Essays on Trading and Manipulation in Financial Markets

Tatiana Franus

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Bayes Business School City, University of London
Faculty of Finance

Supervisor:
Prof. Richard Payne
Dr. Malvina Marchese

January 17, 2024

Abstract

The thesis presents four studies in the field of financial microstructure, specifically trading and manipulation in financial markets. The first two studies aim to investigate the effect of spoofing manipulation on intraday market quality and forecast the market state with high spoofing manipulation risk. The latest two studies focus on informed trading and information incorporation into asset prices.

The first paper studies the intraday relationship between spoofing manipulation activity and market quality in the automated equity market on the Moscow Exchange (MOEX). We find that higher spoofing activity is associated with lower intraday market quality (greater quoted and effective spreads and greater volatility). This effect is economically significant and robust to different specifications, endogeneity, and alternative spoofing measures. Our results hold after controlling for volatility, day trading volume, and intensive trading periods during the day.

The second study introduces a data-driven approach to forecast the market state with high spoofing risk. The approach reduces model selection's importance through forecasts combining different machine learning predictors. We apply the algorithm to a unique dataset of suspicious spoofing cases detected on MOEX. We show that learning from the limit order book using machine learning techniques generates an effective manipulation prediction measure. Our study introduces an indicator of real-time risk to trade in a manipulative environment that exchanges and regulators could utilise for their surveillance systems. Our approach achieves significant forecasting accuracy in a high-frequency environment.

The third study is an empirical investigation of informed trading in the futures market. Using the comprehensive data with the customer type indication from MOEX, we examine trading by different customer groups and find that retail traders forecast intraday returns in a high-frequency time dimension. Institutional traders effectively predict short-term returns while losing their forecasting power after four trading days. We find that different customer groups systematically trade in opposite directions, and their order flows are highly informative about intraday returns.

Finally, the fourth study examines price discovery dynamics between Bitcoin exchange-traded products (ETPs) and spot markets on centralised cryptocurrency exchanges. We apply four popular price discovery measures to ETP and spot transaction data. Our results show that price discovery is dominated by the spot market across all measures and sampling frequencies. This implies that ETP markets play a smaller role in incorporating new information about Bitcoin prices and that informed investors largely prefer to trade on spot markets.

Four essays form coherent research motivated by the increasing speed of development of the electronic markets, the rise of high-frequency trading with the introduction of new possibilities for market destabilisation, and natural demand from investors for a higher-quality trading environment and diversification strategies in new asset classes and markets.

Acknowledgements

First and foremost, I want to thank my supervisors who guided me throughout this research, for which I am deeply grateful:

- Prof. Richard Payne for his invaluable advice, endless support and guidance. His immense knowledge and bountiful experience have encouraged me to become a fully-fledged academic researcher. I especially want to thank him for introducing me to strategic thinking and a wise approach to seeing economic sense in any research experiment.
- Dr. Malvina Marchese for her excellent assistance, feedback and encouragement in various stages of my PhD. She introduced me to various statistical testing, the importance of robustness checks and forecast combinations.

I am also grateful to the PhD department and the Faculty of Finance at Bayes Business School, who contributed through advice, assistance and encouragement to the completion of this research. I am also incredibly grateful to Prof. Barbara Casu for her support throughout these years.

Most importantly, I would like to sincerely thank my family for their love and support throughout my PhD journey.

Declaration of Authorship

I, Tatiana Franus, declare that this thesis titled, "Essays on Trading and Manipulation in Financial Markets" and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at Bayes Business School.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: TATIANA FRANUS

Date: January 17, 2024

Contents

List of Figures

List of Tables

Introduction

This thesis brings together a number of studies written over the last four years, all of which focus on financial microstructure in different markets. The motivation for the research comes from the increasing speed of development of the electronic markets and the rise of high-frequency trading with the introduction of new possibilities for market destabilisation and natural demand from investors for a higher-quality trading environment and diversification strategies in different asset classes and markets. All studies in the current thesis use intraday time dimension with tick time increments as a platform for market microstructure investigation.

Chapter 1 and **2** research a world-spread intraday order-based manipulation called spoofing. Significant resources have been invested in automated surveillance systems to investigate and detect manipulative behaviours in recent years. The greater availability of high-frequency and detailed order book data has increased interest. Furthermore, technological growth with novel machine learning approaches brings new methods for complex market microstructure research to life. By developing trading automation, new possibilities of disruptive practices are introduced when traders make tremendous profits by artificially affecting market beliefs and negatively affecting other market participants. They significantly threaten the trust and integrity of capital markets through mispricing and market imperfections. These harmful practices have increased over the last decade, and our study focuses on analysing and forecasting spoofing events.

