https://oslobors.no`

[2] As of the 23rd of March, combined market capitalization is 2,510.12 billion NOK and the exchange rate is NOK 7.7527/USD. Source: `https://oslobors.no`; `https://https://www.norges-bank.no`

Oslo Stock Exchange's history

1818	King Carl Johan signed the first Stock Exchange Act
1819	Christiania Exchange opened its first office as an auction house and currency exchange with rates updated twice a week
1829	The exchange moved to its current location
1856	OSE started receiving exchange rates and commodity prices from Hamburg twice a week
1881	First listing of shares and bonds with monthly quotes
1907	Daily quotes for exchange rates introduced
1916	Daily quotes introduced for shipping and whaling shares
1922	Daily quotes for all shares
1988	Oslo Børs Informasjon (OBI) established
1988	Launched first electronic trading system – allowing for continuous trading of all securities throughout the day
1999	ASTS fully automated trading system implemented
2000	The last of the local Norwegian exchanges – Bergen Stock Exchange – is fully merged with OSE
2000	Regularly updated prices on the Internet with a 15 minutes delay
2001	OSE became a limited company, fully owned by Oslo Børs Holding ASA
2002	OSE changed to the SAXESS trading system
2003	Launched SMS service for stock exchange information
2007	Oslo Børs Holding ASA merged with VPS Holding ASA
2009	Entered a strategic partnership with the London Stock Exchange Group
2010	TradeElect trading system adapted during the period 2009-2010
2012	OSE introduced the Millennium trading platform

Figure 1: Timeline of Oslo Stock Exchange's history (dates from `https://oslobors.no`)

2.3 Earlier findings

Not much has been written about the return-volume relationship on Oslo Stock Exchange, however there are some related studies. Næs, Skjeltorp, and Ødegaard (2008) examined the relationship between the long-term development in liquidity at the exchange and the Norwegian Economy between 1980 and 2007. They state that all liquidity measures that include trading volume show improved liquidity during the sample period, and that the price level and the return volatility are determinants of liquidity (Næs et al., 2008, pp. 24). Further, they find that the development of the stock market is informative of the state of the economy as a whole (Næs et al., 2008, pp. 33). Jørgensen et al. (2017) studied an order-to-trade ratio fee introduced at the OSE in 2012, and found no impact on liquidity or trading volume, which is different from for example the Italian Stock Exchange (Friederich & Payne, 2015). Mikalsen (2014) shows several examples of volume analysis in technical trading on Oslo Stock Exchange, which at least indicates that volume is an important metric for Norwegian traders as well. Karolyi, Lee, and Van Dijk examined the commonality[3] between trading activity and return in several countries and found that for Norway, commonality was 25.4% in returns, 23.3% in liquidity, and 23.8% in turnover (2009).

Næs, Skjeltorp, and Ødegaard (2011, p. 145) found that liquidity of the Norwegian market improved over the sample period from 1980 to 2008, but also varied across sub-periods. Further, they discovered that changes in liquidity on OSE coincide with changes in the portfolio composition of investors. Specifically, before economic recessions they found a flight to quality, where some investors leave the stock market altogether and others shift their stock portfolios into larger and more liquid stocks. Mutual funds have a stronger tendency to realize the value of their portfolios in small stocks during downturns than the general financial investor (Næs et al., 2011, p. 141).

This section will be useful to have in mind going forward with the theory, literature review and methodology.

3 Theory

In this section, we aim to develop a fundamental understanding of the most prominent economic theories and hypotheses which we will later encounter. First, we will elaborate on different market hypotheses for how financial markets work and what dynamics guide the generation of stock returns. Then, we will explain different reasons investors might have for trading, as the

[3] Commonality is the co-movement between securities.

investor's trading generate trading volume, and thus their reasons govern how the volume series behave.

3.1 Market hypotheses

One of the earliest models of financial markets came from the world of gambling, which – like financial investing – also involve calculations of risk and reward (Lo, 2017, p. 17). This model is known as the *martingale*, and is based on the idea that investing in the stock market is a fair game – and thus, winnings and losses cannot be forecasted by looking at past performance. More technically

$$\{z_t\}\ is\ a\ martingale\ if\ \mathbb{E}(z_t \mid z_{t-1}, \dots, z_1) = z_{t-1}\ for\ t \geq 2$$

In 1900, the French doctoral student Louis Bachelier discovered something unusual about stock prices: they must move as if they were completely random (Fan & Yao, 2017, p. 19; Lo, 2017, p. 18). As any stock trade has a buyer and a seller who must agree on a price in order to make a trade, it has to be a fair trade. No one wants to be a fool, and there would be no agreement if one side were always biased against the other. Today, we call this theory the *random walk model* of stock prices (Lo, 2017, p. 19). Bachelier had come up with a general market theory by arguing that an investor could never profit from past price movements. A random walk is defined as

$$\{z_t\}\ is\ a\ random\ walk\ if\ z_t = \sum_{j=1}^{t} \varepsilon_j,\ where\ \{\varepsilon_t\}\ is\ independent\ white\ noise$$

Since ε is independent white noise, we have that $\mathbb{E}(\varepsilon_t \mid \varepsilon_{t-1}, \dots, \varepsilon_1) = \mathbb{E}(\varepsilon_t) = 0$. This implies that, for a random walk

$$\begin{aligned}
\mathbb{E}(z_t \mid z_{t-1}, \dots, z_1) &= \mathbb{E}(z_t \mid \varepsilon_{t-1}, \dots, \varepsilon_1) \\
&= \mathbb{E}(\varepsilon_1 + \cdots + \varepsilon_{t-1} + \varepsilon_t \mid \varepsilon_{t-1}, \dots, \varepsilon_1) \\
&= \varepsilon_1 + \cdots + \varepsilon_{t-1} \\
&= z_{t-1}
\end{aligned}$$

Thus, the random walk is a martingale (but a martingale is not necessarily a random walk).

Since the price movements of the stock market are martingales, the expected return is

$$\begin{aligned}\mathbb{E}(R_t \mid P_{t-1},\ldots,P_1) &= \frac{\mathbb{E}(P_t \mid P_{t-1},\ldots,P_1) - P_{t-1}}{P_{t-1}} \\ &= \frac{P_{t-1} - P_{t-1}}{P_{t-1}} \\ &= 0\end{aligned}$$

By the properties of martingales and random walks, our best prediction for tomorrow's stock price is today's price. Thus, our best predictions for the return is 0. This imply that there is no information about future returns in past prices. Louis Bachelier concluded that the expected profit of speculators were zero – and consistently outperforming the market would be impossible (Lo, 2017, p. 19).

This idea did not take much hold in financial literature until the 1960s, when Samuelson (1965) – using mathematical induction – showed that all the information of an asset's past price changes are bundled in the asset's present price (Lo, 2004, p. 2; Lo, 2017, p. 21). The reasoning is as follows. If investors could incorporate the possible impact of future events on asset prices today, they would have done so. Thus, future price changes could not be predicted based on any of today's information. If they could, investors would have used that information in the first place, and it would have been incorporated into today's prices. If a market is informationally efficient – that is, prices fully incorporate the expectations of all market investors – then future prices will be impossible to forecast. As a result, prices must move unpredictably (Lo, 2017, p. 21).

The same year as Samuelson's article was published, Fama – a supporter of the random walk hypothesis – coined the term *efficient market* as "*a market where there are large numbers of rational, profit maximizers actively competing, with each trying to predict future market values of individual securities, and where important current information is almost freely available to all participants*" (Fama, 1965, p.56). Fama – together with some of his colleges – soon picked up on Samuelson's ideas (see Fama & Blume, 1966; Fama, Fisher, Jensen, & Roll, 1969). In 1970 Fama formalized the Efficient Market Hypothesis (EMH). The EMH has long been the most dominant market theory. It defines financial markets as efficient, where prices fully reflect all available information and new information is incorporated quickly and correctly into security prices (Lim & Brooks, 2011, p. 69). Agents are rational economic beings, acting in their own self-interest and making decisions in an optimal fashion (Lo, 2005, p. 1).

The EMH can be classified into strong-form, semi-strong-form, and weak-form efficiency. In the strong-form efficiency, all information is incorporated into security prices, including private information. Consistently higher returns can only be obtained through taking higher risk. This means that investors cannot earn excess return by trading on information, even asymmetric –

like inside information, as it is already reflected in the prices. If investors do earn excess return, it is due to luck. If the market is semi-strong efficient, all public information is incorporated into the market, and one could earn excess return based on private information. In a weak-form efficient market, prices reflect all information from historical market prices (Fama, 1970, p. 69).

If markets are perfectly efficient, there is no profit to gathering information, in which case there would be little reason to trade and markets would eventually collapse (Grossman, 1976, p. 574; Lo, 2004, p. 6). This has led to several no trade theorems – a class of results showing that, under certain conditions, trade in asset markets between rational agents cannot be explained on the basis of differences in information alone. In short, these theorems reason that if the initial asset allocation is commonly known to be efficient, then any proposed trade – even after the arrival of new information – cannot lead to a Pareto improvement over the initial allocation as long as the traders interpret the information in a similar fashion (Serrano-Padial, 2010, p. 1). Even if the market is only weak-form efficient, stock prices should follow a random-walk. Thus, one should not find patterns in stock returns, and for example technical analysis – based on past prices – would not be profitable[4]. In a semi-strong efficient market, fundamental analysis – using public information like a company's earnings, sales, and book-to-market ratios to pick stocks – would also be pointless (Lo, 2017, p. 23). The strong-form of the EMH is an extreme form which few have ever treated as anything other than a logical completion of the set of possible hypotheses (Jensen, 1978, p. 4).

The concept of arbitrage is one of the main fundaments of the EMH; rational agents will observe mispricing and take actions upon it. Noise traders – investors not picking stocks in a rational manner – do not have a significant effect on prices, and it is impossible to consistently beat the market and earn riskless returns. If arbitrage opportunities do exist, rational agents would pick up on these and trade upon them (ter Ellen & Verschoor, 2017, p. 4). According to EMH-supporters, market forces will always act to bring prices back to rational levels, implying that the impact of irrational behavior on financial markets is generally negligible and, therefore, irrelevant (Lo, 2004, p. 7).

Although classical economic models assume agent rationality, there are several anomalies which are puzzling from the perspective of such models. These include – but are not limited to – the forward premium puzzle, the equity premium puzzle, the excess trade volume, the momentum effect, post earnings announcement drift, long term reversal and the size effect (ter Ellen & Verschoor, 2017, p. 5).

Muth's (1961) Rational Expectations Hypothesis (REH) has attracted much attention and

[4]Or, as Fama (1965, p. 57) state, chartist theories would be "*akin to astrology and of no real value to the investor.*"

states that market participants have equal access to information and form their expectations about future events in a uniform, rational manner based on the 'true' probability of the state of the economy (Muth, 1961; ter Ellen & Verschoor, 2017). The assumption of rational agents implies that agents incorporate all available information in their decision-making process and that they are able to do this in an efficient way because they have full knowledge about the economic models underlying financial markets (Muth, 1961, p. 316; ter Ellen & Verschoor, 2017, p. 4). One reason that the rational expectations paradigm is, and has been, the dominant one for so long is that there is only one way to be rational, while there are infinite ways to deviate from rationality (ter Ellen & Verschoor, 2017, p. 27). Economists considered rationality a necessary assumption in sophisticated economic models. Lately, an interesting new literature in the direction of bounded rationality has emerged (ter Ellen & Verschoor, 2017, p. 2). The emergence of behavioral economics and behavioral finance has challenged the efficient market hypothesis, arguing that markets are not perfectly rational (Lo, 2004). The most enduring critiques of the EMH revolve around the preferences and behavior of market participants; individuals tend to be risk averse in the face of gains and risk seeking in the face of losses (Lo, 2004, pp. 4–5). Economists argued behavioral theories were impractical, as it was impossible to model the complex behavior of human beings (ter Ellen & Verschoor, 2017, p. 6). After several decades of research, no consensus is reached regarding whether financial markets are – in fact – efficient (Lo, 2004, 2005).

The Adaptive Market Hypothesis (AMH) was developed by Lo (2004; 2005) in the early 2000s. The AMH reconciles the EMH and behavioral finance so the two theories can co-exist in an intellectually consistent manner (Lo, 2005, p. 2; Lim & Brooks, 2011, p. 72)[5]. Under the AMH, the EMH can be seen as the "frictionless" ideal that would exist if there were no capital market imperfections such as transactions costs, taxes, institutional rigidities, and limits to the cognitive and reasoning abilities of market participants. Or as the steady-state limit of a population with constant environmental conditions – that is, if market participants were given enough time to adapt to a market which does not change (Lo, 2005, pp. 2, 21). Behavioral biases are viewed as heuristics taken out of context, and are not necessarily counterexamples to rationality. Given enough time and competitive forces, such heuristics will be reshaped to better fit the environment (Lo, 2005, p. 2). This is similar to Taleb's (2018, pp. 26, 211–233) argument that rationality is linked to survival[6]. As behavioral biases and heuristics have survived, they cannot be irrational.

[5]Briefly, the precepts that guide the AMH – as outlined in Lo (2005, p. 18) – are (1) individuals act in their own self-interest; (2) individuals make mistakes; (3) individuals learn and adapt; (4) competition drives adaptation and innovation; (5) natural selection shapes market ecology; and (6) evolution determines market dynamics.

[6]"*What is rational is what allows the collective — entities meant to live for a long time — to survive*" (Taleb, 2018, p. 26).

The AMH states that prices reflect as much information as dictated by the combination of environmental conditions, and the number and nature of the participants in the economy; such as pensions funds, retail investors, and hedge-funds. Individuals make choices based on past experience and their best guess as to what might be optimal, and they learn by receiving positive or negative reinforcement from the outcomes. If they receive no such reinforcement, they do not learn. If the environment changes, the heuristics of the old environment are not necessarily suited to the new (Lo, 2004, p. 17). If a small number of participants are competing for rather abundant resources in a given market, that market will be less efficient. As competition increases unsuccessful traders are eliminated from the population, and the market will become more efficient. Market efficiency cannot be evaluated in a vacuum, but is highly context-dependent and dynamic (Lo, 2004, pp. 18–20).

According to the AMH, arbitrage and profit opportunities do exist from time to time. Although they disappear after being exploited by investors, new opportunities are continually being created as groups of market participants, institutions and business conditions change. Mistakes occur frequently, but individuals are capable of learning from mistakes and adapting their behavior accordingly (Lo, 2005, p. 19). An equilibrium state, without arbitrage or even profit opportunities, might exists at times – but according to the AMH this is neither guaranteed nor likely to occur at any point in time (Lo, 2005, p. 20). This is consistent with the conjecture of Grossman and Stiglitz (1980) that sufficient profit opportunities must exist to compensate investors for the cost of trading and information gathering. In fact, Daniel and Titman (1999) have earlier highlighted the possible co-existence of EMH and behavioral finance by introducing the term adaptive efficiency. If a market is "adaptive efficient", there might be pricing anomalties observed in the historical data, but as investors learn from them, they will not persist for too long (Daniel & Titman, 1999, p. 34).

When we move away from the notion that agents are unboundedly rational, we see that all investors need not have equal expectations. Heterogeneous Agents Models (HAM), first developed by Zeeman (1974), takes advantage of this and divides the market participants into several types. These models perform well in describing and explaining asset market dynamics and has the ability to produce important stylized facts observed in financial time series – such as volatility clustering, fat tails, bull and bear markets (ter Ellen & Verschoor, 2017, p. 1). HAM assumes that agents are at least bounded rational, and use rules of thumb to form expectations about future asset prices (ter Ellen & Verschoor, 2017, p. 2). Such models usually include at least two types of agents: chartists, who uses past information to predict future returns; and fundamentalists, who bases his expectations on the deviation of the asset price from its fundamental value (ter Ellen & Verschoor, 2017). Fundamentalist expect market prices to revert to the fundamental value of the respective assets while chartists extrapolate price trends (ter

Ellen & Verschoor, 2017)[7]. In other words, while chartists and fundamentalists demand has a direct effect on returns, fundamentalists may only start selling when a stock is overvalued by a certain amount, thereby causing bull (chartists driving the price up) and bear (both chartists and fundamentalists selling stocks) markets (ter Ellen & Verschoor, 2017, p. 9). Thus, technical analysis – used by the chartists – can serve as a self-fulfilling mechanism (ter Ellen & Verschoor, 2017, p. 10). Several studies show that investors use more speculative strategies for shorter horizons and more fundamental strategies for longer horizons (e.g., Frankel & Froot, 1990; ter Ellen, Verschoor, & Zwinkels, 2013).

In reality, it is very likely that agents do not only differ in the way they form beliefs, but also in the preferences they have, the shocks that they are hit by, and the information set they have access to (ter Ellen & Verschoor, 2017, p. 27).

To conclude, classical theories suggest that there should be no relationship between stock return and measures of trading volume. This predicts that we should at least not be able to find any causal or predictive relations in our empirical investigations. Newer theories, however, allow such relations to exists.

3.2 Reasons for trading

According to Gagnon and Karolyi (2009, p. 954), the motive behind trading, and thus the cause of trading volume, can be attributed to asymmetries in information across groups, unanticipated liquidity and portfolio-balancing needs of investors, or hedging.

Most no trade theorems focus on three different equilibrium notions: common knowledge, incentive compatible trade, and rational expectations equilibria. The most frequent approaches taken by the literature to elicit trade in models of asset markets under asymmetric information is to either weaken the common knowledge assumption or exogenously introduce liquidity – for example through demand shocks or noise traders. Other approaches allow agents to ‘agree to disagree’ by introducing bounded rationality, or to introduce uncertainty to the market (Serrano-Padial, 2010, pp. 2–3).

If we find a relation between return and volume in our empirical investigation, this will mean that the reasons investors have for trading is important for the formation of prices.

[7]Chartists chases trends, therefore buying when prices go up and selling when prices go down. Fundamentalists, are “aware” of the true fundamental value, and buys (sells) when the stock is currently undervalued (overvalued) (ter Ellen & Verschoor, 2017, p. 8).

3.2.1 The role of information

In early models of volume, volume was interesting for its correlation with other variables, but not important in itself (Blume, Easley, & O'hara, 1994, p. 154). Today, trading volume is viewed by many as the critical piece of information that signals where prices will go (Gagnon & Karolyi, 2009, p. 953). Stock markets are merciless in how they react to news. Investors buy or sell shares depending on whether news is good or bad, and the market will incorporate the news into the prices of publicly traded corporations. Good news is rewarded, bad news is punished, and rumors often have just as much impact as hard information (Lo, 2017, pp. 13–14). Since information is costly, prices cannot perfectly reflect the information which is available. If it did, those who spent resources to obtain it would receive no compensation (Grossman & Stiglitz, 1980, p. 405). Most models trying to explain the return-volume relationship are related to the flow of new information, and the process that incorporates this information into market prices (e.g., Andersen, 1996, p. 170; Brailsford, 1996, p. 95).

The two main hypothesis underlying these models are the sequential information arrival hypothesis (SIAH) and the mixture of distributions hypothesis (MDH). SIAH was first developed by Copeland (1976, 1977) and later expanded by Jennings, Starks, and Fellingham (1981). The hypothesis assumes that investors receive information sequentially at different times, which shift the optimists' demand curve up, and the pessimists' demand curve down. Trading occur as a reaction to this new information. Buy trades are viewed as noisy signals of good news, sell trades as noisy signals of bad news (O'Hara, 2015, p. 263). MDH assumes that daily price changes are sampled from a set of distributions with different variances. In the MDH-model specified by Epps and Epps (1976), investors revise their reservation price when new information enter the market. Volume is viewed as the disagreement between the investors (B.-S. Lee & Rui, 2002, p. 54).

In both models, the arrival of new information causes investors to revise their price reservations. Research has established that since investors are heterogeneous in their interpretation of news, prices may not change even though new information enters the market. This might happen if some investors interpret the news as good and others as bad (e.g., Mestel, Gurgul, & Majdosz, 2003, p. 3; de Medeiros & Van Doornik, 2006, p. 2). Volume is always non-negative and as long as at least one investor makes an adjustment in their price revision, expected trading volume is positive (Brailsford, 1996, pp. 93–94). Therefore, volume can be seen as an indicator of consensus, or the lack thereof (Gallo & Pacini, 2000, p. 167). Average investor-reaction to information is reflected in price movements (e.g., Mestel et al., 2003, p. 3; de Medeiros & Van Doornik, 2006, p. 2).

Blume et al. (1994, p. 177) propose an equilibrium model that emphasizes the informational role of volume. They show that volume provides information about the quality of traders' information that cannot be conveyed by prices, and thus observing the price and the volume statistics together can be more informative than observing the price statistic alone. Learning is an important feature in many microstructure models. Most such models rely on the notion that some traders have private information which they trade on. Other traders see market data and they learn from it. Market prices adjust to efficient levels that reflect all the information (O'Hara, 2015, p. 263). A trader watching only prices cannot learn as much as a trader watching both prices and volume and so faces an unnecessary penalty if he ignores the volume statistic (Blume et al., 1994, p. 171). Dealers who are too slow to detect and incorporate new information into quoted prices face the risk that he buys at too high prices or sells at too low prices to informed traders in subsequent trades. Thus, dealers who adjust stock quotes to full information levels more quickly lose less to informed traders (Boulatov, Hatch, Johnson, & Lei, 2009, pp. 1531–1532).

The intrinsic value of securities can change across time as a result of new information. The new information may involve any actual or anticipated change in a factor which is likely to affect the company's prospects (Fama, 1965, p. 56). In an efficient market, at any point in time, the actual price of a security will be a good estimate of its intrinsic value (Fama, 1965, p. 56). However, due to uncertainty, the intrinsic value of a security can never be determined exactly. Thus, there is room for disagreement among market participants concerning just what the intrinsic value of an individual security is, and such disagreement will give rise to discrepancies between actual prices and intrinsic values.

If investors privately observe different information, they will typically hold distinct opinions. Thus, arrival of asymmetric information should induce agents to trade (Serrano-Padial, 2010, p. 1). The high levels of daily trading activity observed in many financial markets is often attributed to speculation: agents hold different views about how much assets are worth (Serrano-Padial, 2010, p. 1).

If there is no noise trading, there will be very little trading in individual assets. A person with information or insight about individual firms will want to trade, but will realize that only another person with information or insight will take the other side of the trade. A trader with a special piece of information will know that other traders have their own special piece of information, and will therefore not automatically rush out to trade (Black, 1986, pp. 530–531). Thus there must be noise in the price system so that traders can earn a return on information gathering (Grossman, 1976, p. 574). With noise traders in the market, it pays for those with information to trade (Black, 1986, p. 531). People not only trade on information, but also on noise, which is essential to the existence of liquid markets (Black, 1986, p. 529). Information traders can never be sure if they are trading on information or noise. If information is already reflected in stock prices, it will be just like trading on noise (Black, 1986, p. 529).

The price of a stock reflects both information and noise that traders trade on (Black, 1986, p. 532). Thus noise causes markets to be somewhat inefficient, but often prevents us from taking advantage of inefficiencies (Black, 1986, p. 529).

If information flows sequentially into the market rather than simultaneous, we would see this in our analysis as a Granger causality between return volatility and trading volume for a significant part of the market. However, due to noise traders, this effect might be hard to establish.

3.2.2 The role of liquidity

Liquidity traders, unlike other traders, do not trade on information. They trade for reasons that are not directly related to the future returns of securities. A liquidity trader is often a financial institutions or large trader where buying and selling is linked to a liquidity need or to rebalncing a portfolio (Admati & Pfleiderer, 1988, p. 5), which according to Cremers and Mei (2007, pp. 1772, 1778) is an essential reason for trading.

3.2.3 The role of hedging

Llorente, Michaely, Saar, and Wang (2002) developed a model with speculative traders and hedge traders to see how they affected the return-volume relationship. According to their model, if a speculative trader and a hedge trader both sell their stocks, the outcome will not be the same. If a speculative trader sells, pricees will decrease and the trade will reflect negative information about the future return of the stock. When a hedge trader trades, the price will still decrease, but there is just a temporary low return, as the expected future payoff is still the same. Thus one expect a higher return for the next period. Consequently, hedge traders generated a negative autocorrelation for return, and they found the opposite for speculative traders (Llorente et al., 2002).

4 Literature review

In this section, we survey the current literature on the volume-return relationship, liquidity-return relationship, and the new market environment.

4.1 The volume-return relationship

There is an old Wall Street adage stating that "*It takes volume to make prices move*." According to Chandrapala (2011), studies of the price-volume relation dates back to the late 1950s when Osborne (1959) laid the theoretical foundation. One of the earliest empirical studies was performed by Granger and Morgenstern (1963), who found the connection between volume and stock prices on the New York Stock Exchange to be negligible. Ying (1966) was the first to document a positive correlation between volume and price change $(V,\ \Delta p)$, and a positive correlation between the volume and absolute price change $(V,\ |\Delta p|)$. In his extensive literature review, Karpoff (1987) states that numerous empirical findings in the 60s, 70s and 80s support the positive volume-absolute price change correlation. Further, Karpoff describes several similar findings for the relationship between volume and price change variance, price change magnitude, price variability, absolute price change, squared abnormal return and squared price change. However, most of these effects are of little economic impact (Karpoff, 1987).

Karpoff (1987) summarize the research conducted before 1987 with the following conclusions:

1. No volume-price correlation exists
2. A correlation exists between volume and absolute price change $(V,\ |\Delta p|)$
3. A correlation exists between volume and price change $(V,\ \Delta p)$
4. Volume is higher when prices increase than when prices decrease

He further suggests that it is likely that the relationship between volume and price changes stems from their common ties to the flow of information or their common ties to a directing process which can be interpreted as the flow of information (Karpoff, 1987).

In Table 2 we have summarized the data used, methodology, and results of several other papers on the volume-return relationship.

Author	Year	Data	Model	Conclusion
Heteroscedasticity in stock Return Data: Volume versus GARCH effects				
Lamoureux & Lastrapes	1990	U.S.	ARCH & GARCH	ARCH and GARCH parameters are dramatically reduced when volume is included in the model. The results suggest that lagged squared residuals have little information about the variance of return after accounting for the rate of information flow, measured as V_t
Stock Prices and Volume.				
Gallant et al.	1992	NYSE: D	VAR, ARCH	Contemporaneous volume-volatility correlation. Large price movements associated with higher subsequent volume. Volume-leverage interaction. Positive conditional risk-return relation after conditioning on lagged volume.
The effects of trading activity on market volatility				
Gallo & Pacini	2000	U.S.	GARCH, EGARCH	Structure of GARCH-type models of conditional heteroskedasticity does not manage to capture the quick absorption of large shocks to returns and implies in practice a too high level of persistence of shocks.
Does Trading Volume Contain Information to Predict Stock Returns? China's Stock Markets				
C. F. Lee & Rui	2000	SSE, SZSE: D	GARCH, VAR	Positive contemporaneous correlation between volume and returns. Trading volume do not Granger-cause stock return in any markets. Return Granger-cause volume. Volume helps predict return volatility and vice versa. Trading volume helps predict the volatility of returns but not the level of returns.

Author	Year	Data	Model	Conclusion
The Dynamic Relation between Stock Returns, Trading Volume, and Volatility				
Chen et al.	2001	U.S., Asia, Europe: D	EGARCH, VAR	Granger causality results show that returns cause volume and, although to a lesser extent, that volume causes returns. GARCH effects remains significant when volume is included in the model.
The Dynamic Relationship between stock returns and Trading Volume				
B. -S. Lee & Rui	2002	NY, Tokyo, London: D	GMM, GARCH, VAR	Positive contemporaneous relationship between volume and return. Trading volume do not Granger-cause returns on any of the markets. Returns Granger-cause volume in the U.S. and Japanese markets, but not int he U.K. market. There is a positive feedback relationship between trading volume and return volatility in all three markets.
The empirical relationship between stock returns, return volatility and trading volume: Austrian market				
Mestel et al.	2003	WBAG	GARCH, VAR	The relationship between stock return and trading volume is mostly negligible. Evidence of a relationship between return volatility and trading volume.
Trading Volume and Returns Relationship in Greek Stock Index Futures Market				
Floros & Vougas	2007	ASE, ADEX	GARCH, GMM	Findings indicate that market participants use volume as an indication of prices.
The Price-Volume Relationship in the Chilean Stock Market				
Kamath	2008	IPSA: D		Granger causality running from returns to volume.

Author	Year	Data	Model	Conclusion
The empirical relationship between stock return, return volatility and trading volume: Brazil				
de Mendeiros & Van Doornik	2006	BOVESPA: D	GARCH, VAR	Significant contemporaneous relationship between return volatility and trading volume. Stock return depends on trading volume, not the other way around. Higher trading volume and return volatility relationship is asymmetrical. GARCH effect and high hysteresis in conditional volatility. Granger causality between trading volume and return volatility is strongly evident in both directions.
The Dynamic Relationship between Price and Trading Volume: Indian Stock Market				
Kumar et al.	2009	S&P CNX Nifty Index	GARCH, VAR	ARCH effects decline when trading volume is included in GARCH equation.
Asymmetric Volatility and Trading Volume: The G5 Evidence				
Sabbaghi	2011	G5 stock markets: D	EGARCH	The findings in this paper support prior research that has documented a positive association between trading volume and return volatility. Persistence levels do not decrease with the inclusion of trading volume in the EGARCH.
Relationship between Trading Volume and Asymmetric Volatility in the Korean Stock Market				
Choi et al.	2012	KOSPI	EGARCH, GJR-GARCH	Trading volume is a useful tool for predicting the volatility dynamics of the Korean stock market.

Table 2: Literature overview

Wang, Wu, and Lai (2018) developed a model that allow for the return-volume dependence to switch between positive and negative dependence regimes. They are the first to divide their observations into four different market conditions: rising return/rising volumes, falling returns/falling volumes, rising returns/falling volumes, and falling returns/rising volumes. They find that the volatilities of return and volume are larger for the negative dependence regime than for the positive dependence regimes. They also find support for heterogeneous investors with short-sale constraints. The return-volume dependence is asymmetric. Both the intensity of information and liquidity trading are important in driving the time-varying, return-volume dependence (Wang et al., 2018).

4.2 The liquidity-return relationship

In addition to the volume-return relationship, much literature has been dedicated to the study of liquidity. As it is hard to have a liquid market without trading going on, volume and liquidity are inextricably linked (e.g Benston & Hagerman, 1974; Stoll, 1978; Ødegaard, 2017, p. 30). A market is said to be liquid if traders can quickly buy or sell a large number of shares at low transaction costs with little price impact (Næs et al., 2008, p. 2). In other words, liquidity includes a cost dimension, a quantity dimension, a time dimension, and an elasticity dimension. In 1990, Lawrence Harris – in the monograph *Liquidity, Trading Rules and Electronic Trading Systems* – defined liquidity along the dimensions width, depth, immediacy, and resiliency (as cited in Ødegaard, 2017, p. 5). Trading volume is used as a measure of the market's depth and resiliency (PricewaterhouseCoopers, 2015, p. 19).

The level of liquidity affects expected returns because investors know that in relatively less liquid stocks, transaction costs will erode more of the realized return (see e.g., Amihud & Mendelson, 1986; Anthonisz & Putniņš, 2016). Thus, investors demand a premium for less liquid stocks and so expected returns should be negatively correlated with the level of liquidity (e.g., Chordia et al., 2001, pp. 29–30). Pástor and Stambaugh (2003) found that stocks with higher liquidity betas exhibit higher expected returns – strong evidence that market-wide liquidity represents a priced source of risk.

Similar to the return-volume relationship, liquidity behaves and is priced asymmetrically (e.g., Anthonisz & Putniņš, 2016, p. 3). By assuming symmetry, the importance of liquidity risk in explaining cross-sectional returns might be underestimated. Anthonisz and Putniņš finds that stocks with high downside liquidity risk compensate investors with an substantial expected return premium (2016, p. 3). This is consistent with investors disliking stocks that are more susceptible to liquidity spirals or abandonment during flights to liquidity. Chordia, Roll, and Subrahmanyam (2002) have found that buying activity is more pronounced following market crashes and selling activity is more pronounced following market rises, while Karolyi et al. suggests that common variation in individual stocks tend to rise during financial crises (2009, p. 21). Anthonisz and Putniņš finds that there is a greater dispersion in downside liquidity risk during illiquid market states than liquid states (2016, p. 26). Pástor, Stambaugh, and Taylor (2017, p. 2) finds that funds trade more when stocks are perceived as mispriced. As high liquidity leads to greater market efficiency, stocks should be more susceptible to mispricing during times of low liquidity (Pástor et al., 2017, p. 27). As portfolio rebalancing is an essential motive for stock trading (Cremers & Mei, 2007, pp. 1772, 1778), this might lead to “herding” effects. This is consistent with Pástor et al. (2017, p. 31) findings of a high commonality in turnover among funds, suggesting that periods of low liquidity might increase trading activity.

Several studies suggest that market microstructure directly influences the liquidity or available supply of a tradable asset which in turn impacts the pricing of the asset (e.g., Abrol, Chesir, & Mehta, 2016, p. 116). Thus, market microsturcture factors can be important as determinants of stock returns. Further, their results suggest a strong incentive for the firm to invest in increasing the liquidity of the claims it issues; like going public, standardize contracts, or enlist on exchanges (Amihud & Mendelson, 1986, p. 246). All traits known to increase trading volume.

4.3 The new market environment

During the last 15 years, trading activity has increased dramatically. Many believe this is due to electronic, algoritmic, and – especially – high frequency trading. By all accounts, high frequency trading has become very significant in today’s markets (Friederich & Payne, 2015). According to Johnson et al. (2012, p. 5), the stock markets have gradually transitioned from a time when trading occurred between humans, to a mixed phase of humans and machines to an ultrafast mostly-machine phase where machines dictate price changes. According to Ødegaard (2017, p. 8) the most important driving force behind the move to electronic trading is cost. Replacing slow, mistake-prone and relatively expensive human labor with capital is a feature of most industries and the financial industry is finally catching up. O’Hara states that the rise of HFT has also radically changed how non-high frequency (HF) traders behave, and the markets where they trade. The current market structure is highly competitive and very fast (O’Hara, 2015, p. 258). The estimated amount of high frequency trading differs greatly (see e.g., Hagströmer & Norden, 2013; Brogaard, Hendershott, & Riordan, 2014; O’Hara, 2015). There is a general, but not universal, agreement that HFT market making enhances market quality by reducing spreads and enhancing informational efficiency (O’Hara, 2015, p. 259). The bid-ask spread narrows, leading to a more efficient price discovery process and increased trading volumes (Hendershott, Jones, & Menkveld, 2011; Abrol et al., 2016). However, many are concerned that HFT induce market instability. In a simulation study, Leal, Napoletano, Roventini, and Fagiolo (2016, p. 49) finds that the presence of HF traders increase market volatility, and several authors points out that HFT might lead to periodic illiquidity (see e.g., Kirilenko & Lo, 2013, p. 63; O’Hara, 2015, p. 259; Van Kervel, 2015, p. 1).

The ability of high frequency traders to enter and cancel orders faster than others makes it hard to discern where liquidity exists in the markets (O'Hara, 2015, p. 258). Abrol et al. finds that the high speeds enables sub second injections and withdrawals of liquidity (2016, p. 126), which is faster than humans can notice and physically react to (Johnson et al., 2012, p. 2). If the investors adapt their strategies on a slower time scale than the time scale on which the trading process takes place, this will lead to positive autocorrelation in volatility and volume, which we might see in our analysis (Brock & LeBaron, 1995). Further, HF orders are sent to and from the exchange as part of complex dynamic trading strategies, and it is now common for upward of 98% of all orders to be canceled instead of of being executed as trades (O'Hara, 2015, p. 259). From a computer perspective, HF trading algorithms in the sub-second regime need to be executable extremely quickly and hence be relatively simple, without calling on much memory concerning past information (Johnson et al., 2012, p. 6). There is therefore a question of how much information such trades incorporate. O'Hara argues that with algorithmic trading, trades are no longer the basic unit of information – the underlying orders are (2015, p. 263).

5 Data

In this section we aim to provide a thorough understanding of the data we have used. First, we explain where we obtained the data, and how the data was calculated originally. Then, we take a look at the sample period, and give a few comments about things to watch out for. Next, we comment on our data preparation process, before we detail our filtering of the data.

5.1 Variables and data sources

When writing this thesis, we got access to *Oslo Børs Informasjon AS / BI's Database*. From this database, we downloaded daily returns and daily trading volume of all equity instruments on the Oslo Stock Exchange from the start of January 1980 to the end of November 2017.

According to the notes at *Oslo Børs Informasjon AS / BI's Database*, the stock returns are "raw" returns, calculated as

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

adjusted for dividends and corporate events like stock splits. The returns are not annualized. *Oslo Børs Informasjon AS / BI's Database* state that P_t was found using the algorithm in Figure 2.

The two main reasons why we investigate returns rather than prices is that investors are

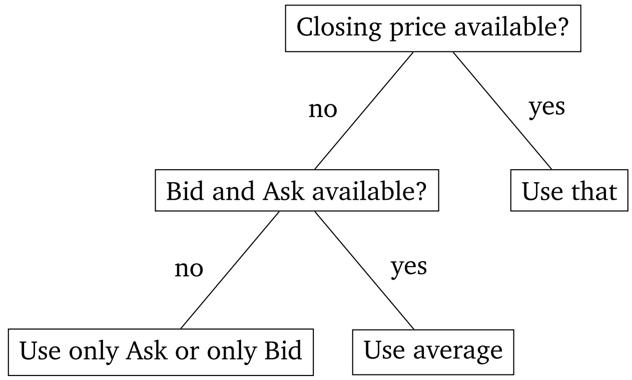

Figure 2: Algorithm for calculating daily returns

mostly interested in returns for their investment decisions and that the properties of returns are in general easier to handle than the properties of prices.

The volume data gives us number of trades for days were trading occurred. That is, days with no trades are not recorded at all, and will show up in our dataset as missing values when the return and volume data are combined.

5.2 Sample period

Our sample period is quite long, spanning 38 years of daily data. This is positive, as it allow us to include a lot of information in our models. Further, we wanted our analysis to cover several full business cycles. There are, however, some drawbacks. First, as detailed in the literature review the market environment seems to have changed, and what happened in the 80s might not be very relevant for today and the near future. The long sample might have time-varying properties which makes it hard to draw conclusions valid for the full period. Over long sample periods, changes in market structure, competition, technology, and activity in financial markets can potentially generate non-stationarities in financial time series (Næs et al., 2011, p. 147). As seen in Figure 1 in Section 2.1, there have been several technological changes at the Oslo Stock Exchange. For example, the launch of an electronic trading system in 1988 and the fully automatic trading system in 1999. Further, there have been changes in the availability of information. Today, everyone can find the last stock price down to the minute online for free, while only 18 years ago there was a 15 minutes delay for this information.

In the 38 years we have data, Norway has been through several full business cycles. We will here comment on some extreme events for this period.

Before 1980, Norwegian economic politic had been characterized by creating a welfare state and building up the petroleum industry (Steigum, 2010). Price regulations in the real estate market was abolished in the early 80s, and restrictions on cross-border capital flows was gradually removed during the 1980s towards a full liberalization in 1990 (Steigum, 2010, pp. 13–14). In October 1987 the markets crashed. The main index dropped by 20% in one day and by the end of October the Norwegian stock market had declined by 28% (Næs et al., 2008, p. 30).

Next up was the banking crisis lasting six years from 1988 to 1993. Banks representing 95% of all commercial bank assets in Norway became insolvent, and the government was forced to bail out numerous financial institutions (Ongena, Smith, & Michalsen, 2003, p. 81). The event that marked the beginning of the crisis, was an earnings report issued by Sunnmørsbanken on March 18th, 1988, stating that it had lost all of its equity capital. The last distress announcements occurred in 1991, but the banking sector did not really stabilize until 1993 when the banks began to record improved results (Næs et al., 2008, p. 31). Although the banks experienced a large and permanent downward revision in their equity capital during the period, the firms that maintained relationships with the banks did only experience small and temporary changes in their stock prices (Ongena et al., 2003, p. 81). Overall, the aggregate impact of bank distress appears small (Ongena et al., 2003, p. 81), and should not affect our sample too much.

Most recently was the financial crisis of 2007-2008. In July and August 2007, the main index at the Oslo stock exchange fell by 2.3 and 4.3 percent respectively. The drop in the market was related to increased uncertainty surrounding the U.S. sub-prime market and potential long run effects of this crisis (Næs et al., 2008, p. 33). However, the full impact of the crisis would not hit Norway before 2008. According to Oslo Stock Exchange, the fall of 2008 would be characterized as one of the worst periods for the exchange, as the value of the stocks at OSE plunged by over 40%, as can be seen in Figure 3.

5.3 Data structure and preparation

There are four main data-files from *Oslo Børs Informasjon AS / BI's Database* we will rely on: a daily returns dataset, a daily volume dataset, a dataset for identifying securities and companies based on a set of names and ID-numbers, and a dataset with monthly observations of stock prices and number of outstanding shares – used for filtering our data later.

Unfortunately, none of these files were in a format optimal for data analysis or for matching the correct volume and return observations when merging the datasets. Therefore, a large portion of our thesis was to structure these datasets, before we could combine and clean them.

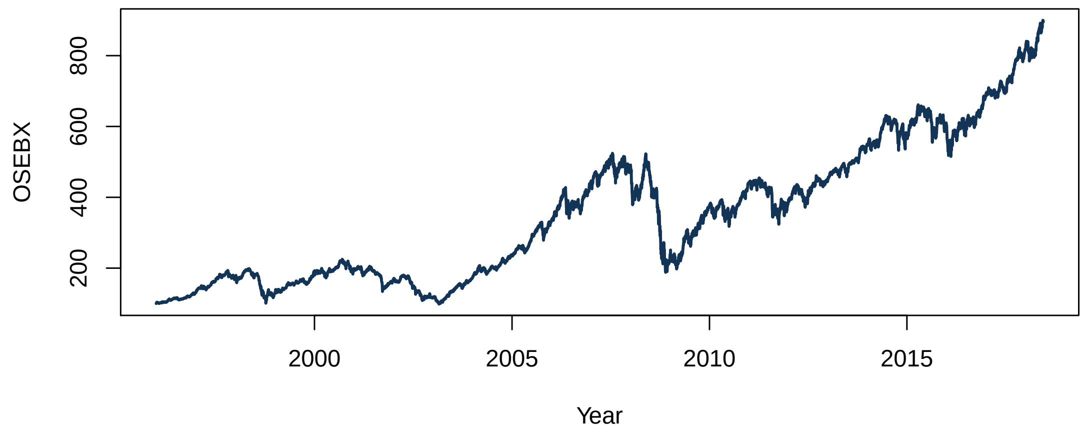

Figure 3: OSEBX – historical levels

According to de Jonge and van der Loo (2013, p. 7), the data preparation process may profoundly influence the statistical statements based on the data and should be considered a statistical operation to be performed in a reproducible manner. We have based our data preparation process on statistical literature, and provided both explanation and justification for the steps we have taken. Documentation of this process is necessary for control and reproducibility of our thesis, but as this part is rather lengthy, and with no direct relevance for the research question at hand, we have detailed our data preparation process in Appendix A.

All data handling in this process was performed using the open source statistical software R (R Core Team, 2017). All R packages used are cited in Appendix A, while the complete R-code for importing, structuring, combining, cleaning, and filtering our data can be found in Appendix B.

After structuring, combining, and cleaning our data, we have a dataset of almost 1.7 million rows and 14 columns: date, year, month, ticker, last company name, last security name, ISIN, OBI security ID, return, volume, last price of the month, number of shares outstanding at the end of the month, the market capitalization (MCAP) at the end of the month, and a dummy variable equal to 1 if the volume is positive and 0 if the volume is 0.

5.4 Filtering

Not all stocks traded at the OSE should necessarily be used in empirical investigations, and it is common to apply certain filters before analyzing the data (Ødegaard, 2018, p. 17). We have used the following filters:

1. We only include stocks which are in the sample at the end of each month.
2. Only companies with an average market capitalization above NOK 1 Million each year.
3. A stock needs an average price above NOK 10 each year – so called “penny stocks” are removed.
4. A stock needs an average price below NOK 8.000 each year.
5. A stock need to have at least 20 trading days a year.
6. Norwegian Savings banks – issuing equity certificates and not stocks – are removed.
7. Other non-stock equities are removed.
8. Securities with less than 500 observations in total was removed.

First, we remove observations where we lack the MCAP. As the MCAP is calculated on a monthly basis, this means that at least one month worth of observations is removed for the stocks in question. The missing MCAP is either due to the price lacking – which is the case the last month of trading for companies that were delisted from the exchange – or due to numbers of shares outstanding missing. We note that most of the cases were we lack the number of shares outstanding are foreign companies noted on the OSE, preferred shares, or equity certificates of small savings banks. In this process, all companies not showing up at the end of a month were also removed from that month, thus fulfilling filter 1.

Next, we followed Ødegaard’s (2018, p. 17) suggestion to remove companies with an MCAP below NOK 1 Million. We defined a vector of company names which at some point during our sample period had an MCAP below NOK 1 Million, and used this vector to check the average yearly MCAP of these companies. We found that for most of these companies, their MCAP were low the first few years of their listing at the OSE, before they grew in size. We decided to remove just the years of observations where the yearly average MCAP was below NOK 1 Million.

Another suggestion by Ødegaard (2018, p. 17) were to remove penny stocks. This is due to the volatile behavior of such stocks’ returns. For the opposite reason, Chordia, Roll, and Subrahmanyam (2011, p. 245) recommended to remove stocks with a value above USD 999. Stocks with a yearly average price below NOK 10 or above NOK 8.000 were consequently removed from that corresponding year.

According to Ødegaard (2018, p. 17), stocks which are seldom traded are especially problematic in empirical asset pricing investigations. Following his advise, we define seldom traded stocks as those with less than 20 trading days a year. We created a dummy variable which were

1 if a stock were traded at a given day – and 0 otherwise – and removed stocks for the full year if total trading days within that year were below 20 days.

Next, we decided to remove all Norwegian Savings banks due to their different ownership structure and issuance of equity certificates rather than stocks. We did this by filtering our data for company and security names that included the substring "*spare*". We made sure not to remove Sparebank 1 SR-Bank post 2011, as it was transformed from a Savings bank to a commercial bank. Similarly, we made sure to remove Sandsvær banken and Sparabanken Rogaland, two Savings banks without "*spare*" in their name.

As suggested by Chordia et al. (2011), we wanted to remove all non-stock equities, as their trading characteristics might differ from stocks. However, in the prior filtering process, all such instruments had been removed, which we checked for extensively.

Last, as some statistical measures – such as skewness and kurtosis – and a number of time series models are sensitive to small samples, we wanted to remove securities with few observations. Hwang and Valls Pereira (2006) suggest that the sample size should be at least 500 if one wants to estimate a GARCH(1,1) model. We choose to remove all securities where we do not have at least 500 observations – about 2 years of daily observations during our 38 year sample period.

After filtering, we are left with the daily return and volume of 511 stocks. A full list of the companies included in our sample can be found in Appendix C.

6 Methodology, analysis, and results

In this section we will present and interpret our results. To make sure that our results can be validated as well as replicated, we also detail the methodology behind our analysis. We will start with a descriptive and exploratory analysis before analyzing different models, which will tend to both a potential contemporaneous and causal relationship.

Although our analytical approach was developed along the way dependent on our findings, the research design was originally inspired by B.-S. Lee and Rui (2002), Mestel et al. (2003), and de Medeiros and Van Doornik (2006). The implementation in R was occasionally inspired by Kleiber and Zeileis (2008), Arratia (2014), and Ruppert and Matteson (2015).

As we are analyzing over 500 stocks individually, we will report summary statistics from models and regressions from all these stocks. When discussing results and parameters, we are referencing the mean/median level of these unless otherwise is stated. A significance level of 5% will be used throughout the thesis.

The analysis has been performed using the open source statistical software R (R Core Team, 2017), and the full code is available in Appendix F. R packages used will be cited consecutively as different packages may have different specifications. Complete results are available upon request. Most tables have been created directly from R using the package `stargazer` (Hlavac, 2018), and shows the return and turnover in whole percentages.

6.1 Measures

6.1.1 Volume

The goal of this thesis is to explore the empirical relationship between stock return and trading volume. One of the first decisions we had to make was to decide upon a measure for volume. A much applied measure of trading activity is *turnover* – the number of shares traded over the number of shares outstanding – sometimes referred to as *relative volume* (Campbell, Grossman, & Wang, 1993; Lo & Wang, 2000). This measure was suggested by, among others, Lo and Wang (2000), and have for example been used by Næs et al. (2008) and Skjeltorp, Ødegaard, et al. (2009) when studying the Oslo Stock Exchange. Other measures suggested in the literature, such as number of shares traded (Gallant et al., 1992), were considered but discarded due to the lack of standardization. Turnover, as a relative measurement, will allow us to compare our results between securities.

The number of shares outstanding for the stocks used in the analysis was only possible to obtain at an end-of-month basis, while we could get number of shares traded on a daily basis. As mentioned in Section 4.3 of the literature review, the stock markets have been through a major change over the last couple of years with the entry of high frequency traders. Thus, we believe that daily data would be the most interesting to analyze. As the number of outstanding shares changes rather seldom, we decided to calculate a daily turnover measure as

$$turnover_{i,d} = \frac{number\ of\ shares\ traded_{i,d}}{number\ of\ shares\ outstanding_{i,m-1}}$$

where i denotes the company, d denotes the day and $m-1$ denotes the last day of the previous month.

Turnover will be used in all of our analysis. The terms turnover, relative volume, trading activity, and volume will be used interchangeably.

6.1.2 Volatility

Both the Mixed Distribution Hypothesis and the Sequential Information Arrival Hypothesis – discussed in Section 3.2.1 – link trading volume with return volatility. The MDH model of Clark (1973) use volume as a measure for flow of information and predict that there is a contemporaneous but not a causal relationship between the two variables (Ahmed, Hassan, & Nasir, 2005, p. 148). SIAH with its sequential flow of information to traders show that past trading volume provides information on current volatility (Lu & Lin, 2010, p. 93).

To explore the return-volume relationship we need a measure for volatility. A popular, often used measured of volatility is squared return. According to Andersen and Bollerslev (1998), squared return is an unbiased estimator for volatility. Brailsford (1996), B.-S. Lee and Rui (2002), and Mestel et al. (2003) all use squared return as a proxy for volatility in their model, and so will we.

6.2 Exploratory analysis

We start with a descriptive analysis of stock return and turnover for the full sample, summarized in Table 3.

Statistic	Return	Turnover
Mean	0.11	0.25
Max	$1,200.00$	624.77
Pctl(75)	1.36	0.20
Median	0.00	0.04
Pctl(25)	-1.34	0.0004
Min	-95.00	0.00

Table 3: Descriptive statistics – whole sample

The mean daily stock return equals 0.11%, with a majority of observations concentrated around $\pm$ 1.35%, and with an extreme maximum of 1,200%. For turnover, the range goes from 0% to almost 625% for a stock a day. The mean of 0.25% is largely affected by the extreme values, as the median and the 75th percentile expose that most stocks have a much lower turnover.

The minimum turnover of 0% and the minimum return of -95% confirm that we did not miss any obvious errors when cleaning the data, detailed in Section A.2.3 of Appendix A.

6.2.1 Descriptive statistics - Stock return

(a)

Statistic	Mean	St. Dev.	Max	Pctl(75)
Mean	0.11	4.34	55.05	1.50
Max	1.30	34.86	1,200.00	3.85
Pctl(75)	0.16	5.05	51.96	1.73
Median	0.09	3.73	31.95	1.43
Pctl(25)	0.04	2.85	21.88	1.18
Min	-0.49	0.85	3.24	0.47

(b)

Statistic	Median	Pctl(25)	Min	Kurtosis	Skewness
Mean	-0.01	-1.52	-28.22	54.58	1.98
Max	0.07	0.00	-3.13	3,044.36	47.02
Pctl(75)	0.00	-1.11	-16.43	28.27	1.74
Median	0.00	-1.41	-23.75	11.03	0.79
Pctl(25)	0.00	-1.82	-34.41	6.49	0.27
Min	-0.62	-3.99	-95.00	0.84	-7.61

Table 4: Descriptive statistics – Return – Individual securities

Table 4 contains descriptive statistics for daily stock return of individual stocks throughout the time series. That is, we calculated the statistics mean, standard deviation, maximum, 75$^{\text{th}}$ percentile, median, 25$^{\text{th}}$ percentile, minimum, kurtosis, and skewness for all the 511 stocks and saved this to a 511×9 matrix, before we calculated summary statistic of each of the columns of the matrix. The skewness and kurtosis was calculated using the R package `e1071` (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2017). The skewness was calculated as

$$\frac{m_3}{s^3}$$

and the kurtosis[8] as

$$\frac{m_4}{s^4} - 3,$$

where m_3 and m_4 is the third and fourth sample moments respectively and s is the standard deviation.

From Table 4 we see that the different return series have a mean (median) kurtosis of 54.58 (11.03), ranging from 0.84 to 3,044. The high excess kurtosis, way above 0, suggest a

[8] As is common in finance, we will use excess kurtosis – kurtosis minus three – when referring to kurtosis.

leptokurtic distribution for most stocks. The skewness range from -7.61 to 47.02, with a mean (median) of 1.98 (0.79). A positive skewness indicates that the right tale is fatter and/or longer than the left one. We found some of these values to be surprisingly high, and checked the series manually. We found nothing suspicious, except that most series were highly concentrated around zero.

6.2.2 Descriptive statistics - Turnover

Table 5 contains descriptive statistics for daily turnover of individual stocks throughout the time series, calculated the same way as Table 4. Turnover has a mean (median) kurtosis of 493 (234) ranging from 3 to 7,055, indicating a leptokurtic distribution for all stocks. As with returns we found the very high kurtosis to be surprising and checked the series manually. We find nothing suspicious about the series, except that a lot of them are highly concentrated around 0% turnover due to 0 trades. The skewness range from 1.5 to 80.6, with a mean (median) of 16.4 (12.9). As turnover is always non-negative, it comes as no surprise that the distribution is skewed to the right.

(a)

Statistic	Mean	St. Dev.	Max	Pctl(75)
Mean	0.28	1.02	27.27	0.25
Max	3.54	20.43	624.77	4.44
Pctl(75)	0.36	1.33	33.69	0.32
Median	0.20	0.67	15.88	0.13
Pctl(25)	0.09	0.36	6.90	0.04
Min	0.00	0.01	0.15	0.00

(b)

Statistic	Median	Pctl(25)	Min	Kurtosis	Skewness
Mean	0.10	0.04	0.00	492.97	16.40
Max	3.15	2.05	0.58	7,054.72	80.62
Pctl(75)	0.12	0.04	0.00	600.14	22.06
Median	0.03	0.01	0.00	234.17	12.89
Pctl(25)	0.00	0.00	0.00	88.76	7.68
Min	0.00	0.00	0.00	3.06	1.48

Table 5: Descriptive statistics – Turnover – Individual securities

According to Engle (2002, p. 428) there are two conventional approaches to modeling non-negativity: ignore the non-negativity, or take the logarithms. As we have values of exactly 0 in our turnover data, we cannot model the logarithms without modifying the values somewhat[9]. Further, most of our data – 75% – is distributed between [0, 0.2]. The already short range does not favor taking the logarithms. Although the high skewness might argue for taking the logarithm (Kleiber & Zeileis, 2008, p. 57), we will resort to the first approach of ignoring the non-negativity. Instead, we will examine the outliers of the data, and discuss what to do with them.

6.2.3 Outlier handling

There are many technical definitions of outliers, but an intuitive one is "*an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*" (Hawkins, 1980, p. 1). Outliers can create problems as they can shift the estimates and the p-values for both linear regression and maximum likelihood estimates. Apart from ignoring them, there are several ways to handle outliers. One is to trim the data. That is, to simply remove observations too far from the median; one trims the tails of the distribution. Another is to use winsorization. Winsorizing the data is to censor outliers by reducing them, so they are more in line with the bulk of the data, instead of removing them. Since financial time series often are heavy tailed, outliers represent valid observations and should be kept in the sample (Hawkins, 1980, p. 5). As we want to keep the information in the tails, we decide to use winsorization rather than trimming the series.

We are interested in the relationship between return and trading activity. As pairwise observations can be outliers together – without any of the single observations being so in their separate distributions – we need to take correlation outliers into account. Thus, we will use the bivariate winsorization method suggested by Khan, Van Aelst, and Zamar (Khan et al., 2007, p. 1291). This method handle correlation outliers much better than the univariate winsorization method.

To allow for different outlier-levels for the different stocks, we winsorized the series on a stock-by-stock basis. The implementation in R was done using the package `robustHD` (Alfons, 2016). The borders of the main part of the data are defined using the median and median absolute deviation, with a fallback option to use the mean and standard deviation for stocks where the robust measures where too small to calculate. A normal distribution is assumed, and the data is shrunken towards a boundary of a tolerance ellipse with coverage probability of 99%[10].

[9] By, for example, adding a small constant (Engle, 2002, p. 429).

[10] The function used for winsorization introduced negative values of turnover for about 1.2% of the observations, with a minimum value of approximately -0.0000000000000001%. As neg-

Summary statistics of the winsorized data can be found in Table 6, 7, and 8. The statistics are calculated the same way as in Table 3, 4, and 5.

Statistic	Return	Turnover
Mean	-0.004	0.15
Max	29.28	12.16
Pctl(75)	1.12	0.15
Median	0.00	0.03
Pctl(25)	-1.19	0.002
Min	-27.73	0.00

Table 6: Descriptive statistics – whole sample – winsorized

(a)

Statistic	Mean	St. Dev.	Max	Pctl(75)
Mean	-0.02	2.67	6.83	1.23
Max	0.79	8.19	29.28	3.57
Pctl(75)	0.04	3.13	8.04	1.52
Median	-0.01	2.46	6.38	1.18
Pctl(25)	-0.07	2.02	5.24	0.91
Min	-0.55	0.75	1.79	0.10

(b)

Statistic	Median	Pctl(25)	Min	Kurtosis	Skewness
Mean	-0.01	-1.33	-6.84	0.71	0.06
Max	0.07	0.00	-1.79	3.85	0.63
Pctl(75)	0.00	-0.96	-5.24	0.98	0.11
Median	0.00	-1.26	-6.38	0.62	0.05
Pctl(25)	0.00	-1.65	-8.04	0.35	0.01
Min	-0.62	-3.45	-27.73	-0.63	-0.30

Table 7: Descriptive statistics – Winsorized Return – Individual securities

As expected, we see that the maximum values of both turnover and return has decreased drastically after winsorizing, and the minimum value of return has increased much as well. The mean has also changed quite a bit for both measures, while the median – robust to outliers – barely moved. Another striking feature of the winsorized data is that the of return has a mean (median) kurtosis of 0.71 (0.62) which is much closer to a mesokurtic distribution than the mean (median) of 54.58 (11.03) return used to have before winsorizing. This, together with

ative turnover values does not make economic sense and are not present in the original data, we change these values to exactly 0.

(a)

Statistic	Mean	St. Dev.	Max	Pctl(75)
Mean	0.16	0.17	0.87	0.24
Max	3.44	1.80	12.16	4.40
Pctl(75)	0.20	0.23	1.05	0.31
Median	0.08	0.11	0.39	0.12
Pctl(25)	0.03	0.04	0.13	0.04
Min	0.00	0.00	0.00	0.00

(b)

Statistic	Median	Pctl(25)	Min	Kurtosis	Skewness
Mean	0.10	0.04	0.00	17.50	2.03
Max	3.15	2.05	0.58	2,007.19	40.69
Pctl(75)	0.12	0.04	0.00	-0.02	0.96
Median	0.03	0.01	0.00	-0.63	0.85
Pctl(25)	0.01	0.00	0.00	-1.19	0.68
Min	0.00	0.00	0.00	-2.00	0.00

Table 8: Descriptive statistics – Winsorized Turnover – Individual securities

the new mean (median) skewness of 0.06 (0.05), looks much closer to a normal distribution than before winsorizing.

The kurtosis and skewness of turnover is also greatly reduced after winsorizing. As seen by the maximum kurtosis, some turnover series still have a very leptokurtic distribution. This seems to be because the series is so highly concentrated around 0% in turnover (due to many days of 0 trades). The 75th percentile, median, and minimum tells us that a lot of the winsorized turnover series have a platykurtic distribution with very thin tails compared to the normal distribution. Some of these cases seems to be because as the outliers where compressed down to the bulk of the data, the distribution became rather uniform.

Hereafter, we will use the winsorized data without specifying that it is winsorized.

6.2.4 Jarque-Bera test for normality

Financial time series tend to display non-normal tendencies, which we will check our data for as well. The Jarque-Bera (JB) test takes into consideration skewness and kurtosis when checking if the distribution can be classified as normal. A normal distribution have expected skewness and expected excess kurtosis of 0.

The test is defined as

$$H_0 : data\ is\ normally\ distributed$$
$$H_a : data\ is\ not\ normally\ distributed$$

With test statistic

$$JB = \frac{n}{6}\left[(\sqrt{b_1})^2 + \frac{(b_2-3)^2}{4}\right]$$

where $\sqrt{b_1}$ is the skewness, b_2 is the kurtosis, and n is the number of observations (Jarque & Bera, 1987). Under the null hypothesis, the JB statistic asymptotically has a chi-squared distribution with two degrees of freedom.

The test was calculated using the R package `normtest` (Gavrilov & Pusev, 2014). The package use 2,000 Monte Caro simulations to estimate the P-values.

Although the kurtosis and skewness of the return looked rather normal, we find that the null hypothesis is rejected for all but 130 stocks at a 1% significance level. Thus, only about 1/4 of our sample have normally distributed returns. For the turnover, the null hypothesis for normality was rejected for all stocks at a 1% significance level.

We will need to take this limitation into account when choosing models and methodology going forward.

6.2.5 Ljung-Box test for serial dependence

It is often found in financial research that returns do not exhibit significant serial dependence while volume measures – such as turnover – do. We apply The Ljung-Box test (Ljung & Box, 1978) to test for autocorrelation in our data. The test is defined as

$$H_0 : data\ is\ independently\ distributed$$
$$H_a : data\ is\ not\ independently\ distributed$$

With test statistic

$$Q = n(n+2)\sum_{k=1}^{m}\frac{(\hat{r}_k)^2}{n-k}$$

where k is the lag, m is the maximum of lags tested, n is the number of observations, and $\hat{r}$

is the correlation between the series at time t and time $t-k$. Under H_0, the test distribution is

$$Q \sim \chi^2_m$$

and H_0 is rejected if

$$Q > \chi^2_{1-\alpha,m}$$

where $1-\alpha$ is the quantile, with a significant level of α.

In our analysis we have used $\alpha = 5\%$ and $m = 10$, and thus rejecting H_0 if

$$Q > \chi^2_{0.95,10} = 18.307$$

By a Ljung-Box test, we found autocorrelation in the return for 79% of the stocks within the first 10 lags at a 5% significance level. Next, we found that the stocks which displayed autocorrelation had an average of 1.4 significant lags, with a median of 1. Thus, most return series display significant autocorrelation, but mainly of low order.

Further, by a Ljung-Box test, we found a significant turnover autocorrelation for 99% of the stocks within the first 10 lags at a 5% level. When checking, we found that the turnover series had many more significant lags than return.

Although both the stock return and the turnover series display persistence, it is much stronger for turnover than return. This indicates that past trading activity can be used to predict future trading activity to a greater extent than past values of return can predict future returns. If the variables can be predicted, they cannot be completely random. This would argue against the efficient market hypothesis, as described in Section 3.1.

6.2.6 Unit root

As mentioned in Section 5.2, long sample periods can potentially generate non-stationarities in financial time series (Næs et al., 2011, p. 147). Generally data have to be stationary before any empirical analysis, and we are later going to use models like the Vector Autoregressive (VAR) model, which is sensitive to this. Hence we need to test whether our time series are non-stationary.

To do so, we test for a unit root by using a Phillips-Perron (P-P) test and an augmented Dickey-Fuller (ADF) test, where a rejection of the null hypothesis indicates that the time series are stationary. The P-P test is more robust than the ADF to a wide range of serial correlation and time-dependent heteroskadesticity (B.-S. Lee & Rui, 2002, pp. 57–58), while there is a good deal of evidence that the ADF outperforms the P-P test in finite samples (Davidson & MacKinnon, 1999, p. 613).

The P-P test was developed by Phillips and Perron (1988) and is based upon one of the three following regression models (Banerjee, Dolado, Galbraith, Hendry, et al., 1993, p. 109-110)

$$y_t = py_{t-1} + u_t$$

$$y_t = \mu + py_{t-1} + u_t$$

$$y_t = \mu + \gamma(t - T/2) + py_{t-1} + u_t$$

where the null hypothesis, $H_0 : p = 1$, is tested against the one-sided alternative hypothesis, $H_a : p < 1$.

The R package `tseries` (Trapletti & Hornik, 2018) was used in testing for unit root by P-P test. The test has a general regression equation which include the constant and linear trend, μ and γ, similar to the last equation above.

For both stock return and turnover, the null hypothesis of the P-P test was rejected for all companies at a 5% level. No individual stock return or turnover series displayed non-stationarity, according to this test.

The Dickey-Fuller test was developed by Dickey and Fuller (1979) and later modified to fit larger time series. The difference between the Dickey-Fuller test and augmented Dickey-Fuller test is that the regression has been augmented with the lagged changes of y_t. The aim of including lagged values is to control for any serial correlation in Δy_t (Wooldridge, 2016, p. 576). There are three possible models

$$y_t = \gamma y_{t-1} + \gamma_1 \triangle y_{t-1} + \cdots + \gamma_p \triangle y_{t-p} + \varepsilon_t,$$

$$y_t = \mu + \gamma y_{t-1} + \gamma_1 \triangle y_{t-1} + \cdots + \gamma_p \triangle y_{t-p} + \varepsilon_t,$$

$$y_t = \mu + \beta t + \gamma y_{t-1} + \gamma_1 \triangle y_{t-1} + \cdots + \gamma_p \triangle y_{t-p} + \varepsilon_t,$$

with lag length p (Greene, 2012, p. 994; Banerjee et al., 1993).

The null hypothesis is that there is a unit root present in the sample, $H_0 : \gamma = 1$, and is tested against the alternative hypothesis of no unit root, $H_a : \gamma < 1$.

An alternative specification is to subtract y_{t-1} from both sides of the equation to obtain

$$\triangle y_t = \mu + \beta t + \gamma^* y_{t-1} + \sum_{j=1}^{p} \phi_j \triangle y_{t-j} + \varepsilon_t,$$

where

$$\phi_j = -\sum_{k=j+1}^{p} \gamma_k \;\; and \;\; \gamma^* = \left(\sum_{i=1}^{p} \gamma_i \right) - 1$$

and the null hypothesis is $H_0 : \gamma^* = 0$ against the alternative hypothesis $H_a : \gamma^* < 0$.

One R package for running an ADF test is `tseries` (Trapletti & Hornik, 2018). In this package, the general regression equation include a constant and a linear trend, μ and βt, like the last of the three equations above. We want to run the ADF test without the linear time trend, βt, similar to the second of the three equations above.

The R package `aTSA` (Qiu, 2015) has an ADF test with the constant but without linear trend. However, the output from this test did not have the wanted format. The solution was to create a custom function in R by modifying the script behind the ADF test function in the `tseries` using inspiration from `aTSA`. The critical values for an ADF test changes based on what model specification is used. Hence a part of rewriting the function was to change the critical values according to the ones found in Table 4.2(b) p. 103 in Banerjee et al. (1993). The function can be found in Appendix D.

The interpretation if we can reject H_0 with the intercept-only specification of the ADF is that y_t is stationary around a constant – there is no long term growth in the data.

We found that for stock return the null hypothesis was rejected for all companies at a 5% level. None of the return series display a unit root, according to the ADF test.

For the turnover series we found that for six companies the null hypothesis of the ADF test could not be rejected at a 5% level. To figure out why these companies where non-stationary we created plots of their volume, raw turnover, and winsorised turnover. Exploring every individual plot did not reveal any single incident or pattern that could have explained why the null hypothesis was not rejected. However, all the series had some periods of high trading volume and similarly high turnover periods. After winsorizing, these periods with many outliers where reduced so that the stock would display periods of the same high turnover, which might explain the results from the ADF test. As it was only six out of 511 companies that the null hypothesis could not be rejected for, we decided to exclude them from our sample. The loss of information from removing these six companies seems like a fair tradeoff to stay on the safe side with regard to statistical inference.

As earlier studies have found trends in volume data, we were a bit puzzled that we did not find a unit root. We decided to test our raw volume data, but found more or less the same results. Then, we aggregated our volume data for each stock by date and plotted it against time. The plot showed a clear trend. As a result we decided to test the accumulated volume for stationarity using an ADF test with both constant and linear time trend. What we found was that the null hypothesis of non-stationarity could not be rejected. Thus the market as a whole has a trend in volume and a unit root, but individual stocks do not.

As our analysis is based on daily individual time series we do not need to detrend turnover. However, as so many earlier studies have shown a trend in volume, we decided to remove a linear trend from our volume data using the `pracma` package in R (Borchers, 2018).

Hereafter, we will use the detrended turnover, without specifying that it is detrended.

6.3 Cross-correlation analysis

Being done with the exploratory analysis, a cross-correlation analysis is the first step in investigating the relationship between stock return and turnover.

We used the formula below to calculate the cross-correlation between two time series

$$\rho(y_t, x_{t-j}) = \frac{Cov(y_t, x_{t-j})}{\sigma(y_t)\sigma(x_{t-j})}$$

where y_t and x_t are two time series at time t, and j is the lag/lead between them.

Table 9 display a summary of the cross-correlation between return and turnover for each stock, while Figure 4 shows the full distribution. There is a low but mostly positive contemporaneous correlation between stock return and turnover. Although the correlation is low, the contemporaneous relationship between stock return and turnover will be further investigated. According to Kozak (2009), even a weak correlation can be of importance if the expectation was for there to be none – like predicted by the EMH detailed in Section 3.1.

Compared to the contemporaneous correlation, there is an even weaker but mostly positive correlation between lagged/lead turnover and stock return. The lagged turnover correlation is slightly more positive than the lead turnover.

Statistic	$j=-4$	$j=-3$	$j=-2$	$j=-1$	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
Max	0.12	0.10	0.11	0.15	0.29	0.20	0.17	0.14	0.12
Pctl(75)	0.01	0.02	0.02	0.05	0.10	0.07	0.05	0.04	0.03
Median	-0.001	0.001	0.01	0.03	0.06	0.03	0.02	0.02	0.01
Pctl(25)	-0.02	-0.01	-0.01	0.01	0.02	0.01	0.004	0.001	-0.003
Min	-0.11	-0.12	-0.08	-0.17	-0.11	-0.11	-0.07	-0.08	-0.09

Table 9: Cross-Correlation between Return and detrended turnover: $Corr(R_t,\ V_{t-j})$

Under the null hypothesis, the cross-correlation coefficients are asymptotically normal with a variance of approximately $1/n$, where n is the length of the series. At a 5% significance level, correlations larger in magnitude than approximately $\pm 1.96/\sqrt{n}$ are deemed significant (Cryer & Chan, 2008, p. 261). The number and percentage of significant correlations for each lag can be found in Table 10.

Significant	$j=-4$	$j=-3$	$j=-2$	$j=-1$	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
Number	21	25	24	138	291	188	117	90	65
Percentage	4.16	4.95	4.75	27.33	57.62	37.23	23.17	17.82	12.87

Table 10: Securities with cross-correlation different from 0 at 5% significance level

Table 10 shows that the correlation is significantly different from 0 for almost 58% of the securities in the contemporaneous relationship. Furthermore, the lagged turnover is more significant than the leading.

It is often stated that price variation tend to increase if there is high trading activity, thus there might be a link between trading activity and higher order moments of stock returns.

Table 11 display a summary of the cross-correlation between return volatility and turnover, where the proxy for return volatility is squared return. The full distribution of the cross-correlation can be seen in Figure 5. The contemporaneous correlation between return volatility and turnover has a positive median value. The upper bound is higher compared with the contemporaneous correlation between stock return and turnover, seen in Table 9, suggesting that this relationship is stronger for some companies. The potential relationship will be further investigated in the next section.

The correlation between lagged turnover and return volatility is for the majority of stocks positive, with relative low negative values compared to the positive ones. For lead turnover and return volatility the correlation is mostly negative for all but the first lead.

Thus Table 11 indicate that there might be a contemporaneous and/or causal relationship

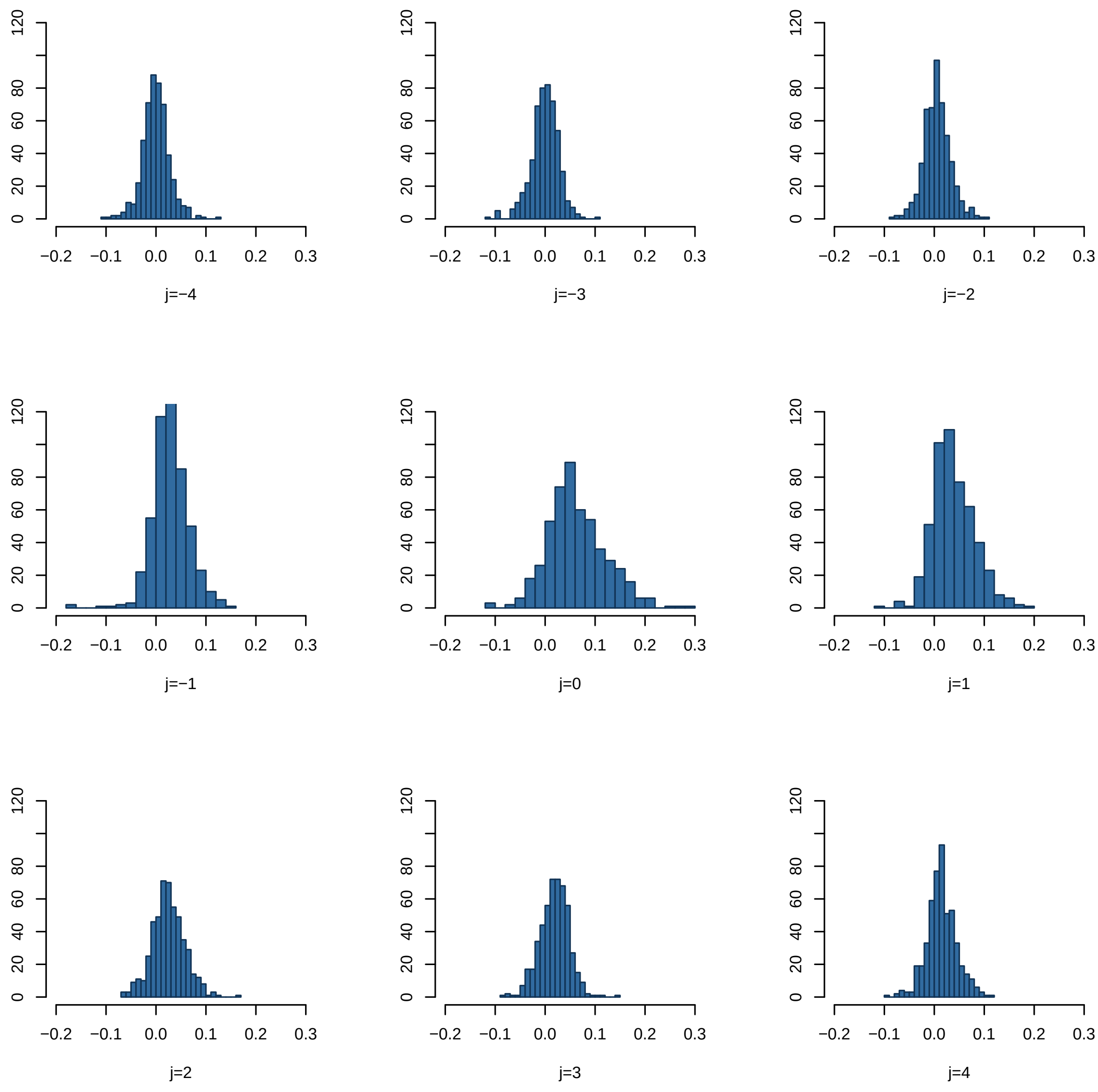

Figure 4: Histogram Log Volume – all stocks

between return volatility and turnover. The next sections will further investigate these findings.

Statistic	$j=-4$	$j=-3$	$j=-2$	$j=-1$	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
Max	0.23	0.23	0.23	0.24	0.35	0.41	0.41	0.24	0.25
Pctl(75)	0.02	0.02	0.02	0.04	0.10	0.10	0.07	0.05	0.05
Median	-0.01	-0.01	-0.01	0.003	0.04	0.04	0.02	0.01	0.01
Pctl(25)	-0.03	-0.03	-0.03	-0.02	0.01	-0.003	-0.01	-0.02	-0.02
Min	-0.13	-0.13	-0.13	-0.13	-0.09	-0.09	-0.10	-0.10	-0.11

Table 11: Cross-Correlation between Squared Return and detrended Turnover: $Corr(R_t^2,\ V_{t-j})$

Significant	$j=-4$	$j=-3$	$j=-2$	$j=-1$	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
Number	119	116	137	143	241	242	196	165	156
Percentage	23.56	22.97	27.13	28.32	47.72	47.92	38.81	32.67	30.89

Table 12: Securities with cross-correlation different from 0 at 5% significance level

Jointly, the findings in this section report stronger results for the simultaneous correlation between the variables than for the subsequent correlation. Although Table 9 and Table 11 report mostly a low correlation, this does not rule out any relationship.

6.4 Contemporaneous relationship

As the cross-correlation showed some correlation between return, return volatility, and turnover in the same period, we explore the contemporaneous relationship further.

6.4.1 Multivariate model

First, we test a multivariate model suggested by B.-S. Lee and Rui (2002), which studies the contemporaneous relationship between two time series variables. The model is applied by, among others, Mestel et al. (2003) and de Medeiros and Van Doornik (2006), and consist of the following two equations

$$R_t = b_0 + b_1 V_t + b_2 V_{t-1} + b_3 R_{t-1} + \varepsilon_t$$

$$V_t = a_0 + a_1 R_t + a_2 V_{t-1} + a_3 V_{t-2} + u_t$$

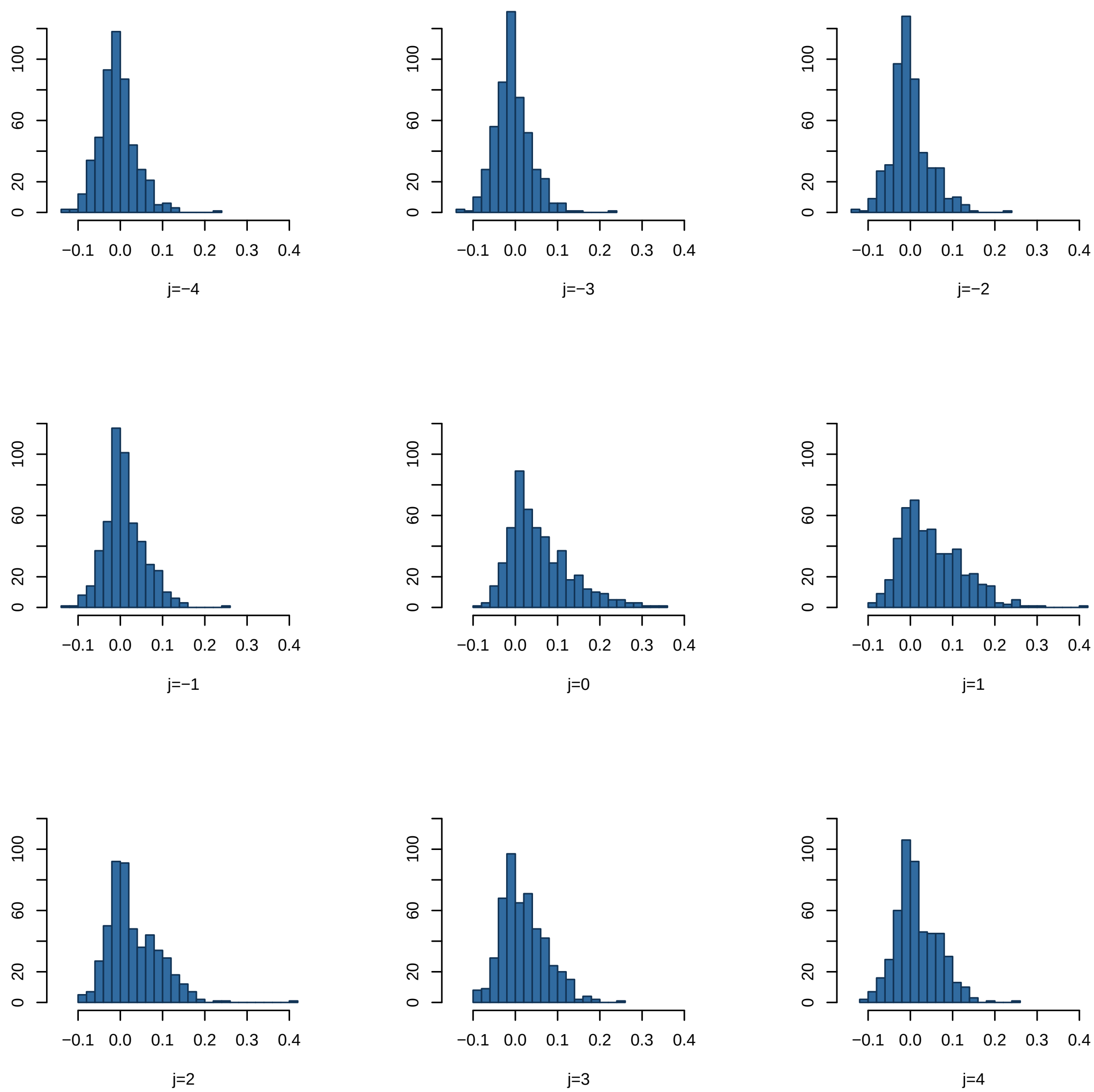

Figure 5: Cross-correlation volatility and volume

where V_t and R_t are the volume and the return at time t. Further, a_i and b_i are model coefficients for $i = 1, 2, 3$ and ε_t and u_t are white noise error terms.

According to economic theory and findings from other markets, trading activity is affecting return and return is affecting trading activity. Thus, we might have a simultaneous bias and problem with endogeny in the model. To avoid this, we estimate the simultaneous equation model using the two-stage least squares (2SLS) instrumental variable approach, which is a structural estimation used to establish whether a model derived from theory has a close fit to the sample data (Dion, 2008, p. 365). The 2SLS works by first regressing equation 1 and obtaining the fitted values for Return. Then these fitted values for Return from regression 1 are used as input to the second regression. A summary of the results can be found in Table 13 and Table 14, while Figure 6 and 7 shows the distribution of the t-statistics from the regressions. The t-statistics give an idea of the direction and the significance of the effect. The red bands in Figure 6 and 7 indicates a 5% significance level.

Statistic	b_0	b_1	b_2	b_3
Max	0.99	$1,034.47$	202.39	0.17
Pctl(75)	0.05	2.37	1.45	-0.03
Median	-0.01	1.26	0.21	-0.14
Pctl(25)	-0.08	0.40	-0.44	-0.24
Min	-0.61	-117.90	-582.72	-0.40

Table 13: $R_t = b_0 + b_1V_t + b_2V_{t-1} + b_3R_{t-1} + \varepsilon_t$

Statistic	a_0	a_1	a_2	a_3
Max	0.39	3.90	7.73	0.41
Pctl(75)	0.003	0.37	0.37	0.17
Median	0.0000	0.07	0.27	0.10
Pctl(25)	-0.003	0.004	0.16	0.03
Min	-0.28	-3.78	-2.87	-0.07

Table 14: $V_t = a_0 + a_1R_t + a_2V_{t-1} + a_3V_{t-2} + u_t$

At a 5% significant level, parameter b_1 was significant for 56.8% of the 505 stocks in sample. The parameter b_1 is positive for the majority of stocks, meaning that – all else equal – an increase in turnover will be accompanied by increased stock returns. The parameter b_2 was significant for 29.7% of the stocks, thus an increase in turnover will be followed by an increase in return for some companies. Lagged stock return b_3 was significant for 76.2%, hence yesterdays stock return have a significant effect on todays stock return.

Parameter a_1, is significant for 90.5% of all stocks. For almost all stocks it is then true

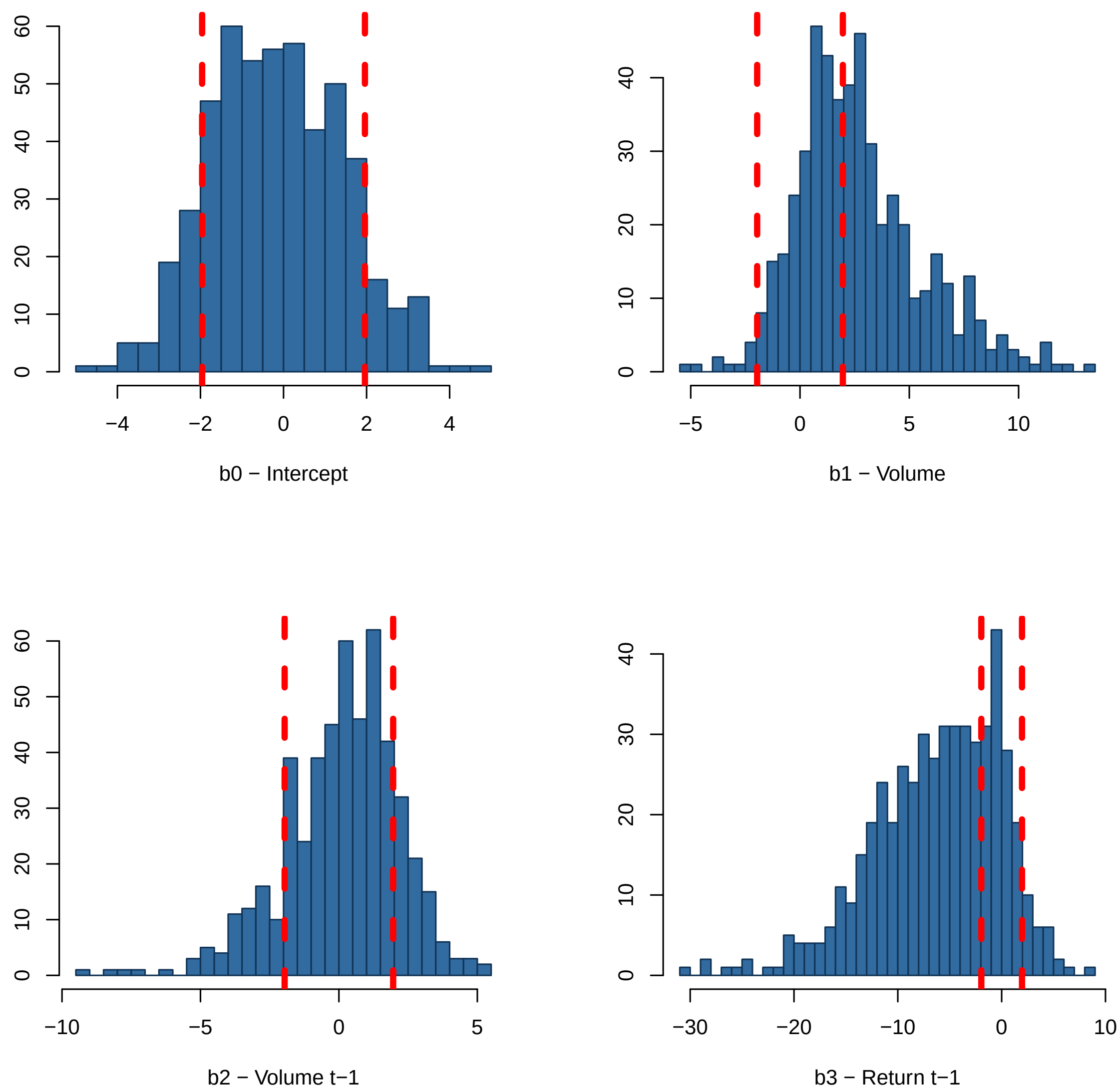

Figure 6: t-statistics from Lee & Rui's equation 1

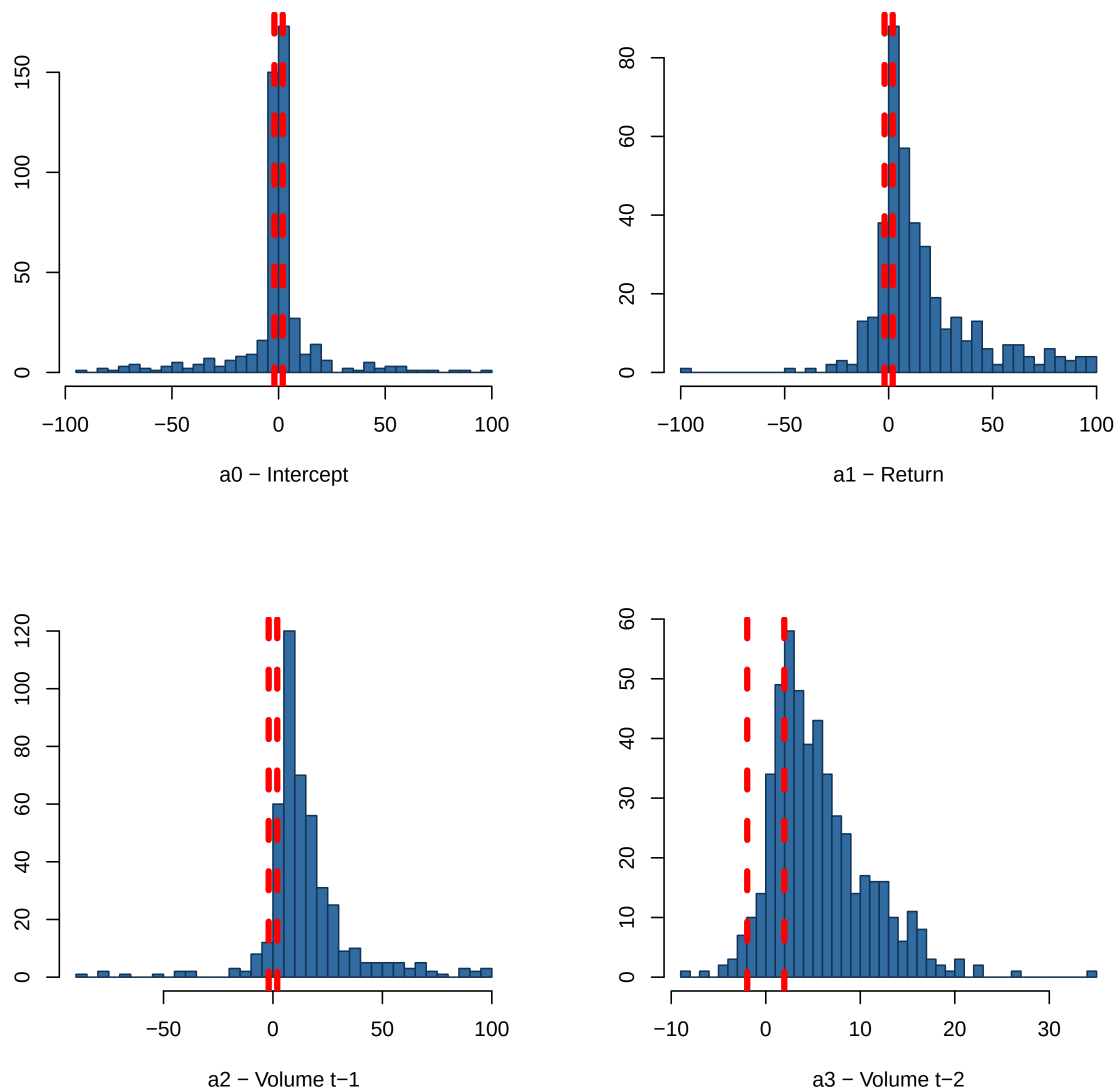

Figure 7: t-statistics from Lee & Rui's equation 2

that an increase in stock return go together with an increase in tturnover. Taking b_1 and a_1 together, it confirm the cross-correlation analysis that there are evidence that point towards a contemporaneous relation between stock return and turnover.

For lagged turnover, both a_2 and a_3 has a significant impact on turnover with 95.6 and 79.2% significant cases respectively. Hence a_2 and a_3 document strong time dependency for the turnover time series, consistent with the Ljung-Box results in Section 6.2.2.

Our results here provide evidence of a contemporaneous relationship between turnover and stock return, supported by earlier findings from the cross-correlation analysis. This is similar to B.-S. Lee and Ruis (2002) and de Medeiros and Van Doorniks (2006) findings from the US, UK, Japanese, and Brazilian stock market, but contradicting to what Mestel et al. (2003) found for the Austrian stock market. It is interesting to see that Return affect Volume and Volume affect Return on the Norwegian stock exchange.

6.4.2 Multivariate model with dummy

Empirical research often report that when trading acticity is high, price fluctuations tend to increase, especially in bullish markets. This suggests that there exist a relationship between higher order moments of stock return and trading activity. We check for this by means of another multivariate model.

The following model is an extension of the model by Brailsford (1996), among others used by Mestel et al. (2003, p. 9) and de Medeiros and Van Doornik (2006, p. 4). The model regress the contemporaneous relationship between turnover and volatility, using squared return as a proxy for volatility. As Brailsford (1996), we added a dummy variable to account for the degree of asymmetry. The regression is given by

$$V_t = \alpha_0 + \phi_1 V_{t-1} + \phi_2 V_{t-2} + \alpha_1 R_t^2 + \alpha_2 D_t R_t^2 + e_t$$

where D_t denotes a dummy variable that equals 1 if the corresponding return R_t is negative and 0 otherwise. V_t is the turnover at time t and R_t^2 is the squared return as a proxy for volatility. The parameter e_t is a white noise error term.

The lagged values of V_t up to lag 2 are included to avoid a problem with serially correlated residuals, as documented by Brailsford (1996).

With this model, we care mostly about catching the degree of asymmetry, and not about any potential endogeny.

A summary of the parameter variables can be seen in Table 15, and histograms of the t-statistics can be seen in Figure 8. Parameter ϕ_1 and ϕ_2 tells a similar story to what we saw in Section 6.4.1; turnover is highly dependent on past tturnover. At a 5% level, ϕ_1 was significant for 97.6% and ϕ_2 for 92.9% of the time series.

Statistic	α_0	ϕ_1	ϕ_2	α_1	α_2
Max	0.03	0.66	0.41	0.03	0.002
Pctl(75)	0.003	0.37	0.23	0.001	-0.0000
Median	0.001	0.30	0.18	-0.0000	-0.0004
Pctl(25)	-0.002	0.23	0.13	-0.0003	-0.002
Min	-0.16	-0.004	-0.06	-0.01	-0.03

Table 15: $V_t = \alpha_0 + \phi_1 V_{t-1} + \phi_2 V_{t-2} + \alpha_1 R_t^2 + \alpha_2 D_t R_t^2 + e_t$

The zero value of parameter α_1 tells us that turnover is unaffected by changes in volatility. This term is symmetric, so the effect is regardless of whether the stock returns are falling or increasing. The coefficient is significant at a 5% level for 59.2% of the stocks.

Parameter α_2 measures the asymmetry in the relationship. The negative parameter is significant in 41.2% of the cases, meaning that for around $2/5$ of the stocks in our sample there is an asymmetrical relationship between turnover and stock return. With a negative but small parameter and the given dummy specifications, turnover increases somewhat more when stock return increases than when stock return decreases. This is similar to Brailsford's (1996) findings that α_2 is generally negative but insignificant.

Mestel et al. (2003) and de Medeiros and Van Doornik (2006) have reported similar findings for the contemporaneous relationship between volume and volatility, as they report that increased prices induce more trading volume than price decrease.

6.4.3 Conditional volatility and trading volume

The alphabet soup of volatility models continually amazes.

– Robert Engle (2002)

As there are indications of a contemporaneous relationship between turnover and volatility, turnover might be a factor in the serial correlation of volatility. The contemporaneous relation-

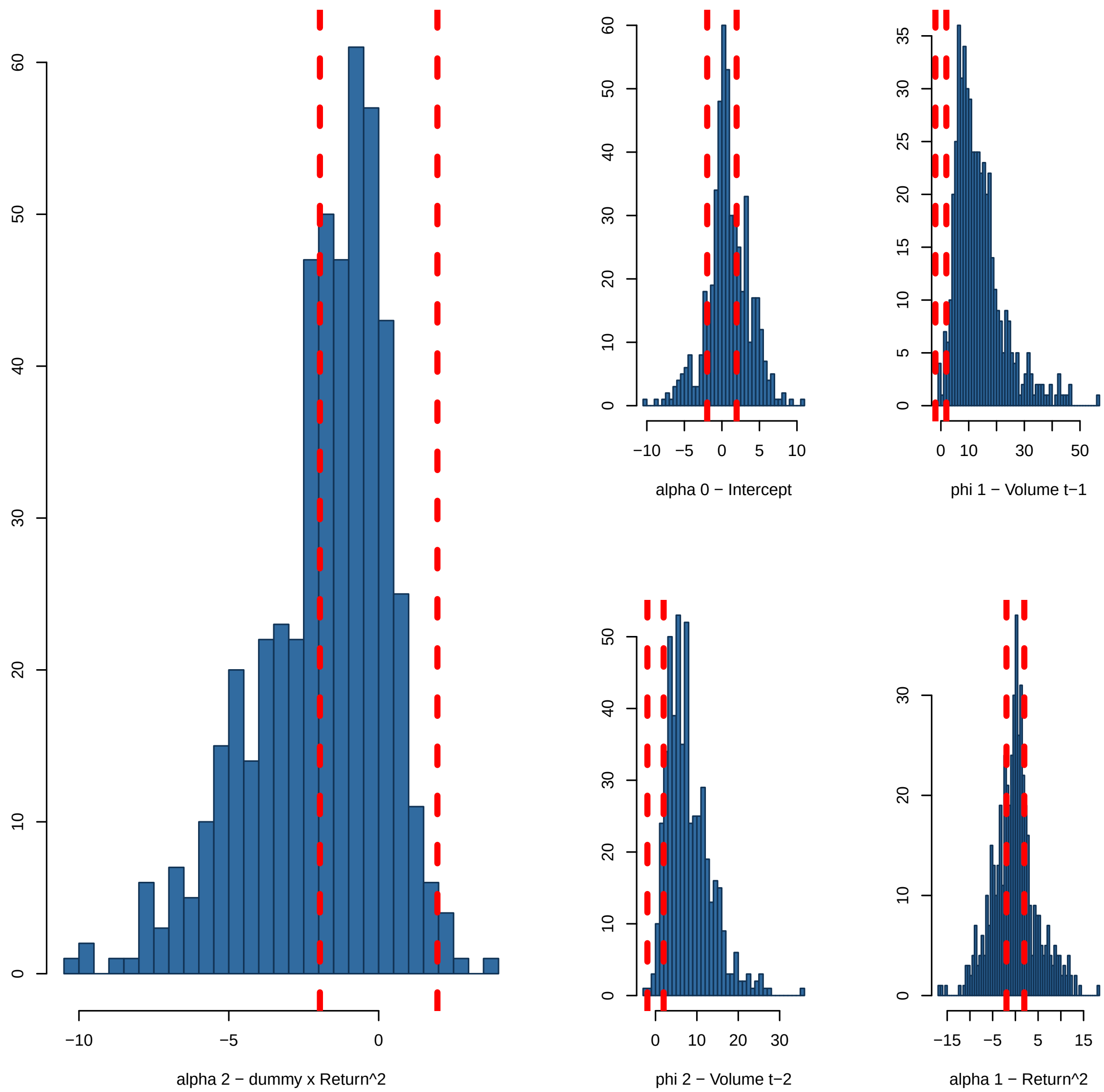

Figure 8: Histogram t-statistics

ship between turnover and volatility is of special interest as both the MDH and the SIAH use trading activity as a measure for the flow of information.

One of the stylized facts found in most financial time series is the clustering of volatility – or conditional heteroskedasticity[11]. The standard warning in the presence of heteroskedasticity is that the regression coefficients for an ordinary least squares (OLS) regression are still unbiased, but the standard errors and confidence intervals estimated by conventional procedures will be too narrow, giving a false sense of precision. Further, volatility – and the forecasting of it – is of great importance to financial economics as it among others is used as input in the Black-Scholes formula. Therefore, the autoregressive conditional heteroskedasticity (ARCH) model by Engle (1982), and its extension into the generalized ARCH (GARCH) model by Bollerslev (1986), are often used. Instead of considering the heteroskedasticity as a problem to be corrected, ARCH and GARCH models treat it as a variance to be modeled. The GARCH model, like the ARCH model, have a weighted average of past squared residuals, but includes declining weights that never reaches zero (Engle, 2001). Further expansions, such as the EGARCH model, were later developed as more evidence indicated that the direction of returns affect volatility (Engle, 2001, p. 166).

All ARCH-type models have been implemented in the analysis using the R package `rugarch` (Ghalanos, 2018).

6.4.4 GARCH(1,1)

The ARCH model by Engle (1982) was the first model of conditional heteroskedasticity. According to Engle (2004, p. 406), he was looking for a model that could assess the validity of the conjecture of Milton Friedman that the unpredictability of inflation was the primary cause of business cycles.

The ARCH model seeks to forecast the conditional variance by modeling it as an AR(q) process of earlier squared error terms.

[11]Volatility clustering: When a series exhibit some periods of low volatility and some periods of high volatility. If the variance is small in one period, it tend to be small in the next period as well, and vice versa. This implies that the series displays time-varying heteroskedasticity (Stock & Watson, 2015, p. 710, 712).

The full ARCH(q) model is given by

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t$$

$$\varepsilon_t | I_{t-1} \sim N(0, \sigma^2)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2$$

where the first equation is the mean model, the second gives the error distribution, and the last equation is the conditional variance model.

According to Engle (2002, pp. 425–426), it took years before the idea of ARCH took off, but when it did, one of the first extentions would also become one of the most influential – the GARCH model by Bollerslev (1986).

Empirically, the ARCH(q) model often require a very large value of q. Therefore, the GARCH model allow the conditional variance to be dependent upon its previous own lags. This is the same as modeling the conditional variance as an ARMA(p,q) process.

The full GARCH(p, q) model is given by

$$y_t = \phi_1 + \phi_2 x_{2,t} + \phi_3 x_{3,t} + \cdots + \phi_k x_{k,t} + \varepsilon_t$$

$$\varepsilon_t | I_{t-1} \sim N(0, \sigma^2)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \beta_2 \sigma_{t-2}^2 + \cdots + \beta_p \sigma_{t-p}^2$$

The parameter restrictions are $\alpha_0 > 0,\ \alpha_j \geq 0\ for\ j = 1, \ldots, q$ and $\beta_j \geq 0,\ for\ j = 1, \ldots, p$. Further, at least one β_j has to be strictly positive for the model specification to be a GARCH model.

A much applied specification of the GARCH model is the GARCH(1,1), which model the conditional variance as

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

In the GARCH(1,1) model, α_1 measures the reaction of conditional volatility to market shocks. Volatility is sensitive to market events if α_1 is relatively large (above 0.1). Parameter β_1 measures the persistence in conditional volatility irrespective of anything happening in the market. If β_1 is relatively large (above 0.9) it takes a long time for the volatility to die out following a crisis in the market (Alexandar, 2008).

The unconditional variance of a GARCH(1,1) model is given by

$$Var(\varepsilon_t) = \frac{\alpha_0}{1-(\alpha_1+\beta_1)},\ for\ (\alpha_1+\beta_1) < 1$$

Taken together, $(\alpha_1+\beta_1)$ decides the rate of convergence of the conditional volatility to the long term average level. When $(\alpha_1+\beta_1) > 1$ we have non-stationarity in variance, so the forecasted conditional variance will not converge to the unconditional variance as the horizon increase[12].

As we found that the return series displayed significant serial correlation of low order in Section 6.2.5, we model our return generating process as an AR(1) model. We use normally distributed errors and model the variance process as a GARCH(1,1) model with volume as an external variable. This is similar to models used by among others Mestel et al. (2003), Ahmed et al. (2005), and de Medeiros and Van Doornik (2006). Our model is given by

$$R_t = \phi_1 + \phi_2 R_{t-1} + \varepsilon_t$$

$$\varepsilon_t | I_{t-1} \sim N(0, \sigma^2)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \zeta_1 V_t$$

Models with volume as external regressor are sometimes referred to as *Volume Augmented* (VA) models, making this an AR(1)–VA–GARCH(1,1) model.

We will test two versions of this model: one version with the restriction $\zeta_1 = 0$, making this an AR(1)–GARCH(1,1) model, and one unrestricted version. The aim is to see whether persistence in volatility decreases when volume is included. The MDH predict that the GARCH effect will disappear when turnover is included in the model. If it does, it will be evidence in favor of the MDG being the correct hypothesis of how information flow into the market. However, if the persistence in volatility does not decrease noticeably, as predicted by the SIAH, it will be evidence in favor of this hypothesis.

Summary statistics of the restricted model can be found in Table 16 and summary statistics of the unrestricted model can be found in Table 17.

In the restricted version of the model, the α_0 was significant for 79.4% of the stocks, the α_1 was significant for 97.8% of the stocks, and the β_1 was significant for 100% of the stocks.

As the mean (median) α_1 is 0.10 (0.09), the conditional volatility is sensitive to market shocks, but the persistence irrespective of anything happening in the market, measured as β_1,

[12]The case where $(\alpha_1+\beta_1) = 1$ is termed integrated GARCH (IGARCH) and has given rise to a model of its own.

Statistic	α_1	β_1	$(\alpha_1 + \beta_1)$
Mean	0.10	0.86	0.96
Max	0.28	1.00	1.00
Pctl(75)	0.13	0.92	0.99
Median	0.09	0.88	0.98
Pctl(25)	0.07	0.83	0.95
Min	0.00	0.38	0.53

Table 16: Restricted model: AR(1)–GARCH(1,1)

is just below the relatively large threshold of 0.90 with its mean (median) value of 0.86 (0.88). However, although the persistence is not classified as large it is still strong. Together, $\alpha_1 + \beta_1$ has a mean (median) of 0.96 (0.98), just below the relatively strong definition of 0.99 (Alexandar, 2008).

Statistic	α_1	β_1	$(\alpha_1 + \beta_1)$
Mean	0.10	0.85	0.95
Max	0.37	1.00	1.14
Pctl(75)	0.13	0.92	0.99
Median	0.10	0.88	0.98
Pctl(25)	0.07	0.82	0.95
Min	0.00	0.00	0.05

Table 17: Unrestricted model: AR(1)–VA–GARCH(1,1)

In the unrestricted version of the model, the α_0 was significant for 70.7% of the stocks, α_1 was significant for 94.7% of the stocks, β_1 was significant for 98.2% of the stocks, and ζ_1 was significant for 2.2% of the stocks.

Although $(\alpha_1 + \beta_1)$ decreased for 45.5% of the stocks when we went from a restricted to an unrestricted model, adding the volume parameter to the model did not decrease the conditional volatilities reaction to market shocks. α_1 stayed more or less unchanged with a mean (median) value of 0.10 (0.10). The same goes for β_1, which has a mean (median) value of 0.85 (0.88), very similar to the restricted model. Furthermore, the turnover coefficient was only significant for 2.2% of the stocks. These findings are very similar to those of Ahmed et al. (2005), and provide evidence against the MDH.

In the unrestricted model, some of the regressions yielded $\alpha_1 + \beta_1 > 1$. This suggests that these processes are not covariance stationary. This is considered an undesirable trait, and could indicate a problem with our model. There could be several explanations for why we got these results. According to Teräsvirta (2009, p. 24), the standard GARCH model often exaggerates the persistence in volatility. Malmsten (2004, p. 13) and Shephard (1996, p 10, 14) report that the probability for estimating this persistence to be greater than one is substantial in small samples. This could be a problem for our shortest return series, where we have only 500 observations. However, it is unlikely to be a problem for the larger series, having daily observations for about 38 years. These series might suffer from another problem, as the assumption that GARCH-models have constant parameters might not be appropriate for such long samples (Mikosch & Stărică, 2004a, 2004b).

Further, the models could be misspecified. The variance process could perhaps be better explained by another GARCH-process, such as the commonly adopted GARCH(1,2) or GARCH(2,1) (Bollerslev, Chou, & Kroner, 1992, p. 22), or the mean process could be unsuited for being modeled as an AR(1)-process. The assumption of normally distributed errors could also be too simplistic, and Nelson (1991, p. 352) suggest using the Generalized Error Distribution (GED) instead. The GED contains the Normal distribution as a special case, but also allow for fatter tails.

Hamilton and Susmel (1994) argue that GARCH-models overestimate the persistence of volatility because they cannot describe large economic shocks properly. As described in Section 5.2, our sample spans the last 38 years and includes shocks such as the October 1987 crash and the 2007 financial crisis, which might make the GARCH(1,1)-model overestimate $(\alpha_1 + \beta_1)$.

Malmsten (2004, p. 13) found that if a GARCH(1,1) model is fitted to data generated by an exponential GARCH(1,1) process, there is a large probability of ending up with $\alpha_1 + \beta_1 \geq 1$. Thus, we decided to test an exponential GARCH model as well.

6.4.5 EGARCH(1,1)

The exponential GARCH (EGARCH) model was first developed by Nelson (1991) to accommodate his three criticisms of the GARCH model: the GARCH model does not allow for an asymmetric response to shocks, the GARCH model impose parameter restrictions that are often violated empirically, and interpreting whether shocks to the conditional variance "persist" or not is too hard in GARCH models.

The full EGARCH(p,q) model may be written as

$$y_t = \phi_1 + \phi_2 x_{2,t} + \phi_3 x_{3,t} + \cdots + \phi_k x_{k,t} + \varepsilon_t$$

$$\varepsilon_t | I_{t-1} \sim N(0, \sigma^2)$$

$$\log(\sigma_t^2) = \alpha_0 + \sum_{j=1}^{q} g_j(z_{t-j}) + \sum_{j=1}^{p} \beta_j \log(\sigma_{t-j}^2)$$

where $g_j(z_{t-j}) = \alpha_j z_{t-j} + \gamma_j \big(|z_{t-j}| - \ \mathbb{E}(|z_{t-j}|) \big)$ and $z_t = \frac{\varepsilon_t}{\sigma_t}$.

As with the GARCH model, the first order EGARCH is most often used in research (Malmsten & Teräsvirta, 2010, p. 447). The EGARCH(1,1) is specified as

$$\log(\sigma_t^2) = \alpha_0 + \alpha_1 z_{t-1} + \gamma_1 \big(|z_{t-1}| - \ \mathbb{E}(|z_{t-1}|) \big) + \beta_1 \log(\sigma_{t-1}^2)$$

Unlike the linear GARCH(1,1) model, there are no restrictions on the parameters α_1 and β_1 to ensure non-negativity of the conditional variances (Bollerslev et al., 1992, p 12).

In the EGARCH(1,1) model, γ_1 measures the magnitude effect – or the symmetric effect – of z_{t-1} on $\log(\sigma_t^2)$. All other equal, the effect is positive (negative) when the magnitude of z_{t-1} is larger (smaller) than its expected value.

The parameter α_1 measures the asymmetry in the relationship. If $\alpha_1 < 0$, then positive shocks generate less volatility than negative shocks, and so if $\alpha_1 > 0$ positive news are more destabilizing than negative news. If $\alpha_1 = 0$ then the model is symmetric. β_1 measures the persistence of shocks and corresponds to $(\alpha_1 + \beta_1)$ in the GARCH(1,1) model.

We test the following EGARCH(1,1) model

$$log(\sigma_t^2) = \alpha_0 + \alpha_1 z_{t-1} + \gamma_1 (|z_{t-1}| - \ \mathbb{E}(|z_{t-1}|)) + \beta_1 log(\sigma_{t-1}^2) + \zeta_1 V_t$$

where ζ_1 is restricted to equal zero in the first regression.

The summary of the restricted model can be found in Table 18.

We found α_0 to be significant for 92.5% of the stocks, α_1 to be significant for 61% of the stocks, β_1 to be significant for 99.8% of the stocks, and γ_1 to be significant for 97.2% of the stocks.

As mostly all stocks had a significant positive γ_1 and a significant negative α_1, we conclude

Statistic	α_0	α_1	β_1	γ_1
Mean	0.12	-0.04	0.94	0.19
Max	1.54	0.22	1.00	0.57
Pctl(75)	0.13	-0.02	0.98	0.24
Median	0.07	-0.03	0.96	0.18
Pctl(25)	0.03	-0.06	0.92	0.12
Min	-0.005	-0.23	0.45	-0.12

Table 18: Restricted exponential model: AR(1)–EGARCH(1,1)

that negative shocks have a higher impact on conditional volatility than positive shocks, all else equal.

Next, we wish to add volume to our model, to investigate how the parameters will change.

A summary of the unrestricted model can be found in Table 19.

Statistic	α_0	α_1	β_1	γ_1	ζ_1
Mean	0.26	-0.04	0.85	0.25	-4.81
Max	4.43	0.22	1.00	8.21	17.57
Pctl(75)	0.24	-0.01	0.98	0.27	-0.002
Median	0.09	-0.03	0.95	0.19	-0.11
Pctl(25)	0.04	-0.05	0.86	0.13	-0.70
Min	-0.01	-1.17	-0.44	-0.06	-100.00

Table 19: Unrestricted exponential model: AR(1)–VA–EGARCH(1,1)

We found α_0 to be significant for 91.1% of the stocks, α_1 to be significant for 51.9% of the stocks, β_1 to be significant for 97.2% of the stocks, γ_1 to be significant for 97.4% of the stocks, and ζ_1 to be significant for 52.7% of the stocks.

As before, mostly all stocks had a significant positive γ_1 and a significant negative α_1. Again, we conclude that negative shocks have a higher impact on conditional volatility than positive shocks, all else equal. As turnover was added, the mean of α_1 stayed the same, but γ_1 increased.

We note that parameter β_1 has a lower value and thus the persistence of shocks have a weakened effect on the conditional volatility when volume is included in the model. The β_1 declined for 71.1% of the stocks when we included volume in the model. Also, the mean of the constant parameter α_0 more than doubled in absolute value, while the median only changed somewhat.

As the effect of past conditional volatility on present conditional volatility is of great interest, we calculated the half-life of β_1 using the formula

$$HL = \frac{\log(0.5)}{\log|\beta_1|}$$

Our findings are summarized in Table 20. We found that the median persistence was reduced from 19.30 to 12.56 for half-value when turnover was included. This means that some of the persistence in volatility attributed to β_1 in the restricted model can be explain by the flow of information, as proxied by turnover.

Statistic	Restricted	Unrestricted
Max	$41,017,099.00$	$65,085,961.00$
Pctl(75)	37.29	29.14
Median	19.30	12.56
Pctl(25)	8.80	4.51
Min	0.87	0.08

Table 20: Summary statistic half-life

6.5 Causal relationship

If there is a causal or dynamical relationship between two variables, one variable tend to influence the other. This type of relationship is of special interest as the notion that one can use todays values of x to predict the future values of y has been heavily debated over the years. In the cross-correlation analysis earlier in this chapter, we found the first signs that there might exists a causal relationship between stock return, volatility and turnover.

6.5.1 Granger causality

The Granger causality test – developed by Granger (1969) – is often used to study the dynamic relationship between two variables and assess in which direction the relationship between them are going. Granger causality has nothing to do with what we normally mean by causality – it is a predictive relation, not a causal one. The variable x is said to Granger-cause the variable y if y can be significantly better predicted using the historical values of both y and x than it can by using only past values of y.

More technically, this can be written as

$$\mathbb{E}(y_t|I_{t-1}) \neq \mathbb{E}(y_t|J_{t-1})$$

where information set I_{t-1} contains information on y and x while J_{t-1} contains only information on past values of y (Wooldridge, 2016, p. 590).

The test is based on the bivariate VAR model

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{p} \beta_i x_{t-i} + \varepsilon_t$$

$$x_t = \gamma_0 + \sum_{i=1}^{p} \gamma_i x_{t-i} + \sum_{i=1}^{p} \delta_i y_{t-i} + \zeta_t$$

which is split into one restricted and one unrestricted version. In the restricted version $\beta_i = 0$ for $i = 1, \ldots, p$ and $\delta_i = 0$ for $i = 1, \ldots, p$.

The null hypothesis and alternative hypothesis for the first equation is defined as

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_a : \beta_i \neq 0 \;\; for\ at\ least\ one\ i = 1, 2, \ldots, p$$

There are several statistical tests one can use to test these hypotheses. Geweke, Meese, and Dent (1983) found that the Wald variants are the most accurate.

We have used the package `lmtest` (Zeileis & Hothorn, 2002) in R to perform the Granger causality test. According to Arratia (2014, p. 79), this function use a Wald test statistic introduced by Toda and Yamamoto (1995) which follows an asymptotically chi-square distribution under the null hypothesis. This is regardless of whether y is stationary or not (Toda & Yamamoto, 1995, p. 230).

We do our investigation of the causal relationship between stock return, volatility and turnover by using a Granger causality test applying a bivariate VAR model of order p. Order p was found by the AIC and BIC for every individual stock.

To select the number of lags or parameters in a model, an information criterion is often applied. We have used the package `vars` (Pfaff, 2008) in R, which provide the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)[13].

For an AR(n) regression including a constant term, T observations and k coefficients, the

[13]Refrenced as Schwarz criterion (SC) in the R package `vars` (Pfaff, 2008).

AIC is given as

$$AIC(n) = \log[\det(\tilde{\Sigma}_u(n))] + k(kn+1)\frac{2}{T}$$

and the BIC is given by

$$BIC(n) = \log[\det(\tilde{\Sigma}_u(n))] + k(kn+1)\frac{\log(T)}{T}$$

where $\tilde{\Sigma}_u$ is the estimated $k \times k$ covariance matrix of the errors from an AR(n) regression, such that the i, j element of $\tilde{\Sigma}_u$ is $\frac{1}{T}\sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt}$, where $\hat{u}_{it}$ and $\hat{u}_{jt}$ are the OLS residuals from the i^{th} and j^{th} equation respectively.

The BIC is very similar to the AIC, but penalize additional parameters somewhat more, and is in that sense stricter.

As the AIC return a very high p for most stocks, we ended up trusting the somewhat stricter BIC, which suggested a VAR(5) model.

The relationship between stock return and tturnover as well as stock volatility and turnover, have been run with 5 lags, and reported at a 5% level. A summary can be found in Table 21.

Direction	% significant
$R \xrightarrow{\text{G.c.}} V$	20.2%
$V \xrightarrow{\text{G.c.}} R$	29.3%
$R^2 \xrightarrow{\text{G.c.}} V$	30.1%
$V \xrightarrow{\text{G.c.}} R^2$	37.6%

Table 21: Granger causality

The data show a weak relationship between return and turnover, where return Granger-cause turnover in only 20.2% of the cases. Thus, one cannot use stock return to predict volume for the majority of stocks at OSE.

In the other direction, we find that volume Granger-cause stock return for 29.3% of the stocks. One can therefore state that volume precedes stock return to a greater extent than stock return precedes volume. As earlier empirical finding have been inconsistent, see Table 2, our findings are not surprising and somewhat similar to Chen et al. (2001).

Running the same Granger test for return volatility and turnover, we find that return volatility Granger cause turnover in 30.1% of the cases. In 37.6% of the cases, turnover Granger cause return volatility.

Also here, turnover comes before return volatility more than the opposite. However, the fact that the Granger effect of turnover on return volatility is significant for just a bit more stocks than the Granger effect of turnover on stock return surprised us. This is due to the majority of earlier finding, summarized in Table 2, who found that the Granger effect of volume on volatility to be more present.

What we have found is that turnover Granger causes return and squared return for approximately 30-40% of the stocks. Thus, turnover has a stronger Granger effect on stock return and volatility than vice versa, which is consistent with earlier findings in the cross-correlation analysis.

As turnover have a Granger effect on stock return and volatility, this can be interpreted as a sign that the weak-form market efficiency does not hold. Also, it seems that arrival of information follows a sequential rather than simultaneous process.

6.6 Robustness check

We have performed the same analysis using the number of trades each day for each stock to check whether our results would be the same as when using turnover.

The few differences we accounted in the analysis was:

1. Due to the large range of the volume data, we decided to take the logarithms in Section 6.2.2.
2. After logging, we did not remove outliers.
3. More companies were found to have a unit root by the ADF test in Section 6.2.6, and was thus removed. This reduced our sample to 483 companies.
4. After logging and removing companies, we did not remove a linear trend.

More companies were found to have a unit root by the ADF test in Section 6.2.6, and was thus removed.

The result of our analysis using number of shares traded was very similar to when we used turnover, and all the sub-conclusion were the same. The results are not included for the sake of brevity and the lack of standardization between securities. The full results are available upon request.

7 Conclusion

In this thesis we examined the empirical relationship between trading volume and stock return on Oslo Stock Exchange. We have done so by means of a cross-correlation analysis, multivariate regressions, GARCH and EGARCH models, and a Granger causality test.

We found evidence of a positive contemporaneous relationship between return and volume, detected by cross-correlation and multivariate regressions. However, our results indicate that this relationship is rather weak. By a two-stage least squares estimation we found that volume had a significant contemporaneous effect on returns in 56.8% of the stocks on OSE, when controlling for lagged values of both volume and return. Further, we found return to have a significant contemporaneous effect on volume in 90.5% of the stocks, when controlling for lagged values of volume. The existence of a contemporaneous relationship between volume and return is in accordance with what one would expect to find if the mixture of distribution hypothesis is true.

Further, we found evidence of a contemporaneous relationship between trading volume and return volatility. The cross-correlation showed a weak but mostly positive contemporaneous relationship. Our multivariate model shows that trading volume is unaffected when volatility increase, regardless of whether the stock return is falling or increasing. This zero symmetric term was found to be significant for 59.2% of the stocks. When accounting for asymmetry in the relationship, we find that volume increase more when returns are positive than when they are negative. The asymmetric effect is significant for 41.2% of the stock at OSE, which is in line with Brailsford's (1996) findings.

By our GARCH(1,1) model we found weak evidence that the persistence in conditional volatility decrease when one includes trading volume as a proxy for information arrival. The coefficients for persistence decreased for 45.6% of the stocks, however the mean value only decreased from 0.96 to 0.95. We concluded that a GARCH model might not be the optimal model, as about 1/4 of the stocks displayed non-stationarity of variance with this model, and decided to test an EGARCH(1,1) also. We found that persistence decreased for 71.1% of the stocks when including trading volume, and that the median half-life of shocks decreased from 19 to 13 days.

We also found evidence of a dynamic relationship. In the cross-correlation analysis we found a positive relationship between return and lag/lead values of volume, however this correlation was weaker than the contemporaneous effect. By a two-stage least squares estimation we found that lagged volume had a significant effect on returns in 29.7% of the stocks on OSE, when controlling for a contemporaneous relationship and lagged values of return. When it comes to

Granger causality we found evidence of a dynamic relationship between return and volume in both directions. We found that return Granger cause volume in 20.2% of the stocks, while the relationship is much stronger in the opposite direction as volume Granger cause return 29.3% of stocks. In the case with volatility and volume we find that volatility Granger cause volatility in 30.1% of the cases, while – again – the other direction is stronger, as volume Granger cause volatility in 37.6% of the cases.

As we found both a contemporaneous and a causal relationship, this lend greater support to the sequential information arrival hypothesis than the mixture of distribution hypothesis. This means that there is some information inefficiency on Oslo Stock Exchange. As in an efficient market, prices already reflect everything that have already occurred and events the market expects to take place in the future, our results lends further credibility to the adaptive market hypothesis and the heterogeneous agents model rather than the efficient market hypothesis.

8 Review of thesis

8.1 Limitations and further research

We found evidence of a contemporaneous and causal relationship between volume and return at Oslo Stock Exchange. However, there are some limitations of our thesis. First, we do not look at cross-sectional differences between the stocks. Maybe the relationship is stronger or more evident for some type of stocks than for others. Second, we did not look at different subsamples in time. Several sources state that these types of findings might be sensitive to the time period. Maybe the relationships were more evident in the 80s than in the 2010s.

Based on this, we have the following suggestion for further research. It would be interesting to study both the cross-sectional and the time varying relationship between volume and return at Oslo Stock Exchange. It would also be very interesting to study the relationship for a short time period but with high frequency data. Further, it would be interesting to follow Wang et al. (2018) and look at asymmetries in tail dependencies in positive and negative dependence regimes.

References

Abbondante, P. (2010). Trading volume and stock indices: A test of technical analysis. *American Journal of Economics and Business Administration*, *2*(3), 287–292.

Abrol, S., Chesir, B., & Mehta, N. (2016). High frequency trading and us stock market microstructure: A study of interactions between complexities, risks and strategies residing in us equity market microstructure. *Financial Markets, Institutions & Instruments*, *25*(2), 107–165.

Admati, A. R., & Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *The Review of Financial Studies*, *1*(1), 3–40.

Ahmed, H. J. A., Hassan, A., & Nasir, A. M. (2005). The relationship between trading volume, volatility and stock market returns: A test of mixed distribution hypothesis for a pre and post crisis on kuala lumpur stock exchange. *Investment Management and Financial Innovations*, *3*(3), 146–158.

Alexandar, C. (2008). *Market risk analysis: Practical financial econometrics (vol. ii)*. Wiley Publishing.

Alfons, A. (2016). robusthd: Robust methods for high-dimensional data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=robustHD` (R package version 0.5.1)

Amihud, Y., & Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of financial Economics*, *17*(2), 223–249.

Amihud, Y., Mendelson, H., & Murgia, M. (1990). Stock market microstructure and return volatility: Evidence from italy. *Journal of Banking & Finance*, *14*(2-3), 423–440.

Andersen, T. G. (1996). Return volatility and trading volume: An information flow interpretation of stochastic volatility. *The Journal of Finance*, *51*(1), 169–204.

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, 885–905.

Anthonisz, S. A., & Putniņš, T. J. (2016). Asset pricing with downside liquidity risks. *Management Science*, *63*(8), 2549–2572.

Arratia, A. (2014). *Computational finance: An introductory course with r* (Vol. 1). Springer Science & Business Media.

Banerjee, A., Dolado, J. J., Galbraith, J. W., Hendry, D., et al. (1993). Co-integration, error correction, and the econometric analysis of non-stationary data. *OUP Catalogue*.

Benston, G. J., & Hagerman, R. L. (1974). Determinants of bid-asked spreads in the over-the-counter market. *Journal of Financial Economics*, *1*(4), 353–364.

Black, F. (1986). Noise. *The journal of finance*, *41*(3), 528–543.

Blume, L., Easley, D., & O'hara, M. (1994). Market statistics and technical analysis: The role of volume. *The Journal of Finance*, *49*(1), 153–181.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307–327.

Bollerslev, T., Chou, R. Y., & Kroner, K. F. (1992). Arch modeling in finance: A review of the theory and empirical evidence. *Journal of econometrics*, *52*(1-2), 5–59.

Borchers, H. W. (2018). pracma: Practical numerical math functions [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=pracma` (R package version 2.1.4)

Boulatov, A., Hatch, B. C., Johnson, S. A., & Lei, A. Y. (2009). Dealer attention, the speed of quote adjustment to information, and net dealer revenue. *Journal of Banking & Finance*, *33*(8), 1531–1542.

Brailsford, T. J. (1996). The empirical relationship between trading volume, returns and volatility. *Accounting & Finance*, *36*(1), 89–111.

Brock, W. A., & LeBaron, B. D. (1995). *A dynamic structural model for stock return volatility and trading volume* (Tech. Rep.). National Bureau of Economic Research.

Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies*, *27*(8), 2267–2306.

Campbell, J. Y., Grossman, S., & Wang, J. (1993). Trading volume and serial correlation in stock returns. *The Quarterly Journal of Economics*, *108*(4), 905–939.

Chandrapala, P. (2011). The relationship between trading volume and stock returns. *Journal of Competitiveness*, *3*(3).

Chen, G.-m., Firth, M., & Rui, O. M. (2001). The dynamic relation between stock returns, trading volume, and volatility. *Financial Review*, *36*(3), 153–174.

Choi, K.-H., Jiang, Z.-H., Kang, S. H., & Yoon, S.-M. (2012). Relationship between trading volume and asymmetric volatility in the korean stock market. *Modern Economy*, *3*(05), 584.

Chordia, T., Roll, R., & Subrahmanyam, A. (2002). Order imbalance, liquidity, and market returns. *Journal of Financial economics*, *65*(1), 111–130.

Chordia, T., Roll, R., & Subrahmanyam, A. (2011). Recent trends in trading activity and market quality. *Journal of Financial Economics*, *101*(2), 243–263.

Chordia, T., Subrahmanyam, A., & Anshuman, V. R. (2001). Trading activity and expected stock returns. *Journal of Financial Economics*, *59*(1), 3–32.

Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica: journal of the Econometric Society*, 135–155.

Copeland, T. E. (1976). A model of asset trading under the assumption of sequential information arrival. *The Journal of Finance*, *31*(4), 1149–1168.

Copeland, T. E. (1977). A probability model of asset trading. *Journal of Financial and Quantitative Analysis*, *12*(4), 563–578.

Cremers, K. M., & Mei, J. (2007). Turning over turnover. *The Review of Financial Studies*, *20*(6), 1749–1782.

Cryer, J. D., & Chan, K.-S. (2008). *Time series analysis: With applications in r*. Springer Science & Business Media.

Daniel, K., & Titman, S. (1999). Market efficiency in an irrational world. *Financial Analysts Journal*, *55*(6), 28–40.

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning* (Vol. 479). John Wiley & Sons.

Davidson, R., & MacKinnon, J. G. (1999). *Foundations of econometrics.* Oxford Press.

de Jonge, E., & van der Loo, M. (2013). An introduction to data cleaning with r. *Heerlen: Statistics Netherlands*.

de Medeiros, O. R., & Van Doornik, B. F. (2006). The empirical relationship between stock returns, return volatility and trading volume in the brazilian stock market.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366a), 427–431.

Dion, P. A. (2008). Interpreting structural equation modeling results: a reply to martin and cullen. *Journal of Business Ethics*, *83*(3), 365–368.

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.

Engle, R. (2001). Garch 101: The use of arch/garch models in applied econometrics. *Journal of economic perspectives*, *15*(4), 157–168.

Engle, R. (2002). New frontiers for arch models. *Journal of Applied Econometrics*, *17*(5), 425–446.

Engle, R. (2004). Risk and volatility: Econometric models and financial practice. *American economic review*, *94*(3), 405–420.

Epps, T. W., & Epps, M. L. (1976). The stochastic dependence of security price changes and transaction volumes: Implications for the mixture-of-distributions hypothesis. *Econometrica: Journal of the Econometric Society*, 305–321.

Fama, E. F. (1965). Random walks in stock market prices. *Financial Analysts Journal*, 55–59.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), 383–417.

Fama, E. F., & Blume, M. E. (1966). Filter rules and stock-market trading. *The Journal of Business*, *39*(1), 226–241.

Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International economic review*, *10*(1), 1–21.

Fan, J., & Yao, Q. (2017). *The elements of financial econometrics*. Cambridge University Press.

Floros, C., & Vougas, D. (2007). Trading volume and returns relationship in greek stock index futures market: Garch vs. gmm. *International Research Journal of Finance and Economics*(12), 98–115.

Frankel, J. A., & Froot, K. A. (1990). Chartists, fundamentalists, and trading in the foreign exchange market. *The American Economic Review*, *80*(2), 181–185.

Friederich, S., & Payne, R. (2015). Order-to-trade ratios and market liquidity. *Journal of Banking & Finance*, *50*, 214–223.

Gagnon, L., & Karolyi, G. A. (2009). Information, trading volume, and international stock return comovements: Evidence from cross-listed stocks. *Journal of Financial and Quantitative Analysis*, *44*(4), 953–986.

Gallant, A. R., Rossi, P. E., & Tauchen, G. (1992). Stock prices and volume. *The Review of Financial Studies*, *5*(2), 199–242.

Gallo, G. M., & Pacini, B. (2000). The effects of trading activity on market volatility. *The European Journal of Finance*, *6*(2), 163–175.

Gavrilov, I., & Pusev, R. (2014). normtest: Tests for normality [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=normtest` (R package version 1.1)

Geweke, J., Meese, R., & Dent, W. (1983). Comparing alternative tests of causality in temporal systems: Analytic results and experimental evidence. *Journal of Econometrics*, *21*(2), 161–194.

Ghalanos, A. (2018). rugarch: Univariate garch models. [Computer software manual]. (R package version 1.4-0.)

Gillespie, C., & Lovelace, R. (2016). *Efficient r programming*. O'Reilly Media, Incorporated.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.

Granger, C. W., & Morgenstern, O. (1963). Spectral analysis of new york stock market prices. *Kyklos*, *16*(1), 1–27.

Greene, W. (2012). *Econometric analysis: International edition: 7th edition*. Pearson Education Limited.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25.

Grossman, S. (1976). On the efficiency of competitive stock markets where trades have diverse information. *The Journal of finance*, *31*(2), 573–585.

Grossman, S., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American economic review*, *70*(3), 393–408.

Hagströmer, B., & Norden, L. (2013). The diversity of high-frequency traders. *Journal of Financial Markets*, *16*(4), 741–770.

Hamilton, J. D., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of econometrics*, *64*(1-2), 307–333.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). Springer.

Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, *66*(1), 1–33.

Hlavac, M. (2018). stargazer: Well-formatted regression and summary statistics tables [Computer software manual]. Bratislava, Slovakia. Retrieved from `https://CRAN.R-project.org/package=stargazer` (R package version 5.2.1)

Hodne, F., & Grytten, O. H. (1992). *Norsk økonomi 1900-1990*. Tano.

Hodne, F., & Grytten, O. H. (2000). Norsk økonomi i det 19. århundre. *Oslo: Fagbokforlaget*.

Hwang, S., & Valls Pereira, P. L. (2006). Small sample properties of garch estimates and persistence. *The European Journal of Finance*, *12*(6-7), 473–494.

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, 163–172.

Jennings, R. H., Starks, L. T., & Fellingham, J. C. (1981). An equilibrium model of asset trading with sequential information arrival. *The Journal of Finance*, *36*(1), 143–161.

Jensen, M. (1978). Some anomalous evidence regarding market efficiency.

Johnson, N., Zhao, G., Hunsader, E., Meng, J., Ravindar, A., Carran, S., & Tivnan, B. (2012). Financial black swans driven by ultrafast machine ecology. *arXiv preprint arXiv:1202.1448*.

Jørgensen, K., Skjeltorp, J., & Ødegaard, B. A. (2017). Throttling hyperactive robots–order-to-trade ratios at the oslo stock exchange. *Journal of Financial Markets*.

Kamath, R. R., & Wang, Y. (2008). The price-volume relationship in the chilean stock market. *International Business & Economics Research Journal*, *7*(10), 7–14.

Karolyi, G. A., Lee, K.-H., & Van Dijk, M. A. (2009). Commonality in returns, liquidity, and turnover around the world. *Ohio State University. Processed*.

Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. *Journal of Financial and quantitative Analysis*, *22*(1), 109–126.

Khan, J. A., Van Aelst, S., & Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, *102*(480), 1289–1299.

Kirilenko, A. A., & Lo, A. W. (2013). Moore's law versus murphy's law: Algorithmic trading and its discontents. *Journal of Economic Perspectives*, *27*(2), 51–72.

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with r*. Springer Science & Business Media.

Kozak, M. (2009). What is strong correlation? *Teaching Statistics*, *31*(3), 85–86.

Kristiania børs. (1919). *Kristiania børs 1819-1919 : et tilbakeblik ved 100 aars jubilæet*. Kristiania: Komiteen.

Kumar, B., Singh, P., & Pandey, A. (2009). The dynamic relationship between price and trading volume: Evidence from indian stock market.

Lamoureux, C. G., & Lastrapes, W. D. (1990). Heteroskedasticity in stock return data: Volume versus garch effects. *The journal of finance*, *45*(1), 221–229.

Leal, S. J., Napoletano, M., Roventini, A., & Fagiolo, G. (2016). Rock around the clock: an agent-based model of low-and high-frequency trading. *Journal of Evolutionary Economics*, *26*(1), 49–76.

Lee, B.-S., & Rui, O. M. (2002). The dynamic relationship between stock returns and trading volume: Domestic and cross-country evidence. *Journal of Banking & Finance*, *26*(1), 51–78.

Lee, C. F., & Rui, O. M. (2000). Does trading volume contain information to predict stock returns? evidence from china's stock markets. *Review of Quantitative Finance and Accounting*, *14*(4), 341–360.

Lim, K.-P., & Brooks, R. (2011). The evolution of stock market efficiency over time: a survey of the empirical literature. *Journal of Economic Surveys*, *25*(1), 69–108.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

Llorente, G., Michaely, R., Saar, G., & Wang, J. (2002). Dynamic volume-return relation of individual stocks. *The Review of Financial Studies*, *15*(4), 1005–1047.

Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective.

Lo, A. W. (2005). Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis.

Lo, A. W. (2017). *Adaptive markets: Financial evolution at the speed of thought*. Princeton University Press.

Lo, A. W., & Wang, J. (2000). Trading volume: definitions, data analysis, and implications of portfolio theory. *The Review of Financial Studies*, *13*(2), 257–300.

Lu, W.-C., & Lin, F.-J. (2010). An empirical study of volatility and trading volume dynamics using high-frequency data.

Malmsten, H. (2004). *Evaluating exponential garch models* (Tech. Rep.). SSE/EFI Working paper Series in Economics and Finance.

Malmsten, H., & Teräsvirta, T. (2010). Stylized facts of financial time series and three popular models of volatility. *European Journal of pure and applied mathematics*, *3*(3), 443–477.

Menkhoff, L. (2010). The use of technical analysis by fund managers: International evidence. *Journal of Banking & Finance*, *34*(11), 2573–2586.

Mestel, R., Gurgul, H., & Majdosz, P. (2003). The empirical relationship between stock returns, return volatility and trading volume on the austrian stock market. *University of Grazbr,*

Institute of Banking and Finance, Research Paper.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017). e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=e1071` (R package version 1.6-8)
Mikalsen, S. (2014). *Aksjer og aksjehandel : hvordan lykkes på børsen.* Oslo: Gyldendal akademisk.
Mikosch, T., & Stărică, C. (2004a). Changes of structure in financial time series and the garch model. *Revstat Statistical Journal*, *2*(1), 41–73.
Mikosch, T., & Stărică, C. (2004b). Nonstationarities in financial time series, the long-range dependence, and the igarch effects. *Review of Economics and Statistics*, *86*(1), 378–390.
Mjølhus, J. O. (2010). *Finansmarkeder*. Oslo: Cappelen akademisk.
Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica: Journal of the Econometric Society*, 315–335.
Müller, K., & Wickham, H. (2017). tibble: Simple data frames [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=tibble` (R package version 1.4.1)
Næs, R., & Ødegaard, B. A. (2009). Liquidity and asset pricing: Evidence on the role of investor holding period.
Næs, R., Skjeltorp, J. A., & Ødegaard, B. A. (2008). Liquidity at the oslo stock exchange.
Næs, R., Skjeltorp, J. A., & Ødegaard, B. A. (2011). Stock market liquidity and the business cycle. *The Journal of Finance*, *66*(1), 139–176.
Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 347–370.
Ødegaard, B. A. (2017). Bond liquidity at the oslo stock exchange.
Ødegaard, B. A. (2018). *Empirics of the oslo stock exchange: Basic, descriptive, results 1980–2017* (Tech. Rep.). University of Stavanger.
O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, *116*(2), 257–270.
Ongena, S., Smith, D. C., & Michalsen, D. (2003). Firms and their distressed banks: lessons from the norwegian banking crisis. *Journal of Financial Economics*, *67*(1), 81–112.
Osborne, M. F. (1959). Brownian motion in the stock market. *Operations research*, *7*(2), 145–173.
Pástor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political economy*, *111*(3), 642–685.
Pástor, L., Stambaugh, R. F., & Taylor, L. A. (2017). Do funds make more when they trade more? *The Journal of Finance*, *72*(4), 1483–1528.
Pfaff, B. (2008). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, *27*(4). Retrieved from `http://www.jstatsoft.org/v27/i04/`
Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*(2), 335–346.
PricewaterhouseCoopers, L. (2015). Global financial markets liquidity study. *report prepared for the Global Financial Markets Association and the Institute of International Finance.[Links]*.
Qiu, D. (2015). atsa: Alternative time series analysis [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=aTSA` (R package version 3.1.2)
R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`
Ruppert, D., & Matteson, D. S. (2015). *Statistics and data analysis for financial engineering: with r examples*. Springer.
Sabbaghi, O. (2011). Asymmetric volatility and trading volume: The g5 evidence. *Global Finance Journal*, *22*(2), 169–181.
Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *IMR; Industrial Management Review (pre-1986)*, *6*(2), 41.
Serrano-Padial, R. (2010). no trade theorems. *The New Palgrave Dictionary of Economics*, *4*.
Shephard, N. (1996). Statistical aspects of arch and stochastic volatility.

Skjeltorp, J. A., Ødegaard, B. A., et al. (2009). The information content of market liquidity: An empirical analysis of liquidity at the oslo stock exchange. *UiS Working Papers in Economics and Finance, University of Stavan-ger*(35).

Steigum, E. (2010). Norsk økonomi etter 1980–fra krise til suksess.

Stock, J. H., & Watson, M. W. (2015). *Introduction to econometrics: Global edition (updated third edition)*. Pearson Education.

Stoll, H. R. (1978). The pricing of security dealer services: An empirical study of nasdaq stocks. *The Journal of Finance*, *33*(4), 1153–1172.

Taleb, N. N. (2018). *Skin in the game: Hidden asymmetries in daily life*. Random House.

Taylor, M. P., & Allen, H. (1992). The use of technical analysis in the foreign exchange market. *Journal of international Money and Finance*, *11*(3), 304–314.

Teräsvirta, T. (2009). An introduction to univariate garch models. In *Handbook of financial time series* (pp. 17–42). Springer.

ter Ellen, S., & Verschoor, W. (2017). Heterogeneous beliefs and asset price dynamics: a survey of recent evidence.

ter Ellen, S., Verschoor, W. F., & Zwinkels, R. C. (2013). Dynamic expectation formation in the foreign exchange market. *Journal of International Money and Finance*, *37*, 75–97.

Tierney, N., Cook, D., McBain, M., & Fay, C. (2018). naniar: Data structures, summaries, and visualisations for missing data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=naniar` (R package version 0.3.1)

Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, *66*(1-2), 225–250.

Trapletti, A., & Hornik, K. (2018). tseries: Time series analysis and computational finance [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=tseries` (R package version 0.10-44.)

Van Kervel, V. (2015). Competition for order flow with fast and slow traders. *The Review of Financial Studies*, *28*(7), 2094–2127.

Wang, Y.-C., Wu, J.-L., & Lai, Y.-H. (2018). New evidence on asymmetric return–volume dependence and extreme movements. *Journal of Empirical Finance*, *45*, 212–227.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from `http://www.jstatsoft.org/v40/i01/`

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23.

Wickham, H. (2017). stringr: Simple, consistent wrappers for common string operations [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=stringr` (R package version 1.2.0)

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=dplyr` (R package version 0.7.4)

Wooldridge, J. (2016). *Introductory econometrics: A modern approach*. Cengage Learning.

Ying, C. C. (1966). Stock market prices and volumes of sales. *Econometrica: Journal of the Econometric Society*, 676–685.

Zeeman, E. (1974). On the unstable behaviour of stock exchanges. *Journal of Mathematical Economics*, *1*(1), 39 - 49. Retrieved from `http://www.sciencedirect.com/science/article/pii/0304406874900342` doi: https://doi.org/10.1016/0304-4068(74)90034-2

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, *14*(6), 1–27. doi: 10.18637/jss.v014.i06

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10. Retrieved from `https://CRAN.R-project.org/doc/Rnews/`

Appendix A Data Preparation

A.1 Data structure

There are four main data-files from *Oslo Børs Informasjon AS / BI's Database* we will rely on: a daily returns dataset, a daily volume dataset, a dataset for identifying securities and companies based on a set of names and ID-numbers, and a dataset with monthly observations of stock prices and number of outstanding shares – used for filtering our data later.

The daily return dataset was available as a white space-delimitered txt-file structured similar to Figure 9, with the return from k securities stacked on top of each other.

$\#$	$ISIN_{Stock_1}$ $Company\ name_{Stock_1}$
$Date_1$	$return$
$Date_2$	$\vdots$
$Date_n$	$\vdots$
$\#$	$ISIN_{Stock_k}$ $Company\ name_{Stock_k}$
$Date_1$	$return$
$Date_2$	$\vdots$
$Date_n$	$\vdots$

Figure 9: Original data structure: Return

The volume data was also given as white space-delimitered txt-files, but was separated into 38 different files – one for each year of our sample period. They were all structured in the way seen in Figure 10, with the volume from k securities stacked on top of each other.

None of these formats are optimal for data analysis or for matching the correct volume and return observations when merging the datasets. Therefore we have to prepare our data.

$\#$	$OBI\ ID_{Stock_1}$
$Date_1$	$Volume$
$Date_2$	$\vdots$
$Date_n$	$\vdots$

$\#$	$OBI\ ID_{Stock_k}$
$Date_1$	$Volume$
$Date_2$	$\vdots$
$Date_n$	$\vdots$

Figure 10: Original data structure: Volume

A.2 Data preparation

Give me six hours to chop down a tree and I will spend the first four sharpening the axe.

– (Attributed to) Abraham Lincoln

Most statistical theory focus on data modeling, prediction, and statistical inference – while it is usually assumed that the data are in the correct state for data analysis already (de Jonge & van der Loo, 2013, p. 7). However, this is not always the case. According to Wickham (2014, p. 5), real datasets are often messy in almost every way imaginable[14]. Our dataset is no exception, and thus it needs to be cleaned[15]. Data cleaning is the process of transforming raw data into consistent data, which can be analyzed (de Jonge & van der Loo, 2013, p. 7). In practice, data preparation is often more time-consuming than the statistical analysis itself (de Jonge & van der Loo, 2013, pp. 3, 7), and according to Dasu and Johnson (2003) it is common to use upwards of 80% of the data analysis on cleaning and preparing data. Data cleaning is an important problem, but it is an uncommon subject of study in statistics (Wickham, 2014, p. 20). Done efficiently at the start of the project – using appropriate tools – the data processing stage can be highly rewarding; working with

[14]Or, as Jenny Bryan stated, "*classroom data are like teddy bears and real data are like a grizzley bear with salmon blood dripping out its mouth*."

[15]There are many words for data processing: cleaning, hacking, manipulating, munging, refining, tidying (Gillespie & Lovelace, 2016, p. 87).

clean data will be beneficial for every subsequent stage of the project (Gillespie & Lovelace, 2016, p. 87). Further, the data preparation process may profoundly influence the statistical statements based on the data, and should therefore be performed in a reproducible manner. Data cleaning methods such as imputation of missing values will influence statistical results and so must be accounted for in the analysis or interpretation thereof (de Jonge & van der Loo, 2013, pp. 7–8). In this subsection, we aim to give the reader a thorough understanding of our data preparation process. For reproducibility, we detail step by step how we combine our datasets, structure and clean them. Where it is appropriate, we will discuss how our choices might affect our statistical analysis.

A.2.1 Tidy data

Smart data structures and dumb code works a lot better than the other way around.

– Eric Raymond

Often, when working with statistics, we like to denote our models and formulas using linear algebra. The daily returns from k stocks and n days could for example be denoted as such

$$R_{n,k} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,k} \end{pmatrix}$$

Thus, it would make sense to structure our data in a spreadsheet like manner, similar to Figure 11.

	$Stock\ 1$	$Stock\ 2$		$Stock\ k$
$Date_1$	$return$	$return$		$return$
$Date_2$	$\vdots$	$\vdots$		$\vdots$
$Date_n$	$\vdots$	$\vdots$		$\vdots$

Figure 11: Spreadsheet structure

However, while data stored in arrays like this can lead to extremely efficient computation when the desired operations can be expressed as matrix operations, combining datasets stored in this way typically requires painstaking alignment before matrix operations can be used, which can make errors very hard to detect (Wickham, 2014, pp. 6, 14).

Therefore, we will structure our data in a way Wickham (2014) calls "tidy". In tidy datasets, (1) each variable forms a column, (2) each observation forms a row, and (3) each type of observational unit forms a table (Wickham, 2014, p. 4). A variable contains all values that measure the same underlying attribute – like date, return, or volume – across units. An observation contains all values measured on the same unit – like a certain stock at a given day – across attributes (Wickham, 2014, p. 3). Tidy datasets provide a standardized way to link the structure of a dataset with its semantics; its physical layout with its meaning (Wickham, 2014, p. 2). Fixed variables – variables with values fixed before the data collection, and not measured or collected – should come first, followed by measured variables (Wickham, 2014, p. 5). For us, this means that date, company name and ID should come first, as these are fixed variables. After this comes daily return and volume. An example of a tidy dataset can be seen in Figure 12.

Date	*OBI ID*	*ISIN*	*Company Name*	*Return*	*Volume*
1^{st} *Feb.* 2018	1	#*NO*1234567890	*Company A*	0.02	2134
2^{nd} *Feb.* 2018	1	#*NO*1234567890	*Company A*	0.04	5732
1^{st} *Feb.* 2018	2	#*NO*0987654321	*Company B*	0.03	98543
2^{nd} *Feb.* 2018	2	#*NO*0987654321	*Company B*	0.07	5432

Figure 12: Tidy data example with two fictional stocks at two dates

However, tidy data is only worthwhile if it makes analysis easier (Wickham, 2014, p. 13). For us, an advantage of tidy data is the ease at which it can be combined with other tidy datasets. When merging – or joining – two datasets, all we need is a "join operator" that works by matching common variables and adding new columns (Wickham, 2014, p. 14). We can use date, OBI security ID and ISIN to do this. Further, tidy datasets are easy to manipulate, model, and visualize. They work with a wide range of tidy tools – tools that use tidy datasets as input and output a new tidy dataset. This is useful because the output of one tool can be used as the input to another (Wickham, 2014, p. 13). Additionally, most modelling tools – such as R's linear regression – work best with tidy datasets (Wickham, 2014, p. 14). Converting data into a tidy form is also advantageous from a computational efficiency perspective, as it is usually faster to run analysis and plotting commands on tidy data (Gillespie & Lovelace, 2016, p. 89). Tidy data is particularly well suited for vectorized programming languages like R, as the layout ensures that the values of different variables from the same observation are always paired (Wickham, 2014, p. 5).

A.2.2 Combining and structuring our data

All data handling done in this section was performed using the open source statistical software R (R Core Team, 2017). All R packages used – a collection of user-created functions downloaded from the Comprehensive R Archive Network (CRAN) – will be cited consecutively, as different packages may have different specifications. The complete R-code for importing, structuring, combining, cleaning, and filtering our data can be found in Appendix B.

First, we import the identification dataset. We tell R to read the txt-file line by line, before we split each entry separated by a tab into different columns. We set the first row as header names, and save the data to a `data.frame`, telling R to treat text entries as text, not as factors. This results in a `data.frame` of almost 7,000 rows – one for each equity instrument which has been on OSE since 1980 – and 4 columns: OBI Security ID, ticker, ISIN and the last registered security name. This dataset was already tidy, so we did not have to change the structure at all. However, following the analogy of de Jonge and van der Loo (2013), we had to make the dataset *technically correct*. Technically correct data is data which is read into an R `data.frame`, with correct names, classes and labels (de Jonge & van der Loo, 2013, p. 7). A dataset is technically correct when each value can be directly recognized as belonging to a certain variable and is stored in a data class that represents the value domain of the real-world variable (de Jonge & van der Loo, 2013, p. 12). That is, a text variable should be stored as text and a numeric variable as a number. The class of an R object is critical to performance, as if the class is incorrectly specified this might lead to incorrect results (Gillespie & Lovelace, 2016, p. 94). To make the identification data technically correct we had to coerce the OBI security ID to the class numeric. Further, we transformed the `data.frame` to a `tibble` (Müller & Wickham, 2017) – a more convenient data frame class for R – before we used the package `naniar` (Tierney, Cook, McBain, & Fay, 2018) to replace all empty values with explicit `NA`-values, which R understands as missing data. The structure of the dataset can be seen in Figure 13.

OBI security ID	*ticker*	*ISIN*	*Last security name*
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

Figure 13: Tidy identification data

Continuing, we look at the volume data. As mentioned in Subsection A.1, the volume data is divided into 38 different files. Thus, our first job is to combine them with each other into one large dataset, which we can structure in a tidy manner before combining it with the return data. We have stored all the volume-files in a folder called daily_volume, and to combine them we first generate a vector of all the file names in that folder that end with ".txt". Then, we use the package `plyr` (Wickham, 2011) to import all the 38 files and combine them in one long dataset. This results in a `data.frame` with almost 1.48 million rows, and 2 columns: one containing the date and the OBI ID stacked on top of each other, and the last containing daily volume. This is a similar structure to the raw return data, so next step is to write an algorithm to tidy both of them.

We start by tidying the volume data. First, we turn the `data.frame` into a `tibble` (Müller & Wickham, 2017) for easier handling, before we name the columns "Date" and "Volume". Next, we create an empty character-vector of the same length as the number of rows in our volume dataset, which we call "OBI.security.ID", and a vector called "condition" which is equal to `TRUE` if the Volume variable is empty, and `FALSE` otherwise. We use a for-loop to go through each row in our volume dataset were the Volume observation is empty, and copy the OBI security ID to the vector of the same name. This results in a vector of approximately 1.48 million entries – where most entries are `NA`s. Using the package `zoo` (Zeileis & Grothendieck, 2005), we replace all the `NA`s with the last non-`NA` entry. Thus, we end up with a vector with no missing values, which we then merge with the volume data. The volume dataset now contains the variables: Date, Volume, and OBI security ID. Then, we use the package `stringr` (Wickham, 2017) to remove the "#" that was at the beginning of all the OBI security IDs before we omit rows that used to hold the OBI security ID, but do not register anything anymore. To make the volume dataset technically correct, we coerce OBI security ID and Volume to the class numeric, Date to a date class of type POSIXct[16]using the package `lubridate` (Grolemund & Wickham, 2011). The structure of the result can be seen in Figure 14.

$Date$	$Volume$	$OBI\ security\ ID$
⋮	⋮	⋮
⋮	⋮	⋮

Figure 14: Tidy volume data

Next, we structure the return data. Similar to the identification dataset we tell R to read the txt-file line by line, before we split each entry separated by a white space consisting of two spaces into different columns. We set header names, and save the data to a `data.frame`, telling R to treat text entries as text, not as factors. This results in a `data.frame` with over 2.5 million rows, and 3 columns: one with #s and the dates stacked on top of each other, one with the ISIN and return stacked on top of each other, and one containing one entry with the last registered company name of each company, but mostly `NA`-values. We processed this dataset in

[16]When converting the date, we can choose between three object types: Date, POSIXlt, and POSIXct. According to de Jonge and van der Loo (2013, pp. 20–21), POSIXct is the most portable way to store such information.

a similar fashion to the volume data. First, we created a character vector called ISIN of the same length as the dataset and a condition vector which is `TRUE` when the Date column contained nothing but a "#", and `FALSE` else. Then we created a for-loop that went through each row of the dataset where the date-column contained only a "#", filling the ISIN-vector with the entry from the ISIN/return-column. The result is a vector of the same length as the number of rows in the dataset, which contains some ISIN numbers, but mostly `NA`s. We use `zoo` (Zeileis & Grothendieck, 2005) to replace all the `NA`s with the last non-`NA` entry. We add the ISIN-vector as a separate column in the dataset, before we use `zoo` (Zeileis & Grothendieck, 2005) again to replace all the `NA`s with the last non-`NA` entry in the column with the last registered company name. Next, we remove all the rows containing only a "#" in the date-column as they are no longer of use. To make the return data technically correct we coerce Return into class numeric, and the Date into a date class of type POSIXct. The final dataset has the structure seen in Figure 15.

Date	*Return*	*Last Company Name*	*ISIN*
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

Figure 15: Tidy return data

Last, we import the monthly data. We read the csv-file line by line, and split the entries separated by a semicolon into columns. This resulted in a dataset with about 85,500 rows and 9 columns: OBI security ID, ISIN, ticker, last security name, date, monthly return, monthly dividend, stock price at the end of the month, and number of shares outstanding at the end of the month. This dataset was already tidy, so in order to make it technically correct we only had to coerce the OBI security ID, monthly return, monthly dividend, last price, and number of shares to class numeric, and the date to class date of type POSIXct. A dummy table can be seen in Figure 16.

Date	*ticker*	*ISIN*	*Security Name*	*Return*	*Dividend*	*Price*	*Shares Outstanding*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 16: Tidy monthly data

The next step is to merge these four datasets. As some ISIN numbers in the identification dataset seemed to be outdated, we updated some of them manually. Using the package `dplyr` (Wickham, Francois, Henry, &

Müller, 2017) and the joining logic of SQL with inner join, semi join and left join[17], we combined the dataset as seen in Figure 17. First, (1) we merged the return and identification dataset using inner join by the ISIN. Then, (2) we merged the volume dataset with the ISIN numbers from the identification dataset using inner join by the OBI security ID. Next, (3) we used semi join by OBI security ID on both the merged datasets to remove entries which are not in both datasets. Then, (4) we merged the shortened dataset with the merged dataset between volume and identification using left join by OBI security ID. Using `lubridate` (Grolemund & Wickham, 2011), we extract the month and year from the date column and add them as two separate columns. We do this for both the merged dataset and the dataset containing monthly data. The last step (5) was to merge the monthly data and the merged dataset containing all the other datasets using inner join by OBI security ID, year, and month.

The final result is a dataset of almost 1.7 million rows and 12 columns: date, year, month, ticker, last company name, last security name, ISIN, OBI security ID, return, volume, last price of the month, and number of shares outstanding at the end of the month.

A.2.3 Data cleaning

Consistent data is the stage where technically correct data is ready for statistical inference; missing values, special values, obvious errors and outliers are either removed, corrected or imputed (de Jonge & van der Loo, 2013, pp. 8, 31). The process towards consistent data involves the following three steps: (1) detection of inconsistencies, (2) selection of the field of fields causing the inconsistency, and (3) correction of the fields that are deemed erroneous[18].

It is imposible to perform statistical analysis on data where one or more values are missing. Thus, one can either omit elements from the dataset or try to impute missing values. Dealing with missing data is something to be dealt with prior to any analysis (de Jonge & van der Loo, 2013, pp. 31–31). Since default imputation may yield unexpected or erroneous results for reasons that are hard to trace, the analyst should decide how empty values are handled (de Jonge & van der Loo, 2013, p. 32). In many datasets, missing values means 0 – such as missing volume in our dataset. If that is the case, it should be explicitly imputed with that value, because it is not unknown, but was coded as empty (de Jonge & van der Loo, 2013, p. 33). Calculations involving special values – such as *Inf* (infinity), *NA* (missing value), *NaN* (Not a Number) or *NULL* (no

[17]Simply put, when combining table A and table B by the join operator x:
1. "A inner join B by x" combine A and B but only for the x they have in common.
2. "A semi join B by x" does not combine the two tables, but remove all x in A which are not also in B.
3. "A left join B by x" keeps A as is, and join B where A and B have common x.

[18]The steps are not necessarily separated, but when (1) and (2) is performed separately, step (2) is usually referred to as error localization (de Jonge & van der Loo, 2013, p. 31).

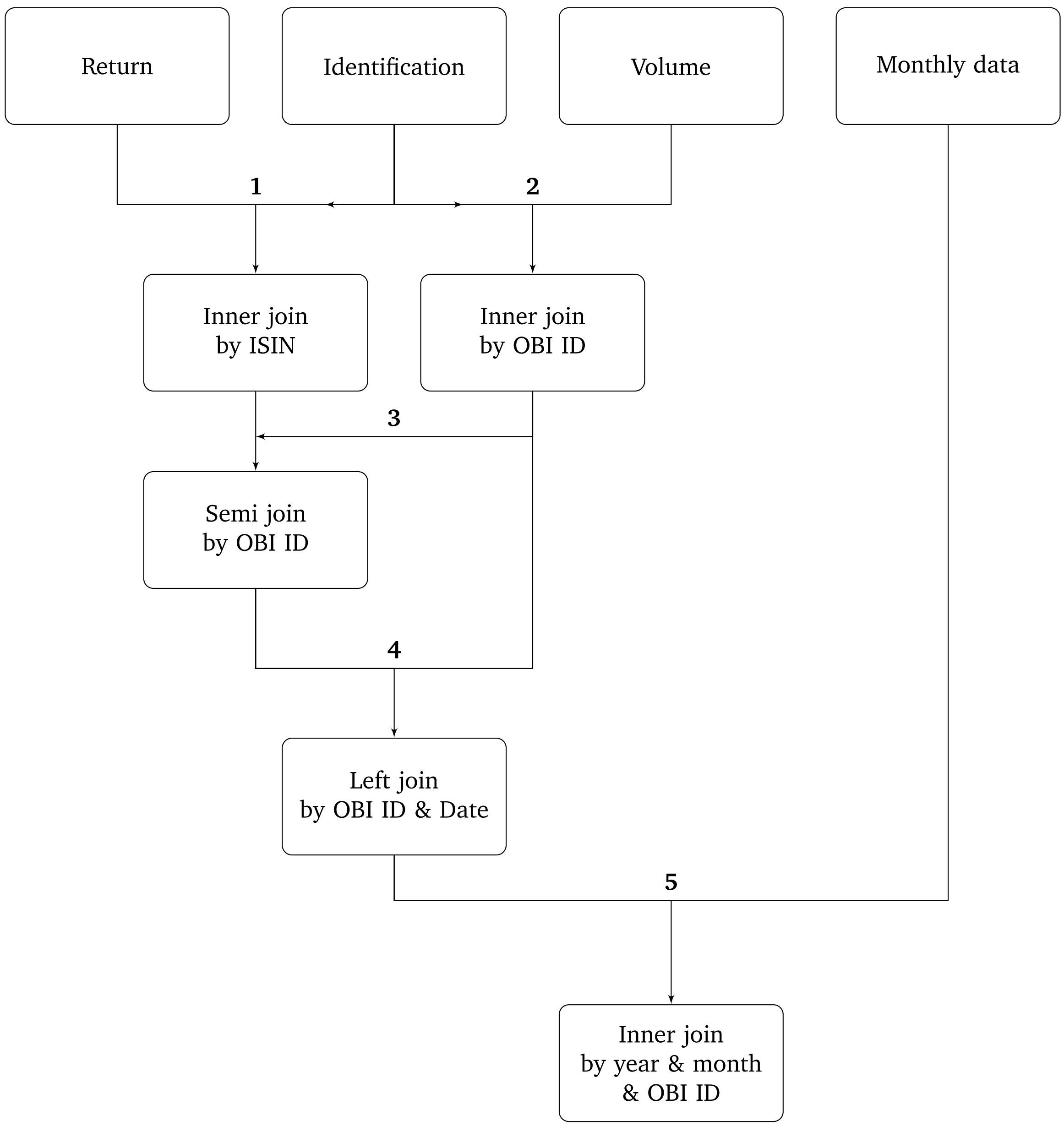

Figure 17: Joining of datasets

value) – often result in special values, and since statistical statements about real-world phenomenons should never include such values, it is desirable to handle them prior to analysis (de Jonge & van der Loo, 2013, p. 33). Obvious inconsistencies occurs when a record contains a value or combination of values that cannot correspond to a real-world situation (de Jonge & van der Loo, 2013, p. 35). For example, trading volume cannot be negative and return cannot be less than -1. As seen in Figure 1 in Section 2.1, trading was not fully automatic at Oslo Stock Exchange before 1999. When the data is registered by people rather than machines, certain typical human-generated errors are likely to occur – such as typing errors, rounding errors, sign errors or variable swaps (de Jonge & van der Loo, 2013, p. 42). We checked extensively for such inconsistencies, but found no such cases.

To clean the data, we start by using `zoo` (Zeileis & Grothendieck, 2005) to replace all missing values of volume with 0s. The values are only missing as there was no trading on that specific day, and as such 0 is the correct value. Further, we add some variables we need for the data filtering in the section 5.4. We add the variable market capitalization (MCAP) as the product of the stock price and number of outstanding shares at the end of each month. Then we add a dummy variable equal to 1 if the volume is a positive value – this dummy can later be used to count how many trading days a stock has each year.

Appendix B Script: Data preperation

Script for importing, structuring, combining, cleaning, and filtering the data.

```
################################################################################
#                Title: Master thesis - cleaning & filtering
#                Author: Jan Petter Iversen
#                Last update: May 22 - 2018
#
#                Requirements:
#                  * Datasets:
#                      - sec_list.txt
#                      - daily_returns.txt
#                      - daily_volume-folder with daily_volume_xxxx.txt-files
#                      - monthly_stock_returns_ose.csv
#                  * Packages
#                      - See: Setup
################################################################################


#################################### Setup #####################################

setwd("C:/Users/Lokal/Desktop/Data Master Thesis")
```

```
library(stringr)
library(lubridate)
library(tibble)
library(Amelia)
library(plyr)
library(dplyr)
library(zoo) # to use na.locf --> carries last non-NA value forward
library(naniar) # for replacing values (e.g. "") with NA
library(parallel) # for using several processors on big data


##################################### Identification data ##################################

# Importing indentification dataset, line by line
df_ident_raw <- readLines('sec_list.txt')

# split the lines by tab
df_ident_raw <- strsplit(df_ident_raw, "\t")

# create a function to assign the values to different fields
assignFields_ident <- function(x){

  out <- character(4)

  out[1] <- x[1]
  out[2] <- x[2]
  out[3] <- x[3]
  out[4] <- x[4]

  out
}

# apply the function
standardFields_ident <- lapply(df_ident_raw, assignFields_ident)

# unlist the list to a matrix
M_ident <- matrix(
  unlist(standardFields_ident)
  , nrow=length(standardFields_ident)
  , byrow=TRUE)

# set columnnames and remove first row (containing column names)
colnames(M_ident) <- (M_ident[1,] %>% str_trim())
M_ident <- M_ident[-1,]

# create data frame from matrix
df_ident_linebyline <- as.data.frame(M_ident, stringsAsFactors=FALSE)


## create technical correct data

# Coerce numeric
```

```
df_ident_linebyline$`OBI security ID` <- as.numeric(df_ident_linebyline$`OBI security ID`)

# create technical correct tibble
df_ident_tech <- as.tibble(df_ident_linebyline)

# remove all functions and variables except the technical correct ident df
rm(assignFields_ident, df_ident_linebyline, df_ident_raw, M_ident, standardFields_ident)

df_ident_tech <- replace_with_na_all(data = df_ident_tech, condition = ~.x == "")

df_ident_tech <-  rename(df_ident_tech, 'OBI.security.ID' = `OBI security ID`)


##################################### Volume data #####################################

# save the name of all .txt-files from the folder in a vector
paths <- dir("daily_volume" , pattern = "\\.txt$", full.names = TRUE)

# add the name of the file to the vector, this will be used for ID later
names(paths) <- basename(paths)

# loop through the files and read them all into a DF
df_volume_raw <-  ldply(paths,
                        read.table,
                        comment.char = "", # ignore comments
                        quote = "",
                        colClasses = c('character'),
                        fill = T, # empty cells gets "NA"
                        header = F, # no headers
                        skip = 5, # skip general informaiton at the top of each file
                        stringsAsFactors = FALSE) # don't read strings as factors


# as tibble, since it's easier
df_volume_raw <-  as.tibble(df_volume_raw)

# remove file ID, for similarity to return data
df_volume_nofilename <- df_volume_raw[,2:3]

# name columns
colnames(df_volume_nofilename) <- c("Date", "Volume")

# Create empty vector for ID
OBI.security.ID <- character(length = nrow(df_volume_nofilename))

# create the condition outside the loop
condition <- df_volume_nofilename$Volume == ""

# loop through the rows and fill with OBI ID
for (i in (1:nrow(df_volume_nofilename))[condition] ) {

  if (condition[i]) {

    OBI.security.ID[i] <- df_volume_nofilename$Date[i]
```

```
  }

}

# replace "" with NAs so na.locf will work
OBI.security.ID[OBI.security.ID == ""] <- NA

# locate forward last non-NA value
OBI.security.ID <- na.locf.default(OBI.security.ID)

# add the vector to the DF as column (same name as in ident-DF)
df_volume_nofilename$OBI.security.ID <- OBI.security.ID

# new data frame to continue work
df_volume_w.ID <-  df_volume_nofilename

# remove '#' from OBI-ID
df_volume_w.ID$OBI.security.ID <- str_sub(df_volume_w.ID$OBI.security.ID, start = 2)

# omit rows with volume == "" --> this will only remove rows where the date is an OBI ID
df_volume_w.ID <- df_volume_w.ID[df_volume_w.ID$Volume != "", ]

# make df technically correct by coercing class
# coerce numeric
df_volume_w.ID$OBI.security.ID <- as.numeric(df_volume_w.ID$OBI.security.ID)
df_volume_w.ID$Volume <- as.numeric(df_volume_w.ID$Volume) # coerce numeric
df_volume_w.ID$Date  <- ymd(df_volume_w.ID$Date, tz = Sys.timezone()) # coerce date

# save technically correct data
df_volume_tech <- df_volume_w.ID

# remove all variables except the technical correct volume df (makes code go faster)
rm(condition, df_volume_nofilename, df_volume_raw,
     df_volume_w.ID, OBI.security.ID, paths, i)



##################################### Return data #####################################

# read daily return line by line
df_return_raw <- readLines('daily_returns.txt')

# remove general information at the top of the file
df_return_raw <- df_return_raw[-c(1:6)]

# split the strings by a double space
df_return_raw <- strsplit(df_return_raw, "  ")

# create function to assign values to different columns
assignFields_ret <- function(x){

  out <- character(3)
```

```
  out[1] <- x[1]
  out[2] <- x[2]
  out[3] <- x[3]

  out
}

# use four processors and apply the newly created function
cluster <- makeCluster(4)
standardFields_ret <- parLapply(cl=cluster, df_return_raw, assignFields_ret)
stopCluster(cluster) # stop clustering processors

# create matrix of values
M_ret <- matrix(
  unlist(standardFields_ret)
  , nrow=length(standardFields_ret)
  , byrow=TRUE)

# name the coulmns of the matrix
colnames(M_ret) <- c("Date","Return","Last.Company.Name")

# save to data frame
df_return_linebyline <- as.data.frame(M_ret, stringsAsFactors=FALSE)

# create vector for ISIN (will become column)
ISIN <- character(length = nrow(df_return_linebyline))

# create condition
condition <- df_return_linebyline$Date == "#"

# loop through the rows and fill with ISIN (less than 6 seconds)
for (i in (1:nrow(df_return_linebyline))[condition] ) {

  if (condition[i]) {

    ISIN[i] <- df_return_linebyline$Return[i]

  }

}

# replace "" with NAs so na.locf will work
ISIN[ISIN == ""] <- NA

# locate forward last non-NA value
ISIN <- na.locf.default(ISIN)


# add the vector to the DF as column (same name as in ident-DF)
df_return_linebyline$ISIN <- ISIN

# locate forward the last non-NA company name in "Last.Company.Name"
df_return_linebyline$Last.Company.Name <-
```

```
        na.locf.default(df_return_linebyline$Last.Company.Name)

# replace "#" with NAs
df_return_linebyline$Date[df_return_linebyline$Date == "#"] <- NA

# remove rows with NAs (will only remove rows which used to be "#" and company names)
df_return_linebyline <- (df_return_linebyline %>% na.omit())

# coerce Return to be numeric
df_return_linebyline$Return <- as.numeric(df_return_linebyline$Return)

## fix Date
# trim for whitespace
df_return_linebyline$Date <- str_trim(df_return_linebyline$Date)

# coerce to date format
df_return_linebyline$Date <- ymd(df_return_linebyline$Date, tz = Sys.timezone())

# new data frame - as tibble for easy reading
df_return_tech <- as.tibble(df_return_linebyline)

# remove the variables and functions we no longer need
rm(assignFields_ret, cluster, condition, df_return_linebyline,
     M_ret, standardFields_ret, df_return_raw, i, ISIN)


##################################### Monthly price and share data #######################

# Importing indentification dataset, line by line
df_monthly_data_raw <- readLines('monthly_stock_returns_ose.csv')

# split the lines by tab
df_monthly_data_raw <- strsplit(df_monthly_data_raw, ";")

# create a function to assign the values to different fields
assignFields_df_monthly_data_raw <- function(x){

  out <- character(9)

  out[1] <- x[1]
  out[2] <- x[2]
  out[3] <- x[3]
  out[4] <- x[4]
  out[5] <- x[5]
  out[6] <- x[6]
  out[7] <- x[7]
  out[8] <- x[8]
  out[9] <- x[9]

  out
}

# apply the function
standardFields_df_monthly_data_raw <-
```

```
        lapply(df_monthly_data_raw, assignFields_df_monthly_data_raw)

# unlist the list to a matrix
M_monthly_data <- matrix(
  unlist(standardFields_df_monthly_data_raw)
  , nrow=length(standardFields_df_monthly_data_raw)
  , byrow=TRUE)

# set columnnames and remove first row (containing column names)
colnames(M_monthly_data) <- (M_monthly_data[1,] %>% str_trim())
M_monthly_data <- M_monthly_data[-1,]

# create data frame from matrix
df_monthly_data_linebyline <- as.tibble(M_monthly_data, stringsAsFactors=FALSE)

## coerce classes

df_monthly_data_linebyline$OBI_SEC_ID <-
                as.numeric(df_monthly_data_linebyline$OBI_SEC_ID)

df_monthly_data_linebyline$MonthlyReturn <-
                as.numeric(df_monthly_data_linebyline$MonthlyReturn)

df_monthly_data_linebyline$MonhlyDividend <-
                as.numeric(df_monthly_data_linebyline$MonhlyDividend)

df_monthly_data_linebyline$LastPrice <-
                as.numeric(df_monthly_data_linebyline$LastPrice)
df_monthly_data_linebyline$NoShares <-
                as.numeric(df_monthly_data_linebyline$NoShares)

df_monthly_data_linebyline$Date <-
                ymd(df_monthly_data_linebyline$Date, tz = Sys.timezone())

# save to new df we will keep
df_monthly_data_tech <- df_monthly_data_linebyline

# remove the variables and functions we no longer need
rm(df_monthly_data_linebyline,
   df_monthly_data_raw,
   standardFields_df_monthly_data_raw,
   assignFields_df_monthly_data_raw,
   M_monthly_data)


####################################Manually fix some entries, before merging ###########

# Some ISIN-numbers are outdated, and we have chosen to fix those
# which we easily could identify the correct ISIN numer for

# DNB
df_ident_tech$ISIN[df_ident_tech$ISIN == "NO0003002008"] <- "NO0010031479"

# Petroleum Geo-Services
```

```
df_ident_tech$ISIN[df_ident_tech$ISIN == "NO0004225004"] <- "NO0010199151"

# Wilh. Wilhelmsen Holding ser. A
df_ident_tech$ISIN[df_ident_tech$ISIN == "NO0003471401"] <- "NO0010571698"

# Wilh. Wilhelmsen Holding ser. B
df_ident_tech$ISIN[df_ident_tech$ISIN == "NO0003471419"] <- "NO0010576010"

# Stolt-Nielsen
df_ident_tech$ISIN[df_ident_tech$ISIN == "LU0081746793"] <- "BMG850801025"

# SAS AB
df_ident_tech$ISIN[df_ident_tech$ISIN == "SE0000805574"] <- "SE0003366871"

# Wentworth Resources
df_ident_tech$ISIN[df_ident_tech$ISIN == "CA04317T1066"] <- "CA9506771042"

# BW Offshore Limited
df_ident_tech$ISIN[df_ident_tech$ISIN == "BMG1190N1002"] <- "BMG1738J1247"

# FLEX LNG
df_ident_tech$ISIN[df_ident_tech$ISIN == "VGG359451074"] <- "BMG359471031"

# Avocet Mining
df_ident_tech$ISIN[df_ident_tech$ISIN == "GB0000663038"] <- "GB00BZBVR613"

 # Archer
df_ident_tech$ISIN[df_ident_tech$ISIN == "BMG0451H1097"] <- "BMG0451H1170"

# Hugo Games
df_ident_tech$ISIN[df_ident_tech$ISIN == "DK0060637999"] <- "DK0060945467"


#################################### Merge datasets ####################################

## merging data

# 1! Merging return and ident, inner by ISIN
merged_return_ident <- inner_join(df_return_tech, df_ident_tech, by = "ISIN")

# 2! Merging volume and ident, inner by OBI ID
merged_volume_ident <-
                inner_join(df_volume_tech, df_ident_tech[,c(1,3)],
                                by = "OBI.security.ID")

# 3! Shortening merged_return_ident (1) with semi_join by ISIN
merged_short_return_ident <-
                semi_join(merged_return_ident, merged_volume_ident,
                                by = "OBI.security.ID")

# 4! Merging merged_short_return_ident (3) with merged_volume_ident (2)
merged_data <-
                left_join(merged_short_return_ident,
                             merged_volume_ident[,c(1,2,3)],
```

```
                              by = c("OBI.security.ID", "Date"))

# Now we want to add the montly price and number of shares outstanding
# Then we need to add the month and year as seperate columns,
# so we can use these to merge the datasets

merged_data$year <- year(merged_data$Date)
merged_data$month <- month(merged_data$Date)

df_monthly_data_tech$year <- year(df_monthly_data_tech$Date)
df_monthly_data_tech$month <- month(df_monthly_data_tech$Date)

# rename so OBI ID has same name in all DFs
df_monthly_data_tech <- rename(df_monthly_data_tech, OBI.security.ID = OBI_SEC_ID)



# merge merged_data and the columns we need from df_monthly_date_tech
merged_data_full <- inner_join(merged_data, df_monthly_data_tech[,c(1,8:11)],
                               by = c("OBI.security.ID", "year", "month"))


rm(merged_return_ident, merged_short_return_ident, merged_volume_ident, merged_data)

#################################### Clean data ####################################

# fill NAs in volume with 0s
merged_data_full$Volume <-  na.fill(merged_data_full$Volume, fill = 0)

## add variables we need

# add MCAP as the product of price and outstanding shares
merged_data_full$MCAP <- (merged_data_full$LastPrice * merged_data_full$NoShares)

# create dummy: 1 if the stock was traded that day, 0 if not.
merged_data_full$trade <- as.integer(ifelse(merged_data_full$Volume > 0, 1, 0))

# Create a combination of year and month to use as an ID.
merged_data_full <-  merged_data_full %>%
                               mutate(ym_id = paste(as.character(year),
                                              as.character(month),
                                              sep = "-"))

# Create dummy for stocks that are going of the exchange
# (those with last month December 2017 are skipped)
merged_data_full$LastMonth <-
        ifelse(((merged_data_full$ym_id != "2017-12")&(is.na(merged_data_full$LastPrice))),
                1, 0)


#################################### Filter Data ####################################

# reshape dataset for intuitive working
merged_data_full <- merged_data_full %>%
```

```
  select(Date,
         year,
         month,
         ym_id,
         OBI.security.ID,
         ISIN,
         ticker,
         `Last Security Name`,
         Last.Company.Name,
         Return,
         Volume,
         trade,
         LastPrice,
         NoShares,
         MCAP,
         LastMonth)


# Remove all december 2017 observations
merged_data_full <- merged_data_full %>% filter(!ym_id == "2017-12")


# remove companies where we cannot calculate MCAP
# (this is, among others, removing the last month of each security)
merged_data_full <- merged_data_full %>% filter(!is.na(MCAP))


# define smallcap companies as those which has at
# least one observation of MCAP below 1M NOK
smallcap <- merged_data_full %>%
              filter(MCAP < 1000000) %>%
              select(Last.Company.Name) %>%
              unique() %>%
              pull()


# filter the data for smallcap companies, find average yearly MCAP,
# and filter those with Mean_MCAP below 1M NOK
smallcap_company_year_pairs <- merged_data_full %>%
  filter(Last.Company.Name %in% smallcap) %>%
  group_by(Last.Company.Name, year) %>%
    summarise(Mean_MCAP = mean(MCAP)) %>%
  filter(Mean_MCAP < 1000000)

# loop through the smallcap_company_year_pairs-dataset
# remove all observations for that stock for that year if it appears in the dataset
# (That is, if the yearly average MCAP was below 1 MNOK)
for (i in 1:nrow(smallcap_company_year_pairs)) {

  merged_data_full <- merged_data_full %>%
    filter(
      !((Last.Company.Name == smallcap_company_year_pairs$Last.Company.Name[i])
      &
      (year == smallcap_company_year_pairs$year[i]))
```

```
    )

}

# remove variables and dfs we do not need anymore
rm(smallcap, i, smallcap_company_year_pairs)

## high and low prices stocks

# define high and low priced stocks as those which have an average price
# during one year of less than 10 NOK or above 8000 NOK
high_low_price <- merged_data_full %>%
  group_by(Last.Company.Name, year) %>%
  summarise(Avg_Price = mean(LastPrice)) %>%
  filter( (Avg_Price < 10)|(Avg_Price >= 8000) )

# Loop through alle rows in high_low_price, remove rows from merged_data_full
# with matching company name and year.
for (i in 1:nrow(high_low_price)) {

  merged_data_full <- merged_data_full %>%
    filter(
      !((Last.Company.Name == high_low_price$Last.Company.Name[i])
      &
      (year == high_low_price$year[i]))
    )

  print(paste(i, " of ", nrow(high_low_price), sep = "")) # just to see some progress

}

# remove variables and dfs we no longer need
rm(i, high_low_price)

# define companies where there are less than 20 days of trading in a year
few_trades <- merged_data_full %>%
  group_by(Last.Company.Name, year) %>%
  summarise(yearly_trading_days = sum(trade)) %>%
  filter(yearly_trading_days < 20)

# loop through and remove the full year of trades if there is less than 20 days of trading
for (i in 1:nrow(few_trades)) {

  merged_data_full <- merged_data_full %>% filter(
    !((Last.Company.Name == few_trades$Last.Company.Name[i])&(year == few_trades$year[i]))
    )

  print(i) # counter to see progress

}

# remove variables and df we no longer need
rm(i, few_trades)
```

```
## filter out Savings banks and non-stock equties

# no observations left with security names including the sub-strings:
# warrant,
# bull, bear,
# DNM, Nordnet

# save a df of all companies with "Spare" in their name
savings_banks <- merged_data_full %>%
  select(`Last Security Name`, Last.Company.Name) %>%
  unique() %>%
  filter(
    str_detect(
      Last.Company.Name, paste(
        c(
          "Spare",
          "spare"
          ),
        collapse = '|'))
    )

# remove "Sparebank 1 SR-Bank" as it is not a savings bank
savings_banks <- savings_banks %>% filter(`Last Security Name` != "SpareBank 1 SR-Bank")


# save a df of other savings banks to add
temp_add_to_savings_banks <- merged_data_full %>%
  select(`Last Security Name`, Last.Company.Name) %>%
  unique() %>%
  filter(`Last Security Name` %in% c("Sandsvaerbanken",
                                     "Sparabanken Rogaland"))

# add the two misisng banks to the savings_banks df
savings_banks <- bind_rows(savings_banks, temp_add_to_savings_banks)

# remove observations with these security names
merged_data_full <- merged_data_full %>%
                        filter(!`Last Security Name` %in%
                               savings_banks$`Last Security Name`)

# remove variables and dfs we no longer need
rm(savings_banks, temp_add_to_savings_banks)

# define small sample as less than 500 observations
small_samples_OBI <-  merged_data_full %>%
  group_by(OBI.security.ID) %>%
  summarise(count = n()) %>%
  filter(count < 500) %>%
  pull(OBI.security.ID)

# remove securities with too small samples
merged_data_full <- merged_data_full %>%
  filter(!OBI.security.ID %in% small_samples_OBI)
```

```
# remove variables we no longer need
rm(small_samples_OBI)


####################################  Save and export data ##############################

# save companies missing and companies included to two variables
companies_missing <- df_return_tech %>%
                                    filter(!ISIN %in% merged_data_full$ISIN) %>%
                                    select(Last.Company.Name) %>%
                                    unique()

companies <- merged_data_full %>% select(Last.Company.Name) %>% unique()

# create txt-files with the companies missing and included
write.table(companies_missing, "missing_companies.txt", sep = "\n",row.names = F)
write.table(companies, "companies_included.txt", sep = "\n",row.names = F)

save(merged_data_full, file = "fully_filtrated_data.RData")
unlink("fully_filtrated_data.RData.RData")

write.csv(merged_data_full, "fully_filtraded_data.csv")
```

Appendix C Companies included

The following 511 companies are included in our sample.

No.	Company name	Security Name
1	Nettbuss Sŕr	Aust–Agder Trafikkselskap
2	Adresseavisen	Adresseavisen
3	Actinor Shipping	Actinor Shipping
4	Actinor	Actinor A/S
5	Adelsten Holding A	Adelsten A
6	Adelsten Holding B	Adelsten B
7	Arendals Fossekompani	Arendals Fossekompani
8	Aker RGI A	Aker RGI
9	Aker RGI B	Aker B–aksjer
10	Aker F	Aker Frie aksjer
11	Ambra	Ambra
12	Arcen	Arcen
13	Atlantica	Atlantica
14	Autronica	Autronica
15	Avantor	Avantor AS
16	Awilco ser. A	Awilco
17	Awilco ser. B	Awilco B
18	Bŕndernes Bank	Bondernes Bank
19	Bergesen d.y ser. A	Bergesen d.y. A–aksjer
20	Bergesen d.y ser. B	Bergesen d.y. B–aksjer
21	Belships	Belships Co.
22	Benor Tankers	Benor Tankers
23	Bik Bok A	Bik Bok Gruppen
24	Bik Bok B	Bik Bok Gruppen B–aksjer
25	Bjŕlvefossen	Bjolvefossen
26	Bjŕlsen Valsemŕlle	Bjolsen Valsemolle
27	NRC Group	Blom A/S
28	Bolig– og NÊringsbanken	Bolig– og Naeringsbanken
29	Bergen Nordhordland Rutelag	Bergen Nordhordaland Rutelag
30	Bonheur	Bonheur
31	Borgestad	Borgestad A
32	Borgestad ser. B	Borgestad B
33	BorgÂ	Borgaa
34	Braathens	Braathens SAFE
35	Bona Shipholding	Bona Shipholding
36	Bergensbanken	Bergensbanken
37	Buskerudbanken	Buskerudbanken
38	Chr. Bank og Kreditkasse	Christiania Bank og Kreditkasse
39	Chr. Bank og Kreditkasse	Christiania Bank og Kreditkasse
40	Winder	Sagatex
41	Color Group	Color Line A.S.
42	Andvord Tybring–Gjedde	C.Tybring–Gjedde A/S
43	E.C.Dahls Bryggeri	E.C.Dahls Bryggeri
44	David Livsforsikringsselskap	David Livsforsikring
45	DNB	Den norske Bank
46	Den Norske Creditbank	Den norske Creditbank (DnC)
47	SAS Norge B	SAS Norge

48	DNO	Det Norske Oljeselskap (DNO)
49	DNO B	Det Norske Oljeselskap (DNO) B–aksjer
50	Det Stavangerske Dampskibss.	Det Stavangerske D/S.
51	Dyno	Dyno Industrier
52	Eiend. Aker Brygge I	Eiendomsselskapet Aker Brygge I
53	Norwegian Car Carriers	Eidsiva
54	Eiendomsutvikling	Eiendomsutvikling
55	Elkjŕp	Elkjop Norge
56	Elektrisk Bureau	Elektrisk Bureau
57	Elkem	Elkem
58	Elkem F	Elkem Frie aksjer
59	Farstad Shipping	Farstad Shipping A/S
60	Forretningsbanken	Forretningsbanken
61	Forenede–Gruppen	Forenede–Gruppen
62	ABG Sundal Collier Holding	Askia Invest
63	Fokus Bank	Fokus Bank
64	Fosen	Fosen Trafikklag
65	First Olsen Tankers	First Olsen Tankers
66	Freia Marabou A	Freia Marabou A–aksjer
67	Freia Marabou B	Freia Marabou B–aksjer
68	Telecast	Industriinvestor
69	Frysja Elektro	Frysja Elektro
70	Gambit	Gambit A/S
71	G. Block Watne	G. Block–Watne
72	Geophysical Comp. of Norway	Geophysical Comp. of Norway A.S (GECO)
73	Grand Hotel	Grand Hotel
74	Grand Hotel F	Grand Hotel Frie aksjer
75	Gimsŕy Kloster	Gimsoy Kloster
76	Christiania Glasmagasin	Christiania Glasmagasin
77	Goodtech	Goodtech
78	GPI	GPI
79	Ganger Rolf	Ganger Rolf
80	Gyldendal	Gyldendal Norsk Forlag
81	H?G	Haag
82	Hansa Bryggeri	Hansa Bryggeri
83	Havtor	Havtor
84	Havtor B	Hav B–aksjer
85	Helly–Hansen	Helly–Hansen
86	Hennes & Mauritz	H&M Hennes & Mauritz
87	Helicopter Services Gr.	Helikopter Service A/S
88	Hafslund ser. A	Hafslund Nycomed A–aksjer
89	Hafslund Nycomed F	Hafslund Nycomed frie A–aksjer
90	Hafslund ser. B	Hafslund Nycomed B–aksjer
91	Tide	Hardanger Sunnhordalandske DS
92	Hunsfos	Hunsfos Fabrikker
93	Ican	Ican a.s.
94	Idun–GjÊrfabrikken	Idun–Gjaerfabrikken
95	International Farvefabrik	International Farvefabrik
96	I.M. Skaugen97	I.M. Skaugen
97	Investa	Investa
98	Finansbanken	Finansbanken
99	Ivarans Rederi	Ivarans Rederi
100	Jonas ßglÊnd	Jonas Oglaend

101	Kaldnes Mek. Verksted	Kaldnes
102	Kaldnes	Kaldnes
103	Kjŕbmandsbanken	Kjopmandsbanken
104	Kristiansand Dyrepark	Kristiansand Dyrepark
105	Kosmos Holding	Kosmos Holding
106	Kirkland	Kirkland (listed etter SUS)
107	Kosmos	Kosmos
108	Kverneland	Kverneland
109	KvÊrner	Kvaerner Industrier
110	KvÊrner B	Kvaerner Industrier B–aksjer
111	KvÊrner F	Kvaerner Industrier Frie aksjer
112	KvÊrner Shipping	Kvaerner Shipping A/S
113	Laboremus	Laboremus
114	Larvik–Fredrikshavnferjen	Larvik–Fredrikshavnferjen
115	Lehmkuhl Elektronikk	Lehmkuhl Elektronikk A/S
116	Leif H?egh & Co	Leif Hoegh & Co A/S
117	Loki	Loki
118	Maritime Group	Maritime Group AS
119	Atea	Merkantildata A/S
120	H.C.A. Melbye	H.C.A. Melbye A/S
121	Mercurius	Mercurius
122	Moss GlasvÊrk A	Moss Glasvaerk A
123	Moelven Industrier	Moelven
124	Mycron	Mycron
125	Den Norske Amerikalinje	Den norske Amerikalinje
126	NTS	Namsos Trafikkselskap
127	Ugland Nordic Shipping	Ugland Nordic Shipping
128	Nordlandsbanken	Nordlandsbanken
129	Norsk Data A	Norsk Data
130	Norsk Data B	Norsk Data B–aksjer
131	Norsk El. & Brown Boveri	NEBB
132	Kongsberg Gruppen	Kongsberg Gruppen
133	Forsikringsselskapet Norge	Norge, Forsikringsselskapet
134	Norges Hypotekinstitutt	Norges Hypotekinstitutt
135	Norsk Hydro	Norsk Hydro
136	Nidar	Nidar
137	Norema A	Norema A–aksjer
138	Nobŕ Fabrikker	Noboe Fabrikker
139	Nora Eiendom	Nora Eiendom a.s
140	Nora Industrier	Nora Industrier
141	Nora Industrier F	Nora Industrier Frie aksjer
142	Norgeskreditt P	Norgeskreditt
143	Norex Offshore	Norex Offshore
144	Norcem	Norcem
145	Reach Subsea	Nomadic Shipping
146	Notodden Elektronikk	Notodden Elektronikk A.S
147	Norse Petroleum	Norse Petroleum
148	Norwegian Rig Consultants	Norwegian Rig Consultants A/S
149	Norske Skog	Norske Skogindustrier
150	Norske Skog B	Norske Skogindustrier B
151	Norske Skogindustrier	Norske Skog A
152	Norske Skog	Norske Skogindustrier
153	Norving	Norving

154	Linstow	Nydalens Compagnie
155	Oslobanken	Oslobanken A/S
156	Oslo Handelsbank	Oslo Handelsbank
157	Oslo Havnelager	Oslo Havnelager
158	Olav Thon Eiendomsselskap	Olav Thon Eiendomsselskap
159	Simrad Optronics	Simrad Optronics
160	Orkla	Orkla
161	Orkla B	Orkla B
162	Orkla F	Orkla Frie aksjer
163	Orkla Industrier	Orkla Industrier
164	Oslo Shipholding	Laly
165	Petroleum Geo–Services	Petroleum Geo–Services
166	Porsgrunds PorselÊn	Porsgrunds Porselaensfabrik
167	Protector Forsikring	Protector Forsikring
168	Pronova	Pronova
169	Raufoss	Raufoss A/S
170	Rogalandsbanken	Rogalandsbanken
171	Realia	Realia
172	Rena Karton	Rena Karton
173	Rieber & S??n	Rieber & Son
174	Rieber & Sŕn B	Rieber & Son B–aksjer
175	Ross Offshore	Ross Offshore
176	Rosshavet	Rosshavet
177	Saga Petroleum	Saga Petroleum A
178	Saga Petroleum B	Saga Petroleum B
179	Saga Petroleum F	Saga Petroleum Frie aksjer
180	Sunnmŕrsbanken	Sunnmorsbanken
181	Stord Bartz	Stord Bartz a.s
182	Schibsted ser. A	Schibsted
183	SDS Shipping og Offshore	SDS Shipping og Offshore A/S
184	Sea Farm	Sea Farm A/S
185	SensoNor	SensoNor
186	DSND Subsea	Det Sondenfjelds Norske D/S
187	Sigmalm	Sigmalm
188	Skiens Aktiem??lle	Skiens Aktiemolle
189	ARK	ARK
190	Simrad A	Simrad A
191	Simrad B	Simrad B
192	Smedvig ser. A	Smedvig a.s
193	Smedvig Tankships Ltd.	Smedvig Tankships Ltd.
194	Solvang	Solvang
195	Sŕrlandsbanken	Sorlandsbanken
196	Scanvest–Ring A	Scanvest–Ring A
197	Scanvest–Ring B	Scanvest–Ring B
198	Stavanger Aftenblad	Stavanger Aftenblad
199	Stentofon	Stentofon
200	Alcatel STK	Alkatel STK
201	Odfjell ser. A	Storli A
202	Odfjell ser. B	Storli B
203	Navia	Navia
204	Sydvaranger	Sydvaranger
205	SE Labels gammel	SE Labels
206	Avenir	Sysdeco Group
207	Tandberg	Tandberg A/S

208 Tandberg Data	Tandberg Data A/S
209 Tiki–Data	Tiki–Data A.S
210 Transocean	Transocean
211 Tofte Industrier	Tofte Industrier A/S
212 Tomra Systems	Tomra Systems
213 Tou	Tou
214 Storebrand P	UNI Storebrand Bundne Pref.
215 Storebrand	Storebrand
216 UNI Storebrand F	UNI Storebrand Frie
217 Unitor	Unitor
218 NCL Holding	NCL Holding
219 Vard B	Vard B–aksjer
220 Vestlandsbanken	Vestlandsbanken
221 Vestenfjelske Bykreditt	Vestenfjeldske Bykreditt
222 Vesteraalens Dampskibsselskab	Vesteraalens D/S
223 Veidekke	Veidekke
224 Vesta–Gruppen	Vesta–gruppen
225 Viking–Askim	Viking Askim, ord. B
226 Vital Forsikring	Vital Forsikring
227 Vital Forsikring F	Vital Forsikring Frie
228 Viking Supply Ships	Viking Supply Ships A.S
229 Voss Veksel– og Landmandsbank	Voss Veksel– og Landmandsbank
230 Western Bulk Shipping	Western Bulk Shipping
231 Wilrig	Wilrig AS
232 Wilh. Wilhelmsen Holding ser. A	Wilh. Wilhelmsen A
233 Wilh. Wilhelmsen Holding ser. B	Wilh. Wilhelmsen B
234 Gresvig	Gresvik
235 Axis Biochemicals	Axis Biochemicals
236 Steen & Strŕm	Steen & Strom
237 Hitec	Hitec
238 Larvik Scandi Line	Larvik Scandi Line
239 Klippen Invest	Jotul
240 Stento	Stento
241 Atlantic Container Line	Atlantic Container Line
242 Avantor	Avantor AS
243 Jinhui Shipping and Transportation	Jinhui Shipping
244 Viking Media	Viking Media
245 Fokus Bank	Fokus Bank
246 Statoil	Statoil
247 Norsk Vekst	Norsk Vekst
248 Nera	Nera
249 Kongsberg Automotive	Kongsberg Automotive
250 A–pressen	A–pressen
251 Ekornes	Ekornes
252 TTS Group	TTS Technology
253 Oslo Reinsurance Co	Oslo Reinsurance Comp.
254 CanArgo Energy Co.	Fountain Oil
255 Crystal Production	Brovig Offshore
256 Fesil	Fesil
257 Legra	Legra
258 Nordic Water Supply	Nordic Water Supply
259 Ivar Holding	Ivar Holding
260 Nordic American Tanker Shipping	Nordic Am. Tanker Shipping
261 Santech Micro Group	Santech Micro Group

262	Selmer	Selmer
263	Agresso Group	Agresso Group
264	Mercur Tankers	Mercur Tankers
265	Visma	Visma
266	Scana Industrier	Scana Industrier
267	Marine Harvest	Pan Fish
268	Stolt–Nielsen B	Stolt–Nielsen B
269	Computer Advances	Computer Advances Group
270	SuperOffice	SuperOffice
271	Norman	Norman Data Def. Sys.
272	Stolt–Nielsen	Stolt Nielsen Ordinaere
273	Opticom	Opticom
274	Altinex	Mercur Subsea Products
275	Nordic Semiconductor	Nordic VLSI
276	Provida	Provida
277	NetCom	NetCom
278	Reitan Narvesen	Narvesen
279	SPCS–Gruppen	PC–Systemer
280	Hydralift	Hydralift
281	ORIGIO	Medi–Cult
282	Wenaas	Wenaas–gruppen
283	Proxima	ASK
284	Smedvig	Smedvig B
285	Transocean Offshore	Transocean Offshore
286	P4 Radio Hele Norge	P4 Radio hele Norge
287	Aker Maritime	Aker Maritime
288	Ocean Rig	Ocean Rig
289	Hexagon Composites	Norwegian Applied Technology
290	Tandberg Television	Tandberg Television
291	Thrane–Gruppen	Thrane–Gruppen
292	I.M. Skaugen	I.M. Skaugen
293	ContextVision	ContextVision
294	KredittBanken	KredittBanken
295	Kitron gammel	Kitron
296	Choice Hotels Scandinavia	Choice Hotels Scandinavia
297	Roxar	CorrOcean
298	Subsea 7	Stolt Comex Seaway
299	Roxar	Multi–Fluid
300	EDB – Elekt.	EDB – Elekt. Databeh.
301	Technor	Technor
302	Norsk Lotteridrift	Norsk Lotteridrift
303	Royal Caribbean Cruises	Royal Caribbean Cruises (RCCL)
304	Tordenskjold	Tordenskjold Shipping
305	Byggma	Norsk Wallboard
306	AF Gruppen	AF Gruppen A
307	Fred. Olsen Energy	Fred. Olsen Energy
308	Hjellegjerde	Hjellegjerde
309	Solstad Farstad	Solstad Offshore
310	TGS–NOPEC Geophysical Company	Nopec International
311	VMetro	VMetro
312	Ignis	Logisoft
313	Aktiv Kapital	Aktiv Inkasso
314	Data Respons	Motegruppen
315	Linde–Group	Fredrik Lindegaard

316	Evercom Network	Evercom Network
317	Team Shipping	Team Shipping
318	Kitron	Sonec
319	Aker BioMarine	Natural
320	Voice	Voice
321	Luxo	Luxo
322	Industrifinans NÊringseiendom	Industrifinans N?űringseiendom
323	Hydralift B	Hydralift B
324	Profdoc	Profdoc
325	Rieber Shipping	Rieber Shipping
326	Amersham	Amersham
327	Norsk Kjŕkkeninvest	Norsk Kj??kkeninvest
328	Stolt Offshore A	Stolt Offshore A
329	Synn??ve Finden	Synn??ve Finden
330	Otrum	Otrum
331	Eltek	Eltek
332	Nortrans Offshore	Nortrans Offshore
333	Software Innovation	Software Innovation
334	Axis–Shield	Axis–Shield
335	Enitel	Enitel
336	EVRY	EVRY
337	StepStone	StepStone
338	Expert	Expert
339	Solon Eiendom	Solon Eiendom
340	Photocure	Photocure
341	InFocus Corporation	InFocus Corporation
342	TeleComputing	TeleComputing
343	Zenitel	Zenitel
344	DOF	DOF
345	Komplett	Komplett
346	Office Line	Office Line
347	Telenor	Telenor
348	Sinvest	Sinvest
349	StrongPoint	StrongPoint
350	Fast Search & Transfer	Fast Search & Transfer
351	SAS AB	SAS AB
352	Golar LNG	Golar LNG
353	Hiddn Solutions	Hiddn Solutions
354	PA Resources	PA Resources
355	Q–Free	Q–Free
356	Ler??y Seafood Group	Ler?Éň?y Seafood Group
357	Techstep	Techstep
358	Subsea 7	Subsea 7
359	Troms Fylkes Dampskibsselskap	Troms Fylkes Dampskibsselskap
360	Norwegian Air Shuttle	Norwegian Air Shuttle
361	NextGenTel Holding	NextGenTel Holding
362	Opera Software	Opera Software
363	Yara International	Yara International
364	Akastor	Akastor
365	Mamut	Mamut
366	Medistim	Medistim
367	STX Europe	STX Europe
368	Jason Shipping	Jason Shipping

369	Norman	Norman
370	Aker	Aker
371	Sevan Marine	Sevan Marine
372	Golden Ocean Group	Golden Ocean Group
373	Bj??rge	Bj?Éň?rge
374	Gaming Innovation Group	Gaming Innovation Group
375	Petrojack	Petrojack
376	GC Rieber Shipping	GC Rieber Shipping
377	Wilson	Wilson
378	APL	APL
379	Imarex	Imarex
380	COSL Drilling Europe AS	Awilco Offshore
381	Vizrt	Vizrt
382	Havfisk	Havfisk
383	Havila Shipping	Havila Shipping
384	Questerre Energy Corporation	Questerre Energy Corporation
385	Kongsberg Automotive	Kongsberg Automotive
386	Eidesvik Offshore	Eidesvik Offshore
387	Wintershall Norge ASA	Wintershall Norge ASA
388	Wentworth Resources	Wentworth Resources
389	American Shipping Company	American Shipping Company
390	Siem Offshore	Siem Offshore
391	Seadrill	Seadrill
392	Unison Forsikring	Unison Forsikring
393	Powel	Powel
394	Biotec Pharmacon	Biotec Pharmacon
395	Norstat	Norstat
396	Cermaq	Cermaq
397	BW Gas	Bergesen d.y. A–aksjer
398	Grenland Group	Grenland Group
399	Fairstar Heavy Transport	Fairstar Heavy Transport
400	Odim	Odim
401	DOF Subsea	DOF Subsea
402	Confirmit	Confirmit
403	DeepOcean	DeepOcean
404	Funcom	Funcom
405	Reservoir Exploration	Reservoir Exploration Technology
406	Petrobank Energy and Resources	Petrobank Energy and Resources
407	Trefoil	Trefoil
408	Aker Drilling	Aker Drilling
409	Scorpion Offshore	Scorpion Offshore
410	Songa Offshore	Songa Offshore
411	SeaBird Exploration	SeaBird Exploration
412	BWG Homes	BWG Homes
413	Navamedic	Navamedic
414	Hurtigruten	Hurtigruten
415	REC Silicon	REC Silicon
416	BW Offshore Limited	BW Offshore Limited
417	Weifa	Weifa
418	Odfjell Invest	Odfjell Invest
419	NextGenTel Holding	NextGenTel Holding
420	InterOil Exploration and Production	InterOil Exploration and Production
421	AGR Group	AGR Group
422	Aker Floating Production	Aker Floating Production

423	Teekay Petrojarl	Teekay Petrojarl
424	Austevoll Seafood	Austevoll Seafood
425	Marine Farms	Marine Farms
426	Codfarmers	Codfarmers
427	Norwegian Property	Norwegian Property
428	AKVA Group	AKVA Group
429	Det norske oljeselskap	Det norske oljeselskap
430	Eitzen Chemical	Eitzen Chemical
431	Deep Sea Supply	Deep Sea Supply
432	Copeinca	Copeinca
433	Comrod Communication	Comrod Communication
434	NEAS	NEAS
435	Algeta	Algeta
436	Electromagnetic Geoservices	Electromagnetic Geoservices
437	Rem Offshore	Rem Offshore
438	Protector Forsikring	Protector Forsikring
439	Bouvet	Bouvet
440	MARITIME INDUSTRIAL SERVICES	MARITIME INDUSTRIAL SERVICES
441	SalMar	SalMar
442	Hunter Group	Badger Explorer
443	Grieg Seafood	Grieg Seafood
444	Tribona	Tribona
445	Aker BP	Aker BP
446	London Mining	London Mining
447	Dockwise	Dockwise
448	Pronova BioPharma	Pronova BioPharma
449	Northern Offshore	Northern Offshore
450	Norwegian Energy Company	Norwegian Energy Company
451	Aqua Bio Technology	Aqua Bio Technology
452	NattoPharma	NattoPharma
453	Infratek	Infratek
454	Philly Shipyard	Philly Shipyard
455	Camposol Holding	Camposol Holding
456	Norway Pelagic	Norway Pelagic
457	Prosafe Production Public	Prosafe Production Public
458	PCI Biotech Holding	PCI Biotech Holding
459	Spectrum	Spectrum
460	Havila Ariel	Havila Ariel
461	Borgestad Industries	Borgestad Industries
462	Polaris Media	Polaris Media
463	FLEX LNG	FLEX LNG
464	Bakkafrost	Bakkafrost
465	S??lvtrans	S?Éň?lvtrans
466	Bridge Energy	Bridge Energy
467	Avocet Mining	Avocet Mining
468	Morpol	Morpol
469	Wallenius Wilhelmsen Logistics	Wilh. Wilhelmsen
470	Storm Real Estate	Storm Real Estate
471	Archer	Archer
472	Gjensidige Forsikring	Gjensidige Forsikring
473	Prospector Offshore Drilling	Prospector Offshore Drilling
474	Norway Royal Salmon	Norway Royal Salmon
475	Awilco Drilling	Awilco Drilling

476	H??egh LNG Holdings	H?Éň?egh LNG Holdings
477	Kv?űrner	Kv?Éňűrner
478	Awilco LNG	Awilco LNG
479	SpareBank 1 SR–Bank	SpareBank 1 SR–Bank
480	Selvaag Bolig	Selvaag Bolig
481	Borregaard	Borregaard
482	Asetek	Asetek
483	EAM Solar	EAM Solar
484	Ocean Yield	Ocean Yield
485	Odfjell Drilling	Odfjell Drilling
486	BW LPG	BW LPG
487	Napatech	Napatech
488	Link Mobility Group	Link Mobility Group
489	Atlantic Petroleum	Atlantic Petroleum
490	Tanker Investments	Tanker Investments
491	Avance Gas Holding	Avance Gas Holding
492	Magseis	Magseis
493	Zalaris	Zalaris
494	NEXT Biometrics Group	NEXT Biometrics Group
495	Cxense	Cxense
496	Havyard Group	Havyard Group
497	Aurora LPG Holding	Aurora LPG Holding
498	Aker Solutions	Aker Solutions
499	Scatec Solar	Scatec Solar
500	XXL	XXL
501	Entra	Entra
502	RenoNorden	RenoNorden
503	Team Tankers International	Team Tankers International
504	Nordic Nanovector	Nordic Nanovector
505	Multiconsult	Multiconsult
506	Schibsted ser. B	Schibsted ser. B
507	Vistin Pharma	Vistin Pharma
508	Europris	Europris
509	Pioneer Property Group	Pioneer Property Group
510	Sbanken	Skandiabanken
511	Kid	Kid

Appendix D Script: Modified augmented Dickey-Fuller test

The following code is the custom function we created by modifying the augmented Dickey-Fuller test from `tseries` (Trapletti & Hornik, 2018) drawing inspiration from the augmented Dickey-Fuller test in `aTSA` (Qiu, 2015). The critical values are from Table 4.2(b) p. 103 in Banerjee et al. (1993).

```
# The following function is a modified version of tseries::adf.test created with
# inspiration from aTSA::adf.test.

function (x, alternative = c("stationary", "explosive"), k = trunc((length(x) -
                                                                    1)^(1/3)))
{
  if ((NCOL(x) > 1) || is.data.frame(x))
    stop("x is not a vector or univariate time series")
  if (any(is.na(x)))
    stop("NAs in x")
  if (k < 0)
    stop("k negative")
  alternative <- match.arg(alternative)
  DNAME <- deparse(substitute(x))
  k <- k + 1
  x <- as.vector(x, mode = "double")
  y <- diff(x)
  n <- length(y)
  z <- embed(y, k)
  yt <- z[, 1]
  xt1 <- x[k:n]
  if (k > 1) {
    yt1 <- z[, 2:k]
    res <- lm(yt ~ xt1 + 1 + yt1)
  }
  else res <- lm(yt ~ xt1 + 1)
  res.sum <- summary(res)
  STAT <- res.sum$coefficients[2, 1]/res.sum$coefficients[2,2]

  # From Table 4.2 (b), p. 103 of Banerjee et al. (1993)
  # A. Banerjee, J. J. Dolado, J. W. Galbraith, and D. F. Hendry (1993):
  # Cointegration, Error Correction, and the Econometric Analysis of
  # Non-Stationary Data, Oxford University Press, Oxford.

  table <- rbind(c(-3.75, -3.33, -3.00, -2.63, -0.37,  0.00, 0.34, 0.72),
                 c(-3.58, -3.22, -2.93, -2.60, -0.40, -0.03, 0.29, 0.66),
                 c(-3.51, -3.17, -2.89, -2.58, -0.42, -0.05, 0.26, 0.63),
                 c(-3.46, -3.14, -2.88, -2.57, -0.42, -0.06, 0.24, 0.62),
                 c(-3.44, -3.13, -2.87, -2.57, -0.43, -0.07, 0.24, 0.61),
                 c(-3.43, -3.12, -2.86, -2.57, -0.44, -0.07, 0.23, 0.60))

  tablen <- dim(table)[2]
  tableT <- c(25, 50, 100, 250, 500, 1e+05)
  tablep <- c(0.01, 0.025, 0.05, 0.1, 0.9, 0.95, 0.975, 0.99)
  tableipl <- numeric(tablen)
```

```
  for (i in (1:tablen)) tableipl[i] <- approx(tableT, table[,i], n, rule = 2)$y

  interpol <- approx(tableipl, tablep, STAT, rule = 2)$y
  if (is.na(approx(tableipl, tablep, STAT, rule = 1)$y))
    if (interpol == min(tablep))
      warning("p-value smaller than printed p-value")
  else warning("p-value greater than printed p-value")
  if (alternative == "stationary")
    PVAL <- interpol
  else if (alternative == "explosive")
    PVAL <- 1 - interpol
  else stop("irregular alternative")
  PARAMETER <- k - 1
  METHOD <- "Augmented Dickey-Fuller Test"
  names(STAT) <- "Dickey-Fuller"
  names(PARAMETER) <- "Lag order"
  structure(list(statistic = STAT, parameter = PARAMETER, alternative = alternative,
                 p.value = PVAL, method = METHOD, data.name = DNAME),
            class = "htest")
}
```

Appendix E Script: Creating turnover variable

The following script was used to create the measure turnover.

```
# lage turnover

load(file = "fully_filtrated_data.RData")

# save to new variable, so we don't mess up the original one :)
data <- merged_data_full


# create column, and set first value
data$NoShares_lastmonth <- numeric(length = length(data$NoShares))
data$NoShares_lastmonth[1] <- NA


# initialize temporary variables, and set equal to first row in data
temp_noshares_lastmonth <- data$NoShares_lastmonth[1]
temp_ymid <- data$ym_id[1]
temp_OBI <- data$OBI.security.ID[1]

imax <- length(data$NoShares_lastmonth)


# code takes about 1h 20 minutes
for(i in 2:imax){

  if (data$OBI.security.ID[i] != temp_OBI) {
    temp_noshares_lastmonth <- NA
    temp_ymid <- data$ym_id[i]
    temp_OBI <- data$OBI.security.ID[i]

  }else if (data$ym_id[i] != temp_ymid) {
    temp_noshares_lastmonth <- data$NoShares[i-1]
    temp_ymid <- data$ym_id[i]
  }

  data$NoShares_lastmonth[i] <- temp_noshares_lastmonth

  setTxtProgressBar(txtProgressBar(min = 0, max = imax, style = 3), i)

}


rm(i, imax, temp_noshares_lastmonth, temp_OBI, temp_ymid)


# create turnover
turnover_data <- data %>% mutate(Turnover = Volume / NoShares_lastmonth)
```

```
save(turnover_data, file = "merged_data_w_turnover.RData")

write.csv(turnover_data, "merged_data_w_turnover.csv")
```

Appendix F Script: Data analysis

The following code was used for the analysis in the thesis.

```
##################################################################
#               Title: Master thesis - Analysis
#               Author: Jan Petter Iversen & Astri Skjesol
#               Last update: August 10 - 2018
#               Approx. time to run full script: > 20 minutes
#
#               Requirements:
#                 * Datasets:
#                     - fully_filtrated_data.RData
#
#                 * Packages
#                     - See: Setup
##################################################################


######################## Setup ########################

setwd("C:/Users/Lokal/Desktop/Data Master Thesis")

set.seed(19503) # for reproducability

library(tibble)
library(e1071)
library(stargazer) # for latex code
library(normtest)
library(tseries) # for adf.test
library(gmm)
library(plm)
library(systemfit)
library(rugarch)
library(lmtest)
library(plyr)
library(dplyr)
library(robustHD) # for winsorization
library(pracma) # for removing linear trend

######################## Import, create, & transform data ########################

load(file = "merged_data_w_turnover.RData")

# save to new variable, so we don't mess up the original one :)
data <- turnover_data


# create list of uniqe OBI IDs
OBI_list <- unique(data$OBI.security.ID)
```

```
# Multiply return and turnover by 100 to get it as a percentage.
# Will improve readability in tables.
data$Return <- data$Return * 100
data$Turnover <- data$Turnover * 100

# as turnover introduced NAs (about 1% of the sample), we remove them
data <- na.omit(data)


######################## Descriptive - whole sample ########################

# print latex-code for descriptive statistics for the whole sample
stargazer(as.data.frame(data[,c("Return", "Turnover")]),
          summary = T,
          summary.stat = c("mean", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2, # if 2 digits round to 0, we can increase to max 4 digits
          align = T,
          colnames = T,
          column.sep.width = "0pt", # space between columns in table
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table") # use "sidewaystable" for flipping the table


# print boxplot of return and utrnover
#pdf(file = "ret_turnover_boxplot_notrim.pdf", width = 8, height = 4)
#boxplot(data[,c("Return", "Turnover")], col = "#316ba0", border = "#123456")
#dev.off()

######################## Descriptive - single securities ########################


# create df with summary statistics of returns of each security
return_df <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(Count = n(),
            Mean = mean(Return),
            St.Dev. = sd(Return),
            Max = max(Return),
            Pctl_75 = quantile(Return , probs = 0.75),
            Median = median(Return),
            Pctl_25 = quantile(Return , probs = 0.25),
            Min = min(Return),
            Kurtosis = kurtosis(Return),
            Skewness = skewness(Return))



# print latex-code for descriptive statistics for individual securities
stargazer(as.data.frame(return_df[,c(3:11)]),
          summary = T,
```

```
          summary.stat = c("mean", "sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2, # max 2+2 = 4 digits
          align = T,
          colnames = T,
          column.sep.width = "0pt", # space between columns in table
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table") # use "sidewaystable" for flipping the table


# create df with summary statistics of Turnover of each security
turnover_df <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(Count = n(),
            Mean = mean(Turnover),
            St.Dev. = sd(Turnover),
            Max = max(Turnover),
            Pctl_75 = quantile(Turnover , probs = 0.75),
            Median = median(Turnover),
            Pctl_25 = quantile(Turnover , probs = 0.25),
            Min = min(Turnover),
            Kurtosis = kurtosis(Turnover),
            Skewness = skewness(Turnover))



# print latex-code for descriptive statistics for individual securities
# subtable 1
stargazer(as.data.frame(turnover_df[,c(3:6)]),
          summary = T,
          summary.stat = c("mean", "sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 0,
          align = T,
          colnames = T,
          column.sep.width = "0pt", # space between columns in table
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table") # use "sidewaystable" for flipping the table

# subtable 2
stargazer(as.data.frame(turnover_df[,c(7:11)]),
          summary = T,
          summary.stat = c("mean", "sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 0,
          align = T,
          colnames = T,
```

```
          column.sep.width = "0pt", # space between columns in table
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table") # use "sidewaystable" for flipping the table


# remove variables we no longer need
rm(turnover_df, return_df)

####################### Winsorization & descriptive #######################

data$Win_return <- numeric(length = length(data$Return))
data$Win_turnover <- numeric(length = length(data$Turnover))

for (stock in OBI_list) {

  winsorized <- data %>%
    filter(OBI.security.ID == stock) %>%
    select(Return, Turnover) %>%
    as.matrix() %>%
    winsorize(fallback = TRUE, prob = 0.99)

  data$Win_return[data$OBI.security.ID == stock] <- winsorized[ ,1]
  data$Win_turnover[data$OBI.security.ID == stock] <- winsorized[ ,2]

}

rm(stock, winsorized)

# remove negatrive turnover-values introduces by winsorization
data$Win_turnover[data$Win_turnover < 0] <- 0

# print summary statistics of the whole sample - winsorized
stargazer(as.data.frame(data[,c("Win_return", "Win_turnover")]),
          summary = T,
          summary.stat = c("mean", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2, # if 2 digits round to 0, we can increase to max 4 digits
          align = T,
          colnames = T,
          column.sep.width = "0pt", # space between columns in table
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table") # use "sidewaystable" for flipping the table



# create df with summary statistics of winsorized returns of each security
win_return_df <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(Count = n(),
```

```
            Mean = mean(Win_return),
            St.Dev. = sd(Win_return),
            Max = max(Win_return),
            Pctl_75 = quantile(Win_return , probs = 0.75),
            Median = median(Win_return),
            Pctl_25 = quantile(Win_return , probs = 0.25),
            Min = min(Win_return),
            Kurtosis = kurtosis(Win_return),
            Skewness = skewness(Win_return))



# print latex-code for descriptive statistics for individual securities
stargazer(as.data.frame(win_return_df[,c(3:11)]),
          summary = T,
          summary.stat = c("mean", "sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2, # max 2+2 = 4 digits
          align = T,
          colnames = T,
          column.sep.width = "0pt", # space between columns in table
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table") # use "sidewaystable" for flipping the table


# create df with summary statistics of Turnover of each security
win_turnover_df <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(Count = n(),
            Mean = mean(Win_turnover),
            St.Dev. = sd(Win_turnover),
            Max = max(Win_turnover),
            Pctl_75 = quantile(Win_turnover , probs = 0.75),
            Median = median(Win_turnover),
            Pctl_25 = quantile(Win_turnover , probs = 0.25),
            Min = min(Win_turnover),
            Kurtosis = kurtosis(Win_turnover),
            Skewness = skewness(Win_turnover))



# print latex-code for descriptive statistics for individual securities
# subtable 1
stargazer(as.data.frame(win_turnover_df[,c(3:6)]),
          summary = T,
          summary.stat = c("mean", "sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 0,
          align = T,
          colnames = T,
```

```
          column.sep.width = "0pt", # space between columns in table
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table") # use "sidewaystable" for flipping the table

# subtable 2
stargazer(as.data.frame(win_turnover_df[,c(7:11)]),
          summary = T,
          summary.stat = c("mean", "sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 0,
          align = T,
          colnames = T,
          column.sep.width = "0pt", # space between columns in table
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table") # use "sidewaystable" for flipping the table


# remove variables we no longer need
rm(win_turnover_df, win_return_df)




####################### Jarque-Bera #######################

## Returns
# do a Jarque-Bera-test for each security, and save the p-value to a df
jarque_bera_test <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(JB = jb.norm.test(Win_return)$p.value)


# still, we check whether any securites have significant
# JB-statistics on 5% and 10% level. None do.
ifelse(jarque_bera_test$JB > 0.05, 1,0) %>% sum()

# remove variables we do not need anymore
rm(jarque_bera_test)

## Volume
# do a Jarque-Bera-test for each security, and save the p-value to a df
jarque_bera_test2 <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(JB = jb.norm.test(Win_turnover)$p.value)

# we check whether any securites have significant
# JB-statistics on 5% level. Some do:
ifelse(jarque_bera_test2$JB > 0.01, 1,0) %>% sum()
```

```
# remove variables we do not need anymore
rm(jarque_bera_test2)

####################### Ljung-Box & Autocorrelation #######################


# do a Ljung-Box-test for each security, and save the p-value to a df
ljung_box_test <-  data %>%
  group_by(OBI.security.ID) %>%
  summarise(LB = Box.test(Win_return, lag = 10, type = "Ljung-Box")$p.value)


# gives percentage of companies with no auto correlation
# in the 10 first lags for 1% and 5% significant levels
ifelse(ljung_box_test$LB > 0.01, 1,0) %>% sum()*100/511
ifelse(ljung_box_test$LB > 0.05, 1,0) %>% sum()*100/511

# removes dfs we no longer need
rm(ljung_box_test)


# do a Ljung-Box-test for each security, and save the p-value to a df
ljung_box_test2 <-  data %>%
  group_by(OBI.security.ID) %>%
  summarise(LB = Box.test(Win_turnover, lag = 10, type = "Ljung-Box")$p.value)

# gives percentage of companies with no auto correlation
# in the 10 first lags for 1% and 5% significant levels
ifelse(ljung_box_test2$LB > 0.01, 1,0) %>% sum()*100/511
ifelse(ljung_box_test2$LB > 0.05, 1,0) %>% sum()*100/511

# removes dfs we no longer need
rm(ljung_box_test2)

## check the order of autocorrelation for stocks

# create an empty df for the OBI IDs and number of significant lags
df_sig_lag <- data.frame(OBI.security.ID = numeric(), sign_lag = numeric())

# use loop to calculate significant lags and fill the df
for (stock in OBI_list) { #loop through all stocks

  # define sample size for each stock
  large_T <- data %>% filter(OBI.security.ID == stock) %>% nrow()

  # create a (16,1) matrix with the autocorrelation from lag 0 to 15
  M <- data %>%
    filter(OBI.security.ID == stock) %>%
    pull(Win_return) %>%
    acf(plot = F, lag.max = 15) %>%
    `$`(acf)

  for (i in 2:16) { # loop through the 15 lags
```

```
      if (abs(M[i]) < (1.96/sqrt( large_T - (i-1) ) ) ) {
        # if the absolute value of the autocorrelation is
        # under the significant treshold,
        # then save the lag before to the
        # df (as this was nececerrely over the treshold)
        df_sig_lag <- rbind(df_sig_lag,
                            data.frame(OBI.security.ID = stock, sign_lag = (i-2)))
        break() # exit inner loop

      }else if (i == 16) { # if some have more than 15 significant lags, write NA to df

        df_sig_lag <- rbind(df_sig_lag,
                            data.frame(OBI.security.ID = stock, sign_lag = NA))

      }
    }
  }

# remove variables we no longer need
rm(i, large_T, M, stock)

# look for number of significant lags for stocks who has significant lags at all
df_sig_lag %>% filter(!sign_lag == 0) %>% pull(sign_lag) %>% mean # mean of 1.4
df_sig_lag %>% filter(!sign_lag == 0) %>% pull(sign_lag) %>% median # median of 1



# create an empty df for the OBI IDs and number of significant lags
df_sig_lag2 <- data.frame(OBI.security.ID = numeric(), sign_lag = numeric())

# use loop to calculate significant lags and fill the df
for (stock in OBI_list) { #loop through all stocks

  # define sample size for each stock
  large_T <- data %>% filter(OBI.security.ID == stock) %>% nrow()

  # create a (16,1) matrix with the autocorrelation from lag 0 to 15
  M <- data %>%
    filter(OBI.security.ID == stock) %>%
    pull(Win_turnover) %>%
    acf(plot = F, lag.max = 100) %>%
    `$`(acf)

  for (i in 2:101) { # loop through the 30 lags


    if (abs(M[i]) < (1.96/sqrt( large_T - (i-1) ) ) ) {
      # if the absolute value of the autocorrelation
      # is under the significant treshold,
      # then save the lag before to the df
      # (as this was nececerrely over the treshold)
```

```
      df_sig_lag2 <- rbind(df_sig_lag2,
                          data.frame(OBI.security.ID = stock, sign_lag = (i-2)))
      break() # exit inner loop

    }else if (i == 101) { # if some have more than 15 significant lags, write NA to df

      df_sig_lag2 <- rbind(df_sig_lag2,
                          data.frame(OBI.security.ID = stock, sign_lag = NA))

    }
  }
}

# remove variables we no longer need
rm(i, large_T, M, stock)

# look for number of significant lags for stocks who has significant lags at all
df_sig_lag2 %>% filter(!sign_lag == 0) %>% pull(sign_lag) %>% mean # mean of 11.8
df_sig_lag2 %>% filter(!sign_lag == 0) %>% pull(sign_lag) %>% median # median of 10

rm(df_sig_lag, df_sig_lag2)



######################## PhillipsâĂŞPerron test ########################

# do a PP-test for each security, and save the p-value to a df
pp_test <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(PP = pp.test(Win_return)$p.value)

# We check for p-values above 0.01. None are found
pp_test %>% filter(PP > 0.01)

# do a PP-test for each security, and save the p-value to a df
pp_test2 <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(PP = pp.test(Win_turnover)$p.value)

# We check for p-values above 0.01. None are found
pp_test2 %>% filter(PP > 0.01)

rm(pp_test, pp_test2)


######################## Augmented Dickey-Fuller ########################

# load the modified ADF-test
load("jp_adf.test.Rdata")

# do a ADF-test for each security, and save the p-value to a df
augmented_dickey_fuller_test <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(ADF = jp_adf.test(Win_return)$p.value)
```

```
# as most has a p.value below 0.01, we only check those above
augmented_dickey_fuller_test %>% filter(ADF > 0.01)


# remove variables we do not need anymore
rm(augmented_dickey_fuller_test)

# then, we check for volume
augmented_dickey_fuller_test2 <- data %>%
  group_by(OBI.security.ID) %>%
  summarise(ADF = jp_adf.test(Win_turnover)$p.value)

# 20 stocks have p-value above 0.05, we save them to a df
maybe_non_stationary <- augmented_dickey_fuller_test2 %>% filter(ADF > 0.05)

# a loop for checking the companies
for (i in 1:length(maybe_non_stationary$OBI.security.ID)) {

  temp_title <- data %>%
    filter(OBI.security.ID == maybe_non_stationary$OBI.security.ID[i]) %>%
    select(Last.Company.Name) %>% unique() %>% pull()

  data %>%
    filter(OBI.security.ID == maybe_non_stationary$OBI.security.ID[i]) %>%
    select(Last.Company.Name) %>%
    unique() %>%
    pull() %>%
    print()

  data %>%
    filter(OBI.security.ID == maybe_non_stationary$OBI.security.ID[i]) %>%
    select(`Last Security Name`) %>%
    unique() %>%
    pull() %>%
    print()

  plot(data$Date[data$OBI.security.ID == maybe_non_stationary$OBI.security.ID[i]],
       data$Volume[data$OBI.security.ID == maybe_non_stationary$OBI.security.ID[i]],
       type = "l",
       main = paste(temp_title, " volume"))

  plot(data$Date[data$OBI.security.ID == maybe_non_stationary$OBI.security.ID[i]],
       data$Turnover[data$OBI.security.ID ==
                          maybe_non_stationary$OBI.security.ID[i]],
       type = "l",
       main = paste(temp_title, " turnover"))

  plot(data$Date[data$OBI.security.ID == maybe_non_stationary$OBI.security.ID[i]],
       data$Win_turnover[data$OBI.security.ID ==
                  maybe_non_stationary$OBI.security.ID[i]],
       type = "l",
       main = paste(temp_title, " winsorised turnover"))
```

```
}

# remove variables we no longer need
rm(temp_title, i)


# we decided to remove those companies which were non-stationary
data <- data %>% filter(!(OBI.security.ID %in% maybe_non_stationary$OBI.security.ID))

# update list of uniqe OBI IDs
OBI_list <- unique(data$OBI.security.ID)

# check if this removed the problem
augmented_dickey_fuller_test2 <-  data %>%
  group_by(OBI.security.ID) %>%
  summarise(ADF = jp_adf.test(Win_turnover)$p.value)

augmented_dickey_fuller_test2 %>% filter(ADF > 0.05)

# This removed the problem, and we no longer have any non-stationary turnover series.

# remove variables we do not need anymore
rm(augmented_dickey_fuller_test2, maybe_non_stationary, jp_adf.test)

######################## Detrending #######################

# initialize column
data$Win_dtrnd_turnover <- numeric(length = length(data$Win_turnover))

# remove linear trend
for (stock in OBI_list) {

  detrended <- data %>%
    filter(OBI.security.ID == stock) %>%
    pull(Win_turnover) %>%
    detrend(tt = "linear")

  data$Win_dtrnd_turnover[data$OBI.security.ID == stock] <- detrended

}

# remove variables we no longer need
rm(stock, detrended)


######################## Cross correlation #######################

## Turnover & Return

M_crosscorr <- matrix(nrow = 0, ncol = 9)
M_significance <- matrix(nrow = 0, ncol = 9)

# use loop to calculate significant lags and fill the df
for (stock in OBI_list) { #loop through all stocks
```

```
  temp_acf_object <- ccf(data$Win_dtrnd_turnover[data$OBI.security.ID == stock],
                         data$Win_return[data$OBI.security.ID == stock],
                         lag.max = 4,
                         plot = F)

  temp_crosscorr <- temp_acf_object$acf %>% as.matrix() %>% t()

  # find significance (qnorm of 1 + 0.95 for 5% significance level)
  temp_significance <- qnorm((1.95)/2)/sqrt(temp_acf_object$n.used) %>%
    rep(9) %>%
    as.matrix() %>%
    t()


  M_crosscorr <- rbind(M_crosscorr, temp_crosscorr)

  M_significance <- rbind(M_significance, temp_significance)

}

crosscorr_names <- paste("j=", c(-4:4), sep = "")

df_crosscorr <- as.data.frame(M_crosscorr)
df_significance <- as.data.frame(M_significance)

names(df_crosscorr) <- crosscorr_names
names(df_significance) <- crosscorr_names

stargazer(df_crosscorr,
          summary = T,
          summary.stat = c("max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")


# print histograms of cross correlation
pdf(file = "crosscorr_hist.pdf", width = 8, height = 8)
# set 3x3 window of graphs
par(mfrow = c(3,3))
for(i in 1:9){
  hist(df_crosscorr[,i],
       breaks = 20,
       main = NULL,
       xlab = names(df_crosscorr)[i],
```

```
        ylab = NULL,
        xlim = c(-0.20,0.30),
        ylim = c(0,120),
        col = "#316ba0",
        border = "#123456")
}
dev.off()

#reset graph window
par(mfrow = c(1,1))

# number/percent of signifiant cross correlations per lag
Number <- apply((abs(df_crosscorr) > df_significance),2, sum)
Percentage <- (apply((abs(df_crosscorr) > df_significance),2, sum)
               *100/length(OBI_list)) %>%
  round(2)


stargazer(as.data.frame(rbind(Number, Percentage)),
          summary = F,
          flip = F,
          digits = 2,
          digits.extra = 0,
          align = F,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")


## turnover & Squared Return

M_crosscorr2 <- matrix(nrow = 0, ncol = 9)
M_significance2 <- matrix(nrow = 0, ncol = 9)

# use loop to calculate significant lags and fill the df
for (stock in OBI_list) { #loop through all stocks


  temp_acf_object2 <-  ccf(data$Win_dtrnd_turnover[data$OBI.security.ID == stock],
                           (data$Return[data$OBI.security.ID == stock])^2,
                           lag.max = 4,
                           plot = F)

  temp_crosscorr2 <- temp_acf_object2$acf %>% as.matrix() %>% t()

  temp_significance2 <- qnorm((1.95)/2)/sqrt(temp_acf_object2$n.used) %>%
    rep(9) %>%
    as.matrix() %>%
    t()

  M_crosscorr2 <- rbind(M_crosscorr2, temp_crosscorr2)
```

```
  M_significance2 <- rbind(M_significance2, temp_significance2)

}


df_crosscorr2 <- as.data.frame(M_crosscorr2)
df_significance2 <- as.data.frame(M_significance2)
names(df_crosscorr2) <- crosscorr_names
names(df_significance2) <- crosscorr_names



stargazer(df_crosscorr2,
          summary = T,
          summary.stat = c("max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")



# print histograms of cross correlation
pdf(file = "crosscorr_hist2.pdf", width = 8, height = 8)
# set 3x3 window of graphs
par(mfrow = c(3,3))
for(i in 1:9){
  hist(df_crosscorr2[,i],
       breaks = 20,
       main = NULL,
       xlab = names(df_crosscorr2)[i],
       ylab = NULL,
       xlim = c(-0.15,0.45),
       ylim = c(0,130),
       col = "#316ba0",
       border = "#123456")
}
dev.off()

#reset graph window
par(mfrow = c(1,1))

# number/percent of signifiant cross correlations per lag
Number <- apply((abs(df_crosscorr2) > df_significance2),2, sum)
Percentage <- (apply((abs(df_crosscorr2) > df_significance2),2, sum)
               *100/length(OBI_list)) %>%
  round(2)
```

```
stargazer(as.data.frame(rbind(Number, Percentage)),
          summary = F,
          flip = F,
          digits = 2,
          digits.extra = 0,
          align = F,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")



rm(M_crosscorr, M_crosscorr2, crosscorr_names,
   temp_crosscorr, temp_crosscorr2,df_crosscorr, df_crosscorr2, stock, i,
   M_significance, M_significance2, temp_acf_object, temp_acf_object2,
   temp_significance, temp_significance2, df_significance, df_significance2,
   Number, Percentage)


######################## Lee & Rui model ########################


# create empty matrices for storing estimates later
step1_estimates <- matrix(nrow = 0, ncol = 4)
step2_estimates <- matrix(nrow = 0, ncol = 4)
step1_pvalues <- matrix(nrow = 0, ncol = 4)
step2_pvalues <- matrix(nrow = 0, ncol = 4)
step1_tvalues <- matrix(nrow = 0, ncol = 4)
step2_tvalues <- matrix(nrow = 0, ncol = 4)

# use loop to do a two-step least squared regression
for (stock in OBI_list) { #loop through all stocks

  # subset and mutate data we will use in each iteration
  temp_lee_rui_data <- data %>%
    filter(OBI.security.ID == stock) %>%
    select(Date, Win_return, Win_dtrnd_turnover) %>%
    mutate(Return_L1 = dplyr::lag(Win_return,1),
           Volume_L1 = dplyr::lag(Win_dtrnd_turnover,1),
           Volume_L2 = dplyr::lag(Win_dtrnd_turnover,2)) %>%
    na.omit()

  # first regression, eq. 1 of Lee & Rui
  step1 <- lm(Win_return ~ Win_dtrnd_turnover + Volume_L1 + Return_L1,
              data = temp_lee_rui_data)

  # extract results we need
  temp_step1_estimates <- step1 %>%
    summary() %>%
    `$`(coefficients) %>%
```

```
    `[`(,1) %>%
    as.matrix() %>%
    t() # estimates


  temp_step1_pvalues <- step1 %>%
    summary() %>%
    `$`(coefficients) %>%
    `[`(,4) %>%
    as.matrix() %>%
    t() # p values

  temp_step1_tvalues <- step1 %>%
    summary() %>%
    `$`(coefficients) %>%
    `[`(,3) %>%
    as.matrix() %>%
    t() # t-values


  # add results to the correct matrix
  step1_estimates <- rbind(step1_estimates, temp_step1_estimates)
  step1_pvalues <- rbind(step1_pvalues, temp_step1_pvalues)
  step1_tvalues <- rbind(step1_tvalues, temp_step1_tvalues)

  # create Return_Hatt as the fitted values from regression 1
  temp_lee_rui_data$Return_Hatt <- predict(step1)

  # second regresseion, eq. 2 of Lee & Rui
  step2 <- lm(Win_dtrnd_turnover ~ Return_Hatt + Volume_L1 + Volume_L2,
              data = temp_lee_rui_data)

  # extract results we need
  temp_step2_estimates <- step2 %>%
    summary() %>%
    `$`(coefficients) %>%
    `[`(,1) %>%
    as.matrix() %>%
    t() # estimates

  temp_step2_pvalues <- step2 %>%
    summary() %>%
    `$`(coefficients) %>%
    `[`(,4) %>%
    as.matrix() %>%
    t() # p values

  temp_step2_tvalues <- step2 %>%
    summary() %>%
    `$`(coefficients) %>%
    `[`(,3) %>%
    as.matrix() %>%
    t() # t-values
```

```
  # add results to the correct matrix
  step2_estimates <- rbind(step2_estimates, temp_step2_estimates)
  step2_pvalues <- rbind(step2_pvalues, temp_step2_pvalues)
  step2_tvalues <- rbind(step2_tvalues, temp_step2_tvalues)

}

# name the variables from the two regressions
step1_names <- c("Intercept","Turnover","Turnover_L1","Return_L1")
step2_names <- c("Intercept","Return_Hatt","Turnover_L1","Turnover_L2")

# save matrices as data frames
df_step1_estimates <- as.data.frame(step1_estimates)
df_step2_estimates <- as.data.frame(step2_estimates)
df_step1_pvalues <- as.data.frame(step1_pvalues)
df_step2_pvalues <- as.data.frame(step2_pvalues)

# give the dfs the correct names
names(df_step1_estimates) <- step1_names
names(df_step1_pvalues) <- step1_names
names(df_step2_estimates) <- step2_names
names(df_step2_pvalues) <- step2_names

# print LaTeX-code for estimates from regression 1
stargazer(df_step1_estimates,
          summary = T,
          summary.stat = c("max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")

# print LaTeX-code for estimates from regression 2
stargazer(df_step2_estimates,
          summary = T,
          summary.stat = c("max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")
```

```
# print percentage of significant variables (5% significance level)
apply((df_step1_pvalues < 0.05), 2, function(x){round((sum(x)*100)/length(x), 1)})
apply((df_step2_pvalues < 0.05), 2, function(x){round((sum(x)*100)/length(x), 1)})


# print histograms of t-stats from step1
pdf(file = "lee_rui_step1_coefficients_hist.pdf", width = 8, height = 8)
# set 2x2 window of graphs
par(mfrow = c(2,2))
for(i in 1:4){
  hist(step1_tvalues[,i],
       breaks = 30,
       main = NULL,
       xlab = c("b0 - Intercept","b1 - Volume","b2 - Volume t-1","b3 - Return t-1")[i],
       ylab = NULL,
       #xlim = c(-35,15),
       #ylim = c(0,120),
       col = "#316ba0",
       border = "#123456")
  abline(v = qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
  abline(v = -qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
}
dev.off()

#reset graph window
par(mfrow = c(1,1))



# print histograms of t-stats from step2
pdf(file = "lee_rui_step2_coefficients_hist.pdf", width = 8, height = 8)
# set 2x2 window of graphs
par(mfrow = c(2,2))
for(i in 1:4){
  temp_hist_data <- step2_tvalues[,i]
  temp_hist_data <- temp_hist_data[(temp_hist_data>-100 & temp_hist_data<100)]
  hist(temp_hist_data,
       breaks = 50,
       main = NULL,
       xlab = c("a0 - Intercept","a1 - Return","a2 - Volume t-1","a3 - Volume t-2")[i],
       ylab = NULL,
       #xlim = c(-10,40),
       #ylim = c(0,120),
       col = "#316ba0",
       border = "#123456")
  abline(v = qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
  abline(v = -qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
}
dev.off()

#reset graph window
par(mfrow = c(1,1))
```

```
# remove variables we no longer need
rm(df_step1_estimates, df_step1_pvalues, df_step2_estimates, df_step2_pvalues,
   step1, step2, step1_names, step2_names, temp_lee_rui_data, temp_step1_estimates,
   temp_step2_estimates, temp_step1_pvalues, temp_step2_pvalues, step1_estimates,
   step1_pvalues, step2_estimates, step2_pvalues, stock, step1_tvalues, step2_tvalues,
   temp_step1_tvalues, temp_step2_tvalues,i, temp_hist_data)


####################### Brailford-extension #######################

# create empty matrices for storing estimates later
brailsford_estimates <- matrix(nrow = 0, ncol = 5)
brailsford_pvalues <- matrix(nrow = 0, ncol = 5)
brailsford_tvalues <- matrix(nrow = 0, ncol = 5)

# use loop to do regression for all stocks
for (stock in OBI_list) { #loop through all stocks

  # subset and mutate data we will use in each iteration
  temp_brailsford_data <- data %>%
    filter(OBI.security.ID == stock) %>%
    select(Date, Win_return, Win_dtrnd_turnover) %>%
    mutate(Volume_L1 = dplyr::lag(Win_dtrnd_turnover,1),
           Volume_L2 = dplyr::lag(Win_dtrnd_turnover,2),
           Dummy = ifelse(Win_return < 0, 1, 0)) %>%
    na.omit()

  # regression
  brailsford <- lm(Win_dtrnd_turnover ~ Volume_L1 + Volume_L2 + I(Win_return^2) +
                     I(Dummy * (Win_return^2)),
                   data = temp_brailsford_data)

  # extract results we need
  temp_brailsford_estimates <- brailsford %>%
    summary() %>%
    `$`(coefficients) %>%
    `[`(,1) %>%
    as.matrix() %>%
    t() # estimates

  temp_brailsford_pvalues <- brailsford %>%
    summary() %>%
    `$`(coefficients) %>%
    `[`(,4) %>%
    as.matrix() %>%
    t() # p values

  temp_brailsford_tvalues <- brailsford %>%
    summary() %>%
    `$`(coefficients) %>%
    `[`(,3) %>%
    as.matrix() %>%
```

```
    t() #  t-values

  # add results to the correct matrix
  brailsford_estimates <- rbind(brailsford_estimates,
                                temp_brailsford_estimates) # estimates

  brailsford_pvalues <- rbind(brailsford_pvalues,
                              temp_brailsford_pvalues) # p values

  brailsford_tvalues <- rbind(brailsford_tvalues,
                              temp_brailsford_tvalues) # t-values

}

# name the variables from the two regressions
brailsford_names <- c("Intercept",
                      "Turnover_1",
                      "Turnover_L2",
                      "Return^2",
                      "Dummy * Return^2")

# save matrices as data frames
df_brailsford_estimates <- as.data.frame(brailsford_estimates)
df_brailsford_pvalues <- as.data.frame(brailsford_pvalues)

# give the dfs the correct names
names(df_brailsford_estimates) <- brailsford_names
names(df_brailsford_pvalues) <- brailsford_names

# print LaTeX-code for estimates
stargazer(df_brailsford_estimates,
          summary = T,
          summary.stat = c("max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")



# print percentage of significant variables (5% significance level)
apply((df_brailsford_pvalues < 0.05), 2, function(x){round((sum(x)*100)/length(x), 1)})

# print histograms of t-stats from step2
pdf(file = "brailsford_coefficients_hist.pdf", width = 8, height = 8)
# set window of graphs
#layout(matrix(c(1,1,2,2,3,3,4,4,4,5,5,5), 2, 6, byrow = TRUE))
layout(matrix(c(5,5,5,5,1,3,2,4), 2, 4, byrow = FALSE))
```

```
for(i in 1:5){
  hist(brailsford_tvalues[,i],
       breaks = 50,
       main = NULL,
       xlab = c("alpha 0 - Intercept",
                "phi 1 - Volume t-1",
                "phi 2 - Volume t-2",
                "alpha 1 - Return^2",
                "alpha 2 - dummy x Return^2")[i],
       ylab = NULL,
       #xlim = c(-10,40),
       #ylim = c(0,120),
       col = "#316ba0",
       border = "#123456")
  abline(v = qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
  abline(v = -qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
}
dev.off()

#reset graph window
par(mfrow = c(1,1))


# remove variables we no longer need
rm(brailsford, brailsford_estimates, brailsford_pvalues, brailsford_names,
   df_brailsford_estimates, df_brailsford_pvalues, temp_brailsford_data,
   temp_brailsford_estimates, temp_brailsford_pvalues, stock, i,
   temp_brailsford_tvalues, brailsford_tvalues)




####################### Restricted GARCH #######################


# RESTRICTED MODEL (NO VOLUME)

# create empty matrices for storing estimates later
garch_restricted_estimates <- matrix(nrow = 0, ncol = 5)
garch_restricted_pvalues <- matrix(nrow = 0, ncol = 5)
garch_restricted_tvalues <- matrix(nrow = 0, ncol = 5)

# set specifications for GARCH model. This is an AR(1)-GARCH(1,1)
# (or, a restricted AR(1)-VA-GARCH(1,1))
spec <- ugarchspec(variance.model = list(model = "sGARCH",
                                         garchOrder = c(1, 1),
                                         external.regressors = NULL),
                   mean.model = list(armaOrder = c(1, 0)),
                   distribution.model = "norm",
                   start.pars = list(),
                   fixed.pars = list())


# use loop to fit model for all stocks
```

```
for (stock in OBI_list) { #loop through all stocks

  # fit model
  garch_restricted <- ugarchfit(spec=spec,
                               data = as.matrix(data$Win_return[data$OBI.security.ID
                                                          == stock]),
                               solver = "hybrid")

  # extract results we need
  temp_garch_restricted_estimates <- garch_restricted@fit$matcoef[,1] %>%
    as.matrix() %>%
    t() # estimates

  temp_garch_restricted_pvalues <- garch_restricted@fit$matcoef[,4] %>%
    as.matrix() %>%
    t() # p values

  temp_garch_restricted_tvalues <- garch_restricted@fit$matcoef[,3] %>%
    as.matrix() %>%
    t() # t-statistic

  # add results to the correct matrix
  garch_restricted_estimates <- rbind(garch_restricted_estimates,
                                      temp_garch_restricted_estimates) # estimates

  garch_restricted_pvalues <- rbind(garch_restricted_pvalues,
                                    temp_garch_restricted_pvalues) # p values

  garch_restricted_tvalues <- rbind(garch_restricted_tvalues,
                                    temp_garch_restricted_tvalues) # t-statistics

}

# name the variables
garch_restricted_names <- c("MU","AR1","Alpha0","Alpha1", "Beta1") # Alpha0 / Omega

# save matrices as data frames
df_garch_restricted_estimates <- as.data.frame(garch_restricted_estimates)
df_garch_restricted_pvalues <- as.data.frame(garch_restricted_pvalues)
df_garch_restricted_tvalues <- as.data.frame(garch_restricted_tvalues)

# give the dfs the correct names
names(df_garch_restricted_estimates) <- garch_restricted_names
names(df_garch_restricted_pvalues) <- garch_restricted_names
names(df_garch_restricted_tvalues) <- garch_restricted_names

# add variable for persitence (Alpha1 + Beta1)
df_garch_restricted_estimates <- df_garch_restricted_estimates %>%
  mutate(alpha1_plus_beta1 = (Alpha1 + Beta1))

# print LaTeX-code for estimates from regression 1
stargazer(df_garch_restricted_estimates[,c("Alpha1", "Beta1", "alpha1_plus_beta1")],
          summary = T,
          summary.stat = c("mean","sd", "max","p75","median", "p25", "min"),
```

```
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")


# print percentage of significant variables (5% significance level)
apply((df_garch_restricted_pvalues < 0.05),
      2,
      function(x){round((sum(x)*100)/length(x),1)})



# print histograms of t-stats from restricted GARCH(1,1)
pdf(file = "hist_restricted_garch.pdf", width = 8, height = 4)
# set 1x2 window of graphs
par(mfrow = c(1,2))
for(i in 1:2){
  hist(df_garch_restricted_tvalues[,(i + 3)] %>%
         subset(df_garch_restricted_tvalues[,(i + 3)] < 100),
       breaks = 50,
       main = NULL,
       xlab = c("alpha 1",
                "beta 1")[i],
       ylab = NULL,
       xlim = c(0,100),
       #ylim = c(0,120),
       col = "#316ba0",
       border = "#123456")
  abline(v = qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
  abline(v = -qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
}
dev.off()

#reset graph window
par(mfrow = c(1,1))


# remove variables we no longer need
rm(df_garch_restricted_pvalues, garch_restricted, garch_restricted_estimates,
   garch_restricted_pvalues, garch_restricted_names, spec, stock,
   temp_garch_restricted_estimates, temp_garch_restricted_pvalues,
   garch_restricted_tvalues, i, temp_garch_restricted_tvalues)



####################### Unrestricted GARCH #######################
```

```
# UNRESTRICTED MODEL (INCLUDING VOLUME)

# create empty matrices for storing estimates later
garch_restricted_estimates2 <- matrix(nrow = 0, ncol = 6)
garch_restricted_pvalues2 <- matrix(nrow = 0, ncol = 6)
garch_restricted_tvalues2 <- matrix(nrow = 0, ncol = 6)

# use loop to fit model for all stocks
for (stock in OBI_list) { #loop through all stocks

  # set specifications for GARCH model. This is an AR(1)-VA-GARCH(1,1)
  spec2 <- ugarchspec(variance.model = list(model = "sGARCH",
                                            garchOrder = c(1, 1),
                                            external.regressors = as.matrix(
                                              data$Win_dtrnd_turnover
                                              [data$OBI.security.ID == stock])),
                      mean.model = list(armaOrder = c(1, 0)),
                      distribution.model = "norm",
                      start.pars = list(),
                      fixed.pars = list())

  # fit model
  garch_restricted2 <- ugarchfit(spec=spec2,
                                 data = as.matrix(data$Win_return[data$OBI.security.ID
                                                                  == stock]),
                                 solver = "hybrid")

  # extract results we need
  temp_garch_restricted_estimates2 <- garch_restricted2@fit$matcoef[,1] %>%
    as.matrix() %>%
    t() # estimates

  temp_garch_restricted_pvalues2 <- garch_restricted2@fit$matcoef[,4] %>%
    as.matrix() %>%
    t() # p values

  temp_garch_restricted_tvalues2 <- garch_restricted2@fit$matcoef[,3] %>%
    as.matrix() %>%
    t() # t-stats

  # add results to the correct matrix
  garch_restricted_estimates2 <- rbind(garch_restricted_estimates2,
                                       temp_garch_restricted_estimates2) # estimates

  garch_restricted_pvalues2 <- rbind(garch_restricted_pvalues2,
                                     temp_garch_restricted_pvalues2) # p values

  garch_restricted_tvalues2 <- rbind(garch_restricted_tvalues2,
                                     temp_garch_restricted_tvalues2) # t-stats

}

# name the variables
```

```
garch_restricted_names2 <- c("MU","AR1","Alpha0","Alpha1", "Beta1", "Turnover")

# save matrices as data frames
df_garch_restricted_estimates2 <- as.data.frame(garch_restricted_estimates2)
df_garch_restricted_pvalues2 <- as.data.frame(garch_restricted_pvalues2)
df_garch_restricted_tvalues2 <- as.data.frame(garch_restricted_tvalues2)

# give the dfs the correct names
names(df_garch_restricted_estimates2) <- garch_restricted_names2
names(df_garch_restricted_pvalues2) <- garch_restricted_names2
names(df_garch_restricted_tvalues2) <- garch_restricted_names2

# add variable for persitence (Alpha1 + Beta1)
df_garch_restricted_estimates2 <- df_garch_restricted_estimates2 %>%
  mutate(alpha1_plus_beta1 = (Alpha1 + Beta1))

# print LaTeX-code for estimates from regression 1
stargazer(df_garch_restricted_estimates2[,c("Alpha1", "Beta1", "alpha1_plus_beta1")],
          summary = T,
          summary.stat = c("mean","sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")



# print percentage of significant variables (5% significance level)
apply((df_garch_restricted_pvalues2 < 0.05),
      2,
      function(x){round((sum(x)*100)/length(x),1)})


difference_alpha_beta <-
  (df_garch_restricted_estimates[,6] - df_garch_restricted_estimates2[,7])

difference_alpha_beta[difference_alpha_beta > 0] %>% length/505*100

# print histograms of t-stats from unrestricted GARCH(1,1)
pdf(file = "hist_unrestricted_garch.pdf", width = 8, height = 4)
# set 1x2 window of graphs
par(mfrow = c(1,2))
for(i in 1:2){
  hist(df_garch_restricted_tvalues2[,(i + 3)] %>%
         subset(df_garch_restricted_tvalues2[,(i + 3)] < 100),
       breaks = 50,
       main = NULL,
       xlab = c("alpha 1",
```

```
               "beta 1")[i],
       ylab = NULL,
       xlim = c(0,100),
       #ylim = c(0,120),
       col = "#316ba0",
       border = "#123456")
  abline(v = qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
  abline(v = -qnorm(1.95/2), col = "red", lwd = 4, lty = 2)
}
dev.off()

#reset graph window
par(mfrow = c(1,1))


# remove variables we no longer need
rm(df_garch_restricted_estimates,df_garch_restricted_estimates2,
   df_garch_restricted_pvalues2, garch_restricted2, garch_restricted_estimates2,
   garch_restricted_pvalues2, garch_restricted_names2, spec2, stock,
   temp_garch_restricted_estimates2, temp_garch_restricted_pvalues2,
   garch_restricted_tvalues2, df_garch_restricted_tvalues,
   df_garch_restricted_tvalues2, difference_alpha_beta,
   temp_garch_restricted_tvalues2, i)



######################## Resticted eGARCH ########################

## Restricted EGARCH -- no volume


# create empty matrices for storing estimates later
egarch_restricted_estimates <- matrix(nrow = 0, ncol = 6)
egarch_restricted_pvalues <- matrix(nrow = 0, ncol = 6)
egarch_restricted_tvalues <- matrix(nrow = 0, ncol = 6)

# set specifications for GARCH model. This is an AR(1)-EGARCH(1,1)
# (or, a restricted AR(1)-VA-EGARCH(1,1))
spec <- ugarchspec(variance.model = list(model = "eGARCH",
                                         garchOrder = c(1, 1),
                                         external.regressors = NULL),
                   mean.model = list(armaOrder = c(1, 0)),
                   distribution.model = "norm",
                   start.pars = list(),
                   fixed.pars = list())


# use loop to fit model for all stocks
for (stock in OBI_list) { #loop through all stocks

  # fit model
  egarch_restricted <- ugarchfit(spec=spec,
                                 data = as.matrix(data$Win_return[data$OBI.security.ID
                                                                  == stock]),
```

```
                              solver = "hybrid")

  # extract results we need
  temp_egarch_restricted_estimates <- egarch_restricted@fit$matcoef[,1] %>%
    as.matrix() %>%
    t() # estimates

  temp_egarch_restricted_pvalues <- egarch_restricted@fit$matcoef[,4] %>%
    as.matrix() %>%
    t() # p values

  #temp_egarch_restricted_tvalues <- egarch_restricted@fit$matcoef[,3] %>%
  #  as.matrix() %>%
  #  t() # t-stats

  # add results to the correct matrix
  egarch_restricted_estimates <- rbind(egarch_restricted_estimates,
                                       temp_egarch_restricted_estimates) # estimates

  egarch_restricted_pvalues <- rbind(egarch_restricted_pvalues,
                                     temp_egarch_restricted_pvalues) # p values

  #egarch_restricted_tvalues <- rbind(egarch_restricted_tvalues,
  #                                   temp_egarch_restricted_tvalues) # t-stats

}



# name the variables
egarch_restricted_names <- c("MU","AR1","Alpha0","Alpha1","Beta1","Gamma1")

# save matrices as data frames
df_egarch_restricted_estimates <- as.data.frame(egarch_restricted_estimates)
df_egarch_restricted_pvalues <- as.data.frame(egarch_restricted_pvalues)
#df_egarch_restricted_vtalues <- as.data.frame(egarch_restricted_tvalues)

# give the dfs the correct names
names(df_egarch_restricted_estimates) <- egarch_restricted_names
names(df_egarch_restricted_pvalues) <- egarch_restricted_names
#names(df_egarch_restricted_tvalues) <- egarch_restricted_names


# print LaTeX-code for estimates from regression 1
stargazer(df_egarch_restricted_estimates[,c("Alpha0","Alpha1", "Beta1", "Gamma1")],
          summary = T,
          summary.stat = c("mean","sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
```

```
          header = F,
          float = T,
          float.env = "table")


# print percentage of significant variables (5% significance level)
apply((df_egarch_restricted_pvalues < 0.05),
      2,
      function(x){round((sum(x)*100)/length(x),1)})


# remove variables we no longer need
rm(df_egarch_restricted_pvalues, egarch_restricted, egarch_restricted_estimates,
   egarch_restricted_pvalues, egarch_restricted_names, spec, stock,
   temp_egarch_restricted_estimates, temp_egarch_restricted_pvalues)



######################## Unresticted eGARCH ########################

## Unrestricted EGARCH -- including volume

# create empty matrices for storing estimates later
egarch_unrestricted_estimates2 <- matrix(nrow = 0, ncol = 7)
egarch_unrestricted_pvalues2 <- matrix(nrow = 0, ncol = 7)

# use loop to fit model for all stocks
for (stock in OBI_list) { #loop through all stocks

  # set specifications for GARCH model. This is an AR(1)-VA-GARCH(1,1)
  spec2 <- ugarchspec(variance.model = list(model = "eGARCH",
                                            garchOrder = c(1, 1),
                                            external.regressors = as.matrix(
                                              data$Win_dtrnd_turnover
                                              [data$OBI.security.ID == stock])),
                      mean.model = list(armaOrder = c(1, 0)),
                      distribution.model = "norm",
                      start.pars = list(),
                      fixed.pars = list())

  # fit model
  egarch_unrestricted2 <- ugarchfit(spec=spec2,
                                    data = as.matrix(data$Win_return
                                                     [data$OBI.security.ID == stock]),
                                    solver = "hybrid")

  # extract results we need
  temp_egarch_unrestricted_estimates2 <- egarch_unrestricted2@fit$matcoef[,1] %>%
    as.matrix() %>%
    t() # estimates

  temp_egarch_unrestricted_pvalues2 <- egarch_unrestricted2@fit$matcoef[,4] %>%
    as.matrix() %>%
    t() # p values
```

```
  # add results to the correct matrix
  # estimates
  egarch_unrestricted_estimates2 <- rbind(egarch_unrestricted_estimates2,
                                          temp_egarch_unrestricted_estimates2)

  # p values
  egarch_unrestricted_pvalues2 <- rbind(egarch_unrestricted_pvalues2,
                                        temp_egarch_unrestricted_pvalues2)

}

df_egarch_unrestricted_estimates2 <- as.data.frame(egarch_unrestricted_estimates2)
df_egarch_unrestricted_pvalues2 <- as.data.frame(egarch_unrestricted_pvalues2)

egarch_unrestricted_names2 <- c("MU",
                                "AR1",
                                "Alpha0",
                                "Alpha1",
                                "Beta1",
                                "Gamma1",
                                "Turnover")

names(df_egarch_unrestricted_estimates2) <- egarch_unrestricted_names2
names(df_egarch_unrestricted_pvalues2) <- egarch_unrestricted_names2


# print LaTeX-code for estimates from regression
stargazer(df_egarch_unrestricted_estimates2[,c("Alpha0",
                                               "Alpha1",
                                               "Beta1",
                                               "Gamma1",
                                               "Turnover")],
          summary = T,
          summary.stat = c("mean","sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")


# print percentage of significant variables (5% significance level)
apply((df_egarch_unrestricted_pvalues2 < 0.05),
      2, function(x){round((sum(x)*100)/length(x),1)})

# remove variables we no longer need
rm(df_egarch_unrestricted_pvalues2, egarch_unrestricted2,
   egarch_unrestricted_estimates2, egarch_unrestricted_pvalues2,
```

```
   egarch_unrestricted_names2, spec2, stock, temp_egarch_unrestricted_estimates2,
   temp_egarch_unrestricted_pvalues2)


#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

difference_beta <-
  (df_egarch_restricted_estimates[,5] - df_egarch_unrestricted_estimates2[,5])

difference_beta[difference_beta > 0] %>% length/505*100

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

## Create half life table

temp_col1 <- (df_egarch_restricted_estimates %>%
                transmute(Restricted = (log(0.5)/log(abs(Beta1))) ) )

temp_col2 <- (df_egarch_unrestricted_estimates2 %>%
                transmute(Unestricted = (log(0.5)/log(abs(Beta1))) ) )



df_half_life <- cbind(temp_col1, temp_col2)
df_half_life <- as.data.frame(df_half_life)


# print LaTeX-code for estimates from regression
stargazer(df_half_life,
          summary = T,
          summary.stat = c("mean","sd", "max","p75","median", "p25", "min"),
          flip = T,
          digits = 2,
          digits.extra = 2,
          align = T,
          colnames = T,
          column.sep.width = "0pt",
          initial.zero = T,
          header = F,
          float = T,
          float.env = "table")

rm(temp_col1, temp_col2, df_half_life,
   df_egarch_restricted_estimates, df_egarch_unrestricted_estimates2)


######################## Granger Causality ########################

## Select order
BIC_order_return <- numeric()
BIC_order_vola <- numeric()
BIC_order_volume <- numeric()

for (stock in OBI_list) {
```

```
  temp_BIC_return <- vars::VARselect(y = (data$Win_return
                                          [data$OBI.security.ID == stock]),
                                     lag.max = 40,
                                     type = "const") %>%
    `$`(selection) %>%
    `[`(3)

  temp_BIC_vola <- vars::VARselect(y = (data$Win_return
                                        [data$OBI.security.ID == stock])^2,
                                   lag.max = 40,
                                   type = "const") %>%
    `$`(selection) %>%
    `[`(3)

  temp_BIC_volume <- vars::VARselect(y = (data$Win_dtrnd_turnover
                                          [data$OBI.security.ID == stock]),
                                     lag.max = 40,
                                     type = "const") %>%
    `$`(selection) %>%
    `[`(3)

  BIC_order_return <- append(BIC_order_return, temp_BIC_return)
  BIC_order_vola <- append(BIC_order_vola, temp_BIC_vola)
  BIC_order_volume <- append(BIC_order_volume, temp_BIC_volume)

}

BIC_order_return %>% quantile(c(0.1, 0.25, 0.5, 0.75, 0.9)) # median: BIC 1 (AIC 4)
BIC_order_vola %>% quantile(c(0.1, 0.25, 0.5, 0.75, 0.9))   # median: BIC 5 (AIC 19)
BIC_order_volume %>% quantile(c(0.1, 0.25, 0.5, 0.75, 0.9))   # median: BIC 5 (AIC 13)

rm(BIC_order_return, BIC_order_vola, BIC_order_volume, stock,
   temp_BIC_return, temp_BIC_vola, temp_BIC_volume)

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# does return granger cause turnover?
ret_gc_vol_pval <- numeric()
for (stock in OBI_list) {

  temp_pval <- grangertest(
    data$Win_dtrnd_turnover[data$OBI.security.ID == stock]~
      data$Win_return[data$OBI.security.ID == stock],
    order = 5) %>%
    `$`("Pr(>F)") %>%
    `[`(2)

  ret_gc_vol_pval <- append(ret_gc_vol_pval, temp_pval)

}
print("Return granger causes volume")
ret_gc_vol_pval[ret_gc_vol_pval < 0.05] %>% length()/505*100
```

```
# does volume granger cause return?
vol_gc_ret_pval <- numeric()
for (stock in OBI_list) {

  temp_pval <- grangertest(
    data$Win_return[data$OBI.security.ID == stock]~
      data$Win_dtrnd_turnover[data$OBI.security.ID == stock],
    order = 5) %>%
    `$`("Pr(>F)") %>%
    `[`(2)

  vol_gc_ret_pval <- append(vol_gc_ret_pval, temp_pval)

}
print("Volume granger causes return")
vol_gc_ret_pval[vol_gc_ret_pval < 0.05] %>% length()/505*100




# does volatility granger cause volume?
vola_gc_vol_pval <- numeric()
for (stock in OBI_list) {

  temp_pval <- grangertest(
    data$Win_dtrnd_turnover[data$OBI.security.ID == stock]~I(
      (data$Win_return[data$OBI.security.ID == stock])^2),
    order = 5) %>%
    `$`("Pr(>F)") %>%
    `[`(2)

  vola_gc_vol_pval <- append(vola_gc_vol_pval, temp_pval)

}
print("Volatility granger causes volume")
vola_gc_vol_pval[vola_gc_vol_pval < 0.05] %>% length()/505*100



# does volume granger cause volatility?
vol_gc_vola_pval <- numeric()
for (stock in OBI_list) {

  temp_pval <- grangertest(
    I(
      (data$Win_return[data$OBI.security.ID == stock])^2)~
      data$Win_dtrnd_turnover[data$OBI.security.ID == stock],
    order = 5) %>%
    `$`("Pr(>F)") %>%
    `[`(2)
```

```
  vol_gc_vola_pval <- append(vol_gc_vola_pval, temp_pval)

}
print("Volume granger causes volatility")
vol_gc_vola_pval[vol_gc_vola_pval < 0.05] %>% length()/505*100


rm(ret_gc_vol_pval, temp_pval, vol_gc_ret_pval,
   vol_gc_vola_pval, vola_gc_vol_pval, stock)
```

Appendix G Preliminary thesis

All following pages are from our preliminary master thesis, delivered the 28th of February to our supervisor. BI Norwegian Business School require us to include this document in the final thesis.

Preliminary Master Thesis
BI Norwegian Business School

What is the empirical relationship between volume and stock returns on Oslo Stock Exchange?

Supervisor:
Costas Xiouros

Examination Code and Name:
GRA 19502 – Master Thesis

Programme:
Master of Science in Business – Major in Finance

Jan Petter Iversen and Astri Skjesol

February 28, 2018

Contents

List of Figures

List of Tables

Abbreviations

ADF	**A**ugmented **D**ickey-**F**uller
AIC	**A**kaike **I**nformation **C**riterion
ARCH	**A**uto**r**egressive **C**onditional **H**eteroskedicity
ADEX	**A**thens **De**rivatives **E**xchange
ASE	**A**thens **S**tock **E**xchange
BIC	**B**ayesian **I**nformation **C**riterion
BOVESPA	**BO**lsa de **V**alores do **E**stado **S**ão **Pa**ulo
CAPM	**C**apital **A**sset **P**ricing **M**odel
EGARCH	**E**xponential **GARCH**
GARCH	**G**eneral **ARCH**
GJR-GARCH	**G**losten-**J**agannathan-**R**unkle **GARCH**
GMM	**G**eneralised **M**hetod of **M**oments
HF	**H**igh **F**requency
HFT	**H**igh **F**requency **T**rading
IPSA	**Í**ndice de **P**recio **S**electivo de **A**cciones
KOSPI	**Ko**rean composite **S**tock **P**rice **I**ndex
MDH	**M**ixed **D**istribution **H**ypothesis
NYSE	**N**ew **Y**ork **S**tock **E**xchange
OSE	**O**slo **S**tock **E**xchange
OLS	**O**rdinary **L**east **S**quares
QGARCH	**Q**uadratic **GARCH**
SARV	**S**tochastic **A**uto**r**egressive **V**olatility
SIAH	**S**equential **I**nformation **A**rrival **H**ypothesis
SSE	**S**hanghai **S**tock **E**xchange
SZSE	**S**hen**z**hen **S**tock **E**xchange
TARCH	**T**reshold **ARCH**
VAR	**V**ector **A**uto**r**egressive
WBAG	**W**iener **B**örse **AG**

1 Introduction

The relationship between return, volatility and trading volume has been a central part of financial research since the late 50s for numerous reasons. First, it is important for the understanding of the microstructure of financial markets, which O'Hara states has become even more important over the last few years (2015). Volume has long been linked to the flow of information, and the role of information in setting security prices is one of the most fundamental research issues in finance (see e.g. Brailsford, 1996, p. 90). Second, because knowledge about this relation might improve short term forecasting of returns, volume or volatility, and might have implications for futures markets. Third, because it might help improve or create liquidity-adjusted risk- or expected shortfall metrics (e.g. Anthonisz & Putniņš, 2016, p. 31). Fourth, because it is often applied in technical analysis as a measure of the strength of stock price movements (e.g. Gallo & Pacini, 2000, p. 167; Abbondante, 2010, p. 287). And last, it has implications for theoretical and empirical asset pricing, established through its effect on liquidity (see Amihud & Mendelson, 1986; Chordia, Subrahmanyam, & Anshuman, 2001). High liquidity reduces the required return by investors and thus reduces the cost of capital for the issuers of securities. An efficient price discovery processes, associated with lower volatility, make market prices more informative and enhance the role of the market in aggregating and conveying information through price signals (Amihud, Mendelson, & Murgia, 1990, p. 439). It also has a socio-economic benefit, as it leads to a more efficient capital allocation.

The relationship between return, volatility and volume has been studied extensively. However, to our knowledge, there has not been conducted any recent studies of this on Oslo Stock Exchange. Additionally, a lot of the literature might be somewhat outdated, given the recent developments in market conditions. Algorithmic trading, and especially high frequency trading (HFT), has changed markets in fundamental ways, and the high speeds gives market microstructure a starring role (O'Hara, 2015, p. 257). The stock markets have gradually transitioned from a time when trading occurred between humans, to a mixed phase of humans and machines to an ultrafast mostly-machine phase where machines dictate price changes (Johnson et al., 2012, p. 5). O'Hara believes that HFT has altered some basic constructs underlying microstructure research (2015, p. 263). She states that research must change to reflect the new realities, and that understanding how markets and trading have changed is important for informing future research (2015, pp. 257–258).

O'Hara argues that with the radically different markets and way of trading there is no lack of things that are not yet understood; both particular, general and conceptual questions demand immediate attention (2015, pp. 263–268). Our aim is to add to the current literature on the volume-return relationship by studying the Norwegian stock exchange. This motivates the following research question: "What is the empirical relationship between volume and stock returns on Oslo Stock Exchange?"

This preliminary thesis is organized in the following way. First we will do a short introduction of Oslo Stock Exchange. Thereafter we will explain the most relevant theories encountered in this thesis. The next section surveys the current literature. Then we will explain what sort of methodology we will use to analyse our data, and what data we want to collect. The last section includes our progression plan.

2 Oslo Stock Exchange

Kristiania Børs – the precursor to what is today Oslo Stock Exchange – was approved by King Carl Johan in 1818. This was Norway's second exchange (Hodne & Grytten, 1992, pp. 53–54) when it opened its doors for the first time in April 1819 – soon two centuries ago (e.g. Mjølhus, 2010, p. 28). At that time, Norway was mainly a country of farmers and fishermen, and the capital had less than 10,000 inhabitants (e.g. Kristiania børs, 1919, p. 1). According to Oslo Stock Exchange's webpage, the exchange originally functioned as an auction house for goods, ships and ship parts, and as an exchange for foreign currencies. Back then, the currency prices were updated twice a week.

The Oslo Stock Exchange introduced stocks in 1881. Although the trade was modest at first, the number of securities exploded between 1891 and 1900, from 40 to 165 (Hodne & Grytten, 2000, p. 170). Daily quotes were introduced in 1916 for some stocks – but not until 1922 for all stocks.

Today, Oslo Stock Exchange list the shares of 187 companies, with a combined market capitalization of almost 329 billion USD. The exchange is a private limited company, which it has been since 2001. Not much has been written about the return-volume relationship on Oslo Stock Exchange. Næs, Skjeltorp, and Ødegaard (2008) examined the relationship between the long-term development in liquidity at the exchange and the Norwegian Economy, and Jørgensen, Skjeltorp, and Ødegaard (2017) wrote about the order-to-trade ratio. Mikalsen (2014) shows several examples of volume analysis in technical trading on Oslo Stock Exchange – at least indicating that volume is an important metric for Norwegian traders as well. Karolyi, Lee, and Van Dijk examined co-movement between trading activity and return in several countries and found that for Norway, commonality was 25.36% in returns, 23.31% in liquidity, and 23.82% in turnover (2009). The sparse literature about the exchange is part of the motivation as to why we want to write about Oslo Stock Exchange and not any other exchange.

3 Theory

The efficient market hypothesis (EMH) emphasize the role of information in setting prices, and defines an efficient market as one in which new information is incorporated quickly and correctly into its current security prices (Lim & Brooks, 2011, p. 69). Therefore, most models trying to explain the return-volume relationship are clearly related to the flow of new information, and the process that incorporates this information into market prices (e.g. Andersen, 1996, p. 170; Brailsford, 1996, p. 95).

The two main hypothesis underlying these models are the sequential information arrival hypothesis (SIAH) and the mixture of distributions hypothesis (MDH). SIAH was first developed by Copeland (1976, 1977) and later expanded by Jennings, Starks, and Fellingham (1981). The hypothesis assumes that investors receive information sequentially at different times, which shift the optimists' demand curve up, and the pessimists' demand curve down. Trading occur as a reaction to this new information. Buy trades are viewed as noisy signals of good news, sell trades as noisy signals of bad news (O'Hara, 2015, p. 263). MDH assumes that daily price changes are sampled from a set of distributions with different variances. In the MDH-model specified by Epps and Epps (1976), investors revise their reservation price when new information enter the market. Volume is viewed as the disagreement between the investors (see e.g. B.-S. Lee & Rui, 2002, p. 54). Andersen note that there is evidence both in support of and against the MDH (1996, p. 170).

The arrival of new information causes investors to revise their price reservations. As investors are heterogeneous in their interpretation of news, prices may not change even though new information enters the market. This might happen if some investors interpret the news as good and others as bad (e.g. Mestel, Gurgul, & Majdosz, 2003, p. 3; de Medeiros & Van Doornik, 2006, p. 2). Volume is always non-negative and as long as at least one investor makes an adjustment in their price revision, expected trading volume is positive (Brailsford, 1996, pp. 93–94). Therefore, volume can be seen as an indicator of consensus, or the lack thereof (Gallo & Pacini, 2000, p. 167). Average investor-reaction to information is reflected in price movements (e.g. Mestel et al., 2003, p. 3; de Medeiros & Van Doornik, 2006, p. 2). However, information arrival is not constant, and displays seasonalities and distinct intraday patterns (e.g. Berry & Howe, 1994).

Learning is an important feature in many microstructure models. Most such models rely on the notion that some traders have private information which they trade on. Other traders see market data and they learn from it. Market prices adjust to efficient levels that reflect all the information (O'Hara, 2015, p. 263).

4 Literature Review

There is an old Wall Street adage stating that "It takes volume to make prices move." Studies of the price-volume relation dates back to the late 1950s (see e.g. Chandrapala, 2011) when Osborne laid the theoretical foundation (1959). One of the earliest empirical studies was performed by Granger and Morgenstern (1963), who found the connection between volume and stock prices on the New York Stock Exchange to be negligible. Ying (1966) was the first to document a positive correlation between volume and price change ($V,\ \Delta p$), and a positive correlation between the volume and absolute price change ($V,\ \mid \Delta p \mid$). In his extensive literature review, Karpoff (1987) state that numerous empirical findings in the 60s, 70s and 80s support the positive volume-absolute price change correlation. Further, Karpoff describes several similar findings for the relationship between volume and price change variance, price change magnitude, price variability, absolute price change, squared abnormal return and squared price change. However, most of these effects have little economic impact (Karpoff, 1987).

Karpoff (1987) summarize the research conducted before 1987 with the following conclusions:

1. No volume-price correlation exists
2. A correlation exists between volume and absolute price change ($V,\ \mid \Delta p \mid$)
3. A correlation exists between volume and price change ($V,\ \Delta p$)
4. Volume is higher when prices increase than when prices decrease

He further suggests that it is likely that the relationship between volume and price changes stems from their common ties to information flows or their common ties to a directing process that can be interpreted as the flow of information (Karpoff, 1987).

In Table 1 we have summarized the data used, methodology and results of several other papers on the volume-return relationship.

Author	**Year**	**Data**	**Model**	**Conclusion**
Transaction data test of the mixture of distribution hypothesis				
Harris	(1987)	NYSE: D		Trading might be self generating.
Heteroscedasticity in stock Return Data: Volume versus GARCH effects				
Lamoureux & Lastrapes	(1990)	U.S.	GARCH	GARCH effects vanish (due to volume).
Stock Prices and Volume.				

Author	Year	Data	Model	Conclusion
Gallant et al.	(1992)	NYSE: D	VAR, ARCH	Contemporaneous volume-volatility correlation. Large price movements associated with higher subsequent volume. Volume-leverage interaction. Positive conditional risk-return relation after conditioning on lagged volume.
Return Volatility and Trading Volume: An information Flow Interpretation of Stochastic Volatility				
Andersen	(1996)	IBM share	GMM, GARCH	Consistent with the MDH.
The effects of trading activity on market volatility				
Gallo & Pacini	(2000)	U.S.	GARCH, EGARCH	Structure of GARCH-type models of conditional heteroskedasticity does not manage to capture the quick absorption of large shocks to returns and implies in practice a too high level of persistence of shocks.
Does Trading Volume Contain Information to Predict Stock Returns? China's Stock Markets				
C. F. Lee & Rui	(2000)	SSE, SZSE: D	GARCH, VAR	Trading volume does not ganger cause stock return on individual markets. US and Hong Kong financial market information contained in returns, volatility and volume has very weak predictive power for Chinese financial market variables.
The Dynamic Relation between Stock Returns, Trading Volume, and Volatility				
Chen et al.	(2001)	U.S., Asia, Europe: D	EGARCH, VAR	GARCH effects remains significant when contemporaneous and lagged volume is included in the model.
The Dynamic Relationship between stock returns and Trading Volume				

Author	**Year**	**Data**	**Model**	**Conclusion**
B. -S. Lee & Rui	(2002)	NY, Tokyo, London: D	GMM, GARCH, VAR	Trading volume does not Granger-cause stock market returns on each of the markets. However, there exists a positive feedback relationship between trading volume and return volatility in all three markets.
The empirical relationship between stock returns, return volatility and trading volume: Austrian market				
Mestel et al.	(2003)	WBAG	GARCH, VAR	The relationship between stock return and trading volume is mostly negligible. Evidence of a relationship (contemporaneous & causal) between return volatility and trading volume.
Trading Volume and Returns Relationship in Greek Stock Index Futures Market				
Floros & Vougas	(2007)	ASE, ADEX	GARCH, GMM	Findings indicate that market participants use volume as an indication of prices.
The Price-Volume Relationship in the Chilean Stock Market				
Kamath	(2008)	IPSA: D		Granger causality running from returns to volume.
The empirical relationship between stock return, return volatility and trading volume: Brazil				
de Mendeiros & Van Doornik	(2006)	BOVESPA: D	GARCH, VAR	Significant contemporaneous relationship between return volatility and trading volume. Stock return depends on trading volume, not the other way around. Higher trading volume and return volatility relationship is asymmetrical. GARCH effect and high hysteresis in conditional volatility. Granger causality between trading volume and return volatility is strongly evident in both directions.
The Dynamic Relationship between Price and Trading Volume: Indian Stock Market				

Author	Year	Data	Model	Conclusion
Kumar et al.	(2009)	S&P CNX Nifty Index	GARCH, VAR	ARCH effects decline when trading volume is included in GARCH equation.
Asymmetric Volatility and Trading Volume: The G5 Evidence				
Sabbaghi	(2011)	G5 stock markets: D	EGARCH	The findings in this paper support prior research that has documented a positive association between trading volume and return volatility. Persistence levels do not decrease with the inclusion of trading volume in the EGARCH.
Relationship between Trading Volume and Asymmetric Volatility in the Korean Stock Market				
Choi et al.	(2012)	KOSPI	EGARCH, GJR-GARCH	Trading volume is a useful tool for predicting the volatility dynamics of the Korean stock market.

Table 1: Literature overview

In addition to the volume-return relationship, much literature is dedicated to the study of liquidity. Volume and liquidity is inextricably linked (e.g Benston & Hagerman, 1974; Stoll, 1978). A market is said to be liquid if traders can quickly buy or sell a large number of shares at low transaction costs with little price impact (Næs et al., 2008, p. 2). In other words, liquidity includes a cost dimension, a quantity dimension, a time dimension and an elasticity dimension. A natural measure of the cost dimension is the bid-ask spread, which indeed has been found to be negatively correlated with other liquidity characteristics such as volume, number of shareholders, number of market makers trading the stock and stock price continuity (Amihud & Mendelson, 1986, pp. 223–224).

The level of liquidity affects expected returns because investors know that in relatively less liquid stocks, transaction costs will erode more of the realized return (see e.g. Amihud & Mendelson, 1986; Anthonisz & Putniņš, 2016). Thus, investors demand a premium for less liquid stocks, and so expected returns should be negatively correlated with the level of liquidity (e.g. Chordia et al., 2001, pp. 29–30). Amihud and Mendelson shows that excess returns are increasing in both β and the spread (1986, p. 238), indicating that part of the effect traditionally attributed to the CAPM β may in fact be due to the spread.

Similar to the return-volume relationship, liquidity behaves and is priced asymmetrically (e.g. Anthonisz & Putniņš, 2016, p. 3). By assuming symmetry, as is implicit in much of the existing theoretical and empirical literature, the importance of liquidity risk in explaining cross-sectional returns might be underestimated. Anthonisz and Putniņš finds that stocks with high downside liquidity risk compensate investors with an substantial expected return premium (2016, p. 3). This is consistent with investors disliking stocks that are more susceptible to liquidity spirals or abandonment during flights to liquidity. Chordia, Roll, and Subrahmanyam (2002) have found that buying activity is more pronounced following market crashes and selling activity is more pronounced following market rises, while Karolyi et al. suggests that common variation in individual stocks tend to rise during financial crises (2009, p. 21). Anthonisz and Putniņš finds that there is a greater dispersion in downside liquidity risk during illiquid market states than liquid states (2016, p. 26). Wang, Wu, and Lai developed a model which allow for the return-volume dependence to switch between positive and negative dependence regimes (2018). They are the first to divide their observations into four different market conditions: rising return/rising volumes, falling returns/falling volumes, rising returns/falling volumes, and falling returns/rising volumes. They find that the volatilities of return and volume are larger for the negative dependence regime than for the positive dependence regimes. They find support for heterogeneous investors with short-sale constraints. The return-volume dependence is asymmetric. Both the intensity of information and liquidity trading are important in driving the time-varying, return-volume dependence (Wang et al., 2018).

If the investors adapt their strategies on a slower time scale than the time scale on which the trading process takes place, this will lead to positive autocorrelation in volatility and volume (Brock & LeBaron, 1995). Chordia et al. finds that liquidity is highly predictable not only by its own past values but also by past market returns (2002). The number of trades and the market return can predict future changes in liquidity. However, controlling for the market return, the predictive power of volatility is only marginal.

Several studies suggest that market microstructure directly influences the liquidity or available supply of a tradable asset which in turn impacts the pricing, valuation and risk measurement of the asset (e.g. Abrol, Chesir, & Mehta, 2016, p. 116). Amihud and Mendelson suggest that liquidity increasing financial policies can reduce the firm's opportunity cost of capital and provide measures for the value of improvements in the trading and exchange process (1986, p. 224). Thus, market-microsturcture factors can be important as determinants of stock returns. Further, their results suggest a strong incentive for the firm to invest in increasing the liquidity of the claims it issues; like going public, standardize contracts or enlist on exchanges (Amihud & Mendelson, 1986, p. 246). Anthonisz and Putniņš finds that firms can also reduce their cost of capital by minimizing their stocks' downside liquidity risk (Anthonisz & Putniņš, 2016, p. 31).

Karolyi et al. find that commonality in returns, liquidity, and turnover is greater in countries that are less economically and financially developed, have weaker investor protection, and are characterized by a less transparent information environment, a smaller equity mutual fund base, and a greater fraction of closely held shares (2009, p. 18). This is consistent with Bhattacharya and Galpin (2011) and Wu (2017). Bhattacharya and Galpin finds that value-weighted portfolios are more popular in developed markets than emerging markets. They speculate that this is because stock picking is less popular when the public information disclosure environment is good. Stock pickers can only make money when they have better information than everyone else, which they will not have when public information disclosure is good (2011, pp. 739–740). There seems to be a general, but not universal, consensus that increased transparency result in better liquidity and reduced transaction costs (e.g. Næs et al., 2008, p. 7). One conflicting opinion is Madhavan who shows that transparency can also reduce liquidity, as transparrent markets might lose out on informed traders who do not want to reveal their trading interests (1995, pp. 593–594).

One of our motivations for this thesis is the changed trading environment. By all accounts, high frequency trading has become very significant in today's markets (Friederich & Payne, 2015). According to O'Hara, the rise of HFT has also radically changed how non high frequency (HF) traders behave, and the markets where this trading occurs. The current market structure is highly competitive, highly fragmented, and very fast (O'Hara, 2015, p. 258). The estimated amount of high frequency trading differs. Brogaard, Hendershott, and Riordan (2014) found that HTF makes up over 42% of traded volume on Nasdaq, while Hagströmer and Norden (2013) estimate that 26-30% of firms trading on Nasdaq-OMX to be pure HF firms, and a total amount of HF trading could be as high as 50%. O'Hara also state that by some estimates, high frequency traders make up half or more of all trading volume (2015, p. 258). There is a general, but not universal, agreement that HFT market making enhances market quality by reducing spreads and enhancing informational efficiency (O'Hara, 2015, p. 259). The bid-ask spread narrows, leading to a more efficient price discovery process, and increased trading volumes has increased market liquidity (see e.g. Hendershott, Jones, & Menkveld, 2011; Abrol et al., 2016). However, many are concerned that HFT induce market instability. When looking at multiple exchanges between 2006 and 2011, Johnson et al. finds on average more than one flash-crash each trading day (2012) and O'Hara points out that HFT might lead to periodic illiquidity (2015, p. 259). Additionally, some HFT strategies are considered predatory. According to Friederich and Payne there is no estimate of how much HF flow might be abusive in nature, because such behaviour is very difficult to detect (2015, p. 4). They further state that there is a suspicion that regulators are overwhelmed by the amount of data that today's markets generate, and that they are lagging behind brokers and exchanges in respect of the skills needed to analyse this data.

The ability of high frequency traders to enter and cancel orders faster than others, makes it hard to discern where liquidity exists in the markets (O'Hara, 2015, p. 258). Abrol et al. finds that the high speeds enables sub second injections and withdrawals of liquidity (2016, p. 126), which is faster than humans can notice and physically react to (Johnson et al., 2012, p. 2). Orders are sent to and from the exchange as part of complex dynamic trading strategies, and it is now common for upward of 98% of all orders to be canceled instead of of being executed as trades (O'Hara, 2015, p. 259). From a computer perspective, HF trading algorithms in the sub-second regime need to be executable extremely quickly and hence be relatively simple, without calling on much memory concerning past information (Johnson et al., 2012, p. 6). There is therefore a question of how much information such trades incorporate. O'Hara argues that with algorithmic trading, trades are no longer the basic unit of information – the underlying orders are (2015, p. 263).

An implication of HFT for regulators is that extreme behaviors on long and very short time scales – such as crashes and flash-crashes – cannot be separated *a priori*. Rules targeted solely at controlling one or the other can induce dangerous feedback effects at the opposite timescale (Johnson et al., 2012, p. 4). Additionally, some regulations targeted at HFT might damage liquidity due to the fact that HF traders may be acting as *de facto* market makers (see Friederich & Payne, 2015, p. 5).

5 Methodology

Before we start on our analysis, we must decide for a measure of trading activity. A main challenge in empirical research on liquidity has been to construct measures that can capture all dimensions of liquidity in a satisfactory way (Næs et al., 2008, p. 2). There is no theoretically or universally accepted measure to determine a market's liquidity, and thus a number of measures must be considered (e.g. Lybek & Sarr, 2002). Also, a range of market-specific factors and peculiarities must be taken into consideration. A much applied measure of trading activity is *turnover* – the number of shares traded over the number of shares outstanding – sometimes referred to as *relative volume* (Campbell, Grossman, & Wang, 1993; Lo & Wang, 2000).

$$turnover = \frac{number\ of\ sharestraded}{number\ of\ shares\ outstanding}$$

When we have decided on such a measure and collected the data we need, we are ready to start with the main part of our thesis. We plan to start with data cleaning and preparation. When our dataset is ready, we start our exploratory data analysis. We plan to test the time series for stationarity using an augmented Dickey-Fuller (ADF) test. If we are working with stock prices, they will most likely display strong traits of non-stationary, and will need to be transformed. There are several ways to achieve stationarity. Some series require detrending, and are called trend-stationary. Others will need to be differenced and are called difference-stationary. Stock prices tend to be difference-stationary, and thus we can log-transform them and first difference them. The daily log-difference series will be a very close proxy of daily returns, and will hopefully be stationary. Several articles state that volume display traits of trend and seasonality. Thus we deem it likely that we need to detrend our volume-data, maybe even using non-linear filters.

Next, we want to move to descriptive statistics. Financial time series tend to display non-normal tendencies, which we would like to test using a Jarque-Bera test for normality. This is important to know about, in case some of our tests or models require normality.

Next step in exploring the relationship between stock return and volume would be to do a cross-correlation analysis to look at the contemporaneous as well as a dynamic (causal) relationship, using a Vector Autoregressive (VAR) model. As VAR models can be sensitive to non-stationarity, this is somewhat dependent on the results we find earlier in our analysis. The dynamic relationship between stock return and trading volume can help in better understanding the microstructure of Oslo Stock Exchange.

Continuing, we plan to develop a multivariate model. It is widely documented that daily financial return series display strong conditional heteroskedasticity. The standard warning is that in the presence of heteroskedasticity, the regression coefficients for an ordinary least squares (OLS) regression are still unbiased, but the standard errors and confidence intervals estimated by conventional procedures will be too narrow, giving a false sense of precision. Therefore, the ARCH model, and its extension into GARCH, is often used – with good results (e.g. Andersen, 1996). Instead of considering this as a problem to be corrected, ARCH and GARCH models treat heteroskedasticity as a variance to be modeled. The GARCH model, like the ARCH model, have a weighted average of past squared residuals, but includes declining weights that never reaches zero (Engle, 2001). The EGARCH and TARCH models where later developed as more evidence indicated that the direction of returns affect volatility (Engle, 2001, p. 166). It is adjacent to expect that these models will be a good fit for us, however there are other possibilities too. Several studies suggest that the return-volume relationship is asymmetric. Other extensions of the ARCH, such as the EGARCH or the QGARCH, takes this asymmetry into consideration. Other models we have seen used is GJR-GARCH, SARV, and GMM models. We would need to study these models closer before deciding on one. Whichever model we use, we plan to use an information criterion such as the AIC or the BIC to decide the order of the model.

Further, we probably need to use control variables in our model. Some earlier studies suggest including day-of-the-week dummies to control for weekly asymmetries. Others suggest including quadratic terms or interaction terms to control for non-linear relationship between the variables. Several studies control for external influences, such as macroeconomic variables, or well-known pricing factors such as the factors in the Fama French model or momentum.

When our model is developed, we plan to look for dynamic relationships. We will test our sample for Granger causality, both from volume to returns and returns to volume.

If the time and scope of our thesis allow for it, it would be interesting to perform the same tests with different frequency data, and look for differences in results. It would also be interesting to follow Wang et al. and look at asymmetry in positive and negative dependence regimes (2018).

When analyzing our data, we will use the open source statistical software R (R Core Team, 2017). As it is not always natural to cite the extensions we have used in our analysis in the running text, we will reference them all here. In this version of the thesis, we have used the following package:

- plan (Kelley, 2013)

6 Data

We need data on stock prices, volume and preferably adjusted returns. All of these data are downloadable from Yahoo Finance going back to the year 2000. We can also find market returns from OSE on Bernt Arne Ødegards homepage, going all the way back to the 1980s. His webpage also includes the risk free rate and the Fama French factors: SMB, HML and UMB, the Charhart Momentum factor and a Liquidity factor for the Norwegian market, which we might need as control variables. We might also need macroeconomic variables as control variables, which we can download from Thomson Reuters Datastream, Statistics Norway or from the webpages of the Central Bank of Norway.

When it comes to frequency, the choice seems somewhat arbitrary. Our initial thought were to look for high frequency data, such as five minute intervals, so that we could pick up on HFT. However, in a high frequency setting, five minutes is a “lifetime”, and is not a meaningful time frame to evaluate trading (O’Hara, 2015, p. 267). Higher than daily frequency poses problems of inter-asset synchronicity which could make it difficult to detect market-wide relations (Chordia et al., 2002) and could result in noisy estimates (e.g. Corwin & Coughenour, 2008, p. 3038). Further, high frequency data would result in a very large data set, and would lead to a lot of work in data cleaning and preparation. Also, since O’Hara believes that issues when analyzing HFT data cannot be solved by better data sets (2015, p. 268), this does not seem like the way to go.

Chordia et al. states that the relationship between liquidity and returns is most likely to manifest itself over short horizons, that is daily as opposed to weekly or monthly, and picked a daily frequency because of this (2002). Wu (2017) chose a weekly horizon as a compromise between maximizing sample size and minimizing day to day volume and return fluctuations that have less direct economic relevance.

Our plan is to first collect daily data, and – if time and scope allow it – also collect and analyse weekly or monthly data. It could be interesting to see if the different frequencies yield different results.

7 Progression Plan

Going forward, we plan to continue our literature review. Especially, we want to include more behavioral theories, and read more about the Adaptive Market Hypothesis (AMH). After we have collected our data, we want to do a deep dive into the financial statistics and economietrics literature, to really get to the bottom of the pros and cons of the different methods, models and tests. In Figure 1 we outline our expected progress going forward.

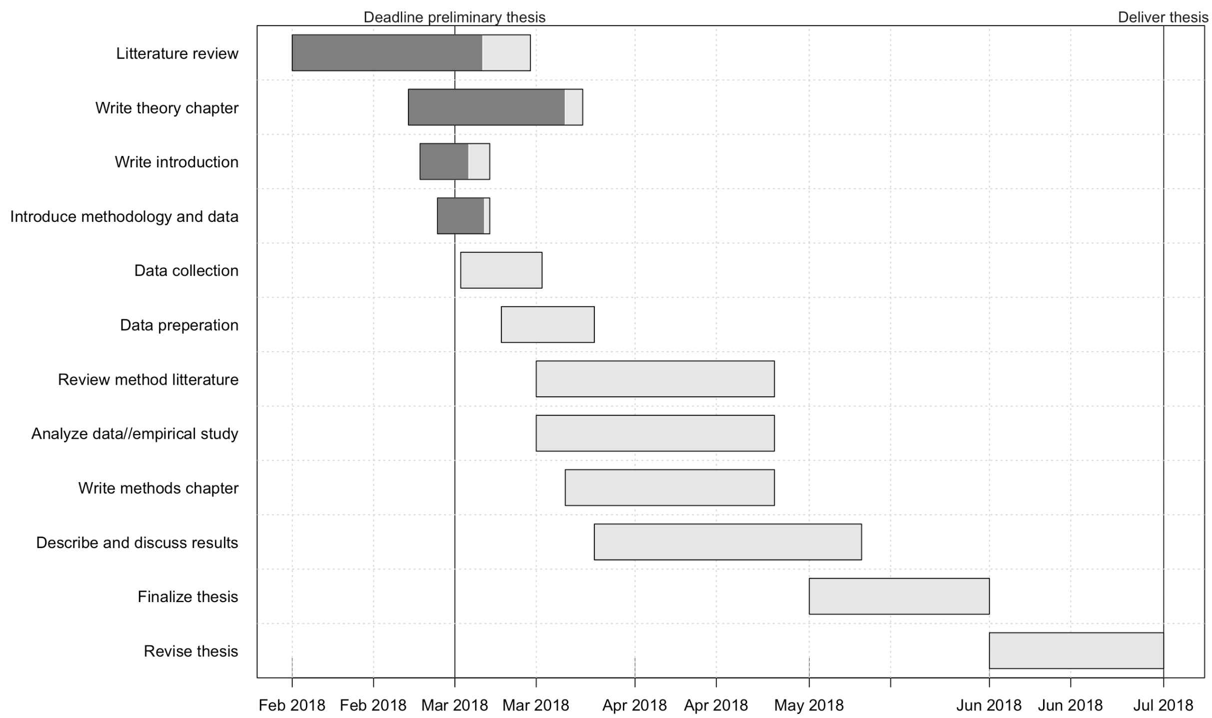

Figure 1: Progression plan

References

Abbondante, P. (2010). Trading volume and stock indices: A test of technical analysis. *American Journal of Economics and Business Administration*, *2*(3), 287–292.

Abrol, S., Chesir, B., & Mehta, N. (2016). High frequency trading and us stock market microstructure: A study of interactions between complexities, risks and strategies residing in us equity market microstructure. *Financial Markets, Institutions & Instruments*, *25*(2), 107–165.

Amihud, Y., & Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of financial Economics*, *17*(2), 223–249.

Amihud, Y., Mendelson, H., & Murgia, M. (1990). Stock market microstructure and return volatility: Evidence from italy. *Journal of Banking & Finance*, *14*(2-3), 423–440.

Andersen, T. G. (1996). Return volatility and trading volume: An information flow interpretation of stochastic volatility. *The Journal of Finance*, *51*(1), 169–204.

Anthonisz, S. A., & Putniņš, T. J. (2016). Asset pricing with downside liquidity risks. *Management Science*, *63*(8), 2549–2572.

Benston, G. J., & Hagerman, R. L. (1974). Determinants of bid-asked spreads in the over-the-counter market. *Journal of Financial Economics*, *1*(4), 353–364.

Berry, T. D., & Howe, K. M. (1994). Public information arrival. *The Journal of Finance*, *49*(4), 1331–1346.

Bhattacharya, U., & Galpin, N. (2011). The global rise of the value-weighted portfolio. *Journal of Financial and Quantitative Analysis*, *46*(3), 737–756.

Brailsford, T. J. (1996). The empirical relationship between trading volume, returns and volatility. *Accounting & Finance*, *36*(1), 89–111.

Brock, W. A., & LeBaron, B. D. (1995). *A dynamic structural model for stock return volatility and trading volume* (Tech. Rep.). National Bureau of Economic Research.

Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies*, *27*(8), 2267–2306.

Campbell, J. Y., Grossman, S. J., & Wang, J. (1993). Trading volume and serial correlation in stock returns. *The Quarterly Journal of Economics*, *108*(4), 905–939.

Chandrapala, P. (2011). The relationship between trading volume and stock returns. *Journal of Competitiveness*, *3*(3).

Chen, G.-m., Firth, M., & Rui, O. M. (2001). The dynamic relation between stock returns, trading volume, and volatility. *Financial Review*, *36*(3), 153–174.

Choi, K.-H., Jiang, Z.-H., Kang, S. H., & Yoon, S.-M. (2012). Relationship between trading volume and asymmetric volatility in the korean stock market. *Modern Economy*, *3*(05), 584.

Chordia, T., Roll, R., & Subrahmanyam, A. (2002). Order imbalance, liquidity, and market returns. *Journal of Financial economics*, *65*(1), 111–130.

Chordia, T., Subrahmanyam, A., & Anshuman, V. R. (2001). Trading activity and expected stock returns. *Journal of Financial Economics*, *59*(1), 3–32.

Copeland, T. E. (1976). A model of asset trading under the assumption of sequential information arrival. *The Journal of Finance*, *31*(4), 1149–1168.

Copeland, T. E. (1977). A probability model of asset trading. *Journal of Financial and Quantitative Analysis*, *12*(4), 563–578.

Corwin, S. A., & Coughenour, J. F. (2008). Limited attention and the allocation of effort in securities trading. *The Journal of Finance*, *63*(6), 3031–3067.

de Medeiros, O. R., & Van Doornik, B. F. (2006). The empirical relationship between stock returns, return volatility and trading volume in the brazilian stock market.

Engle, R. (2001). Garch 101: The use of arch/garch models in applied econometrics. *Journal of economic perspectives*, *15*(4), 157–168.

Epps, T. W., & Epps, M. L. (1976). The stochastic dependence of security price changes and transaction volumes: Implications for the mixture-of-distributions hypothesis. *Econometrica: Journal of the Econometric Society*, 305–321.

Floros, C., & Vougas, D. (2007). Trading volume and returns relationship in greek stock index futures market: Garch vs. gmm. *International Research Journal of Finance and Economics*(12), 98–115.

Friederich, S., & Payne, R. (2015). Order-to-trade ratios and market liquidity. *Journal of Banking & Finance*, *50*, 214–223.

Gallant, A. R., Rossi, P. E., & Tauchen, G. (1992). Stock prices and volume. *The Review of Financial Studies*, *5*(2), 199–242.

Gallo, G. M., & Pacini, B. (2000). The effects of trading activity on market volatility. *The European Journal of Finance*, *6*(2), 163–175.

Granger, C. W., & Morgenstern, O. (1963). Spectral analysis of new york stock market prices. *Kyklos*, *16*(1), 1–27.

Hagströmer, B., & Norden, L. (2013). The diversity of high-frequency traders. *Journal of Financial Markets*, *16*(4), 741–770.

Harris, L. (1987). Transaction data tests of the mixture of distributions hypothesis. *Journal of Financial and Quantitative Analysis*, *22*(2), 127–141.

Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, *66*(1), 1–33.

Hodne, F., & Grytten, O. H. (1992). *Norsk økonomi 1900-1990.* Tano.

Hodne, F., & Grytten, O. H. (2000). Norsk økonomi i det 19. århundre. *Oslo: Fagbokforlaget*.

Jennings, R. H., Starks, L. T., & Fellingham, J. C. (1981). An equilibrium model of asset trading with sequential information arrival. *The Journal of Finance*, *36*(1), 143–161.

Johnson, N., Zhao, G., Hunsader, E., Meng, J., Ravindar, A., Carran, S., & Tivnan, B. (2012). Financial black swans driven by ultrafast machine ecology. *arXiv preprint arXiv:1202.1448*.

Jørgensen, K., Skjeltorp, J., & Ødegaard, B. A. (2017). Throttling hyperactive robots–order-to-trade ratios at the oslo stock exchange. *Journal of Financial Markets*.

Kamath, R. R., & Wang, Y. (2008). The price-volume relationship in the chilean stock market. *International Business & Economics Research Journal*, *7*(10), 7–14.

Karolyi, G. A., Lee, K.-H., & Van Dijk, M. A. (2009). Commonality in returns, liquidity, and turnover around the world. *Ohio State University. Processed*.

Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. *Journal of Financial and quantitative Analysis*, *22*(1), 109–126.

Kelley, D. (2013). plan: Tools for project planning [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=plan` (R package version 0.4-2)

Kristiania børs. (1919). *Kristiania børs 1819-1919 : et tilbakeblik ved 100 aars jubilæet*. Kristiania: Komiteen.

Kumar, B., Singh, P., & Pandey, A. (2009). The dynamic relationship between price and trading volume: Evidence from indian stock market.

Lamoureux, C. G., & Lastrapes, W. D. (1990). Heteroskedasticity in stock return data: Volume versus garch effects. *The journal of finance*, *45*(1), 221–229.

Lee, B.-S., & Rui, O. M. (2002). The dynamic relationship between stock returns and trading volume: Domestic and cross-country evidence. *Journal of Banking & Finance*, *26*(1), 51–78.

Lee, C. F., & Rui, O. M. (2000). Does trading volume contain information to predict stock returns? evidence from china's stock markets. *Review of Quantitative Finance and Accounting*, *14*(4), 341–360.

Lim, K.-P., & Brooks, R. (2011). The evolution of stock market efficiency over time: a survey of the empirical literature. *Journal of Economic Surveys*, *25*(1), 69–108.

Lo, A. W., & Wang, J. (2000). Trading volume: definitions, data analysis, and implications of portfolio theory. *The Review of Financial Studies*, *13*(2), 257–300.

Lybek, M. T., & Sarr, M. A. (2002). *Measuring liquidity in financial markets* (No. 2-232). International Monetary Fund.

Madhavan, A. (1995). Consolidation, fragmentation, and the disclosure of trading information. *The Review of Financial Studies*, *8*(3), 579–603.

Mestel, R., Gurgul, H., & Majdosz, P. (2003). The empirical relationship between stock returns, return volatility and trading volume on the austrian stock market. *University of Graz, Institute of Banking and Finance, Research Paper*.

Mikalsen, S. (2014). *Aksjer og aksjehandel : hvordan lykkes p børsen*. Oslo: Gyldendal akademisk.

Mjølhus, J. O. (2010). *Finansmarkeder*. Oslo: Cappelen akademisk.

Næs, R., Skjeltorp, J. A., & Ødegaard, B. A. (2008). Liquidity at the oslo stock exchange.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, *116*(2), 257–270.

Osborne, M. F. (1959). Brownian motion in the stock market. *Operations research*, *7*(2), 145–173.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Sabbaghi, O. (2011). Asymmetric volatility and trading volume: The g5 evidence. *Global Finance Journal*, *22*(2), 169–181.

Stoll, H. R. (1978). The pricing of security dealer services: An empirical study of nasdaq stocks. *The Journal of Finance*, *33*(4), 1153–1172.

Wang, Y.-C., Wu, J.-L., & Lai, Y.-H. (2018). New evidence on asymmetric return–volume dependence and extreme movements. *Journal of Empirical Finance*, *45*, 212–227.

Wu, Y. (2017). What factors drive trading around the world?

Ying, C. C. (1966). Stock market prices and volumes of sales. *Econometrica: Journal of the Econometric Society*, 676–685.