Chapter 3 and **4** in the current thesis focus on informed trading and information incorporation into asset prices. Information arrives at securities markets through price quotes and trades. As price responses signal informed trades, consistent profits gained from positions or trading activity indicate who is informed. Most previous studies have analysed informed trading with data on different investor types in the foreign exchange, equity and bond markets. In Chapter 3, we analyse how the order flow of different client types affects market outcomes, allowing us to discuss informativeness in the future market. Informed traders impose adverse selection costs on quote suppliers. The process through which new information is efficiently incorporated into asset prices is less straightforward when trading in an asset is fragmented across multiple venues or markets. In such a scenario, it is interesting to identify where price discovery takes place Hasbrouck (1995). In Chapter 4, we analyse the incorporation of new information about Bitcoin prices on the exchange traded products and spot markets.

Chapter 1

In Chapter 1, we discuss the problem of defining spoofing orders in the limit order book and investigate if those orders have any effect on the market quality. For our research, we use historical information on all orders placed on the stock market for 91 of the most traded and liquid stocks on the Moscow Exchange (MOEX) from January to June 2019. We track each order by its identification number and determine if it was cancelled or executed. We adapt the algorithm of Lee et al. (2013) to identify spoofing orders in our empirical analysis. Our spoofing identification algorithm relies on, amongst other things, the following components: the number of ticks the order is away from the best price, the lifetime of the order and the minimum order size. Different combinations of parameters give us 54 versions of the spoofing order identification algorithm. We keep the variety for robustness and discuss how further research could adapt the algorithm depending on the market and asset class.

We then relate spoofing intensity to well-investigated market quality measures such as volatility and liquidity. Volatility is measured as a one-minute standard deviation of trade prices normalized by midquote, while quoted and effective spreads measure liquidity. To understand the effect that spoofing has on market quality, we check market quality measures before and after the spoofing manipulation. We run a series of panel regressions that control for factors associated with changes in market quality. Variables are constructed for 10-, 30- and 60-minute time windows.

Our finding shows that the ratio of cancelled orders to all placed on MOEX is, on average, 83.88%. Our measure of possible spoofing manipulation activity is defined as the “spoofing ratio” or SP, which is a ratio of identified possible spoofing orders by the algorithm to all placed orders for the specific stock. SP ratio varies from 0.84% to 4.6% for different algorithms. Overall, 93% of all identified spoofing orders belong to the group of the most liquid stocks, while 90% of spoofing orders for these stocks lie before the 45$^{\text{th}}$ order book price level and 70% of spoofing orders – before the 23$^{\text{rd}}$ level.

Our key results in are as follows. Our measure of spoofing manipulation, SP, tends to be associated with a decrease in the change in quoted and effective spreads and short-term volatility just before spoofing manipulation periods. However, higher SP tends to be associated with a significant rise in the change in quoted and effective spreads and short-term volatility after spoofing manipulation periods. More spoofing leads to rising spreads and volatility in the next period. Relationships in all panel regressions show similar results.

We explain our finding in a way that manipulation creates a fake or temporary perception in the market and can subsequently make traders less confident about the level of the actual asset price. This effect is economically significant and robust to different specifications, endogeneity tests, and alternative modifications of SP. Our results hold after controlling for volatility, day trading volume, and periods of intensive trading during the day. So, our study suggests that intense spoofing activity is associated with degraded market quality.

Chapter 2

As we find a harmful effect of spoofing on intraday market quality, the problem of spoofing detection in real-time becomes more vital. Significant resources have been invested in automated surveillance systems to detect price-manipulating behaviour, especially since the Dodd-Frank Act made spoofing illegal (Dodd-Frank (2010)). We continue investigating spoofing activity in Chapter 2 and seek to forecast the appearance of disruptive practices.

Chapter 2 introduces a data-driven approach to forecast the market state when spoofing manipulation is likely to appear. We train ML methods on suspect spoofing cases detected on MOEX and present a novel Real-Time Spoofing Probability (RTSP) measure to indicate a spoofing manipulation risk. Our main insight is that learning from the limit order book and data on suspected spoofing cases generates an effective manipulation prediction measure.

In our study, firstly, we identify a hundred predictors of the market state and use lasso and elastic net algorithms as variable selection methods. Secondly, we match suspect spoofing orders with all orders and trades, allowing us to track the manipulative order's lifetime and order book price level. We choose a ML methodology as spoofing events are not equally spaced within an extensive dataset. Our selected models (Random Forest, XGBoost, and Decision Trees) correctly predict over 70% spoofing events for balanced data. We use cross-validation to avoid overfitting and keep the analysis out-of-sample. We also use different cross-validation checks for imbalanced data. Finally, we introduce our RTSP measure, which forecasts intraday manipulative activity. To test our measure on out-of-sample data, we train ML algorithms on five previous trading days and run a rolling cross-validation forecast for the next 10, 30, and 60 minutes. We endogenize the RTSP measures as a simple average of ML outcomes.

Consequently, RTSP shows a probability in real time that the market state is preferable for the spoofer to place their order. Moreover, we compare the forecasting performance of RTSP to the performance of other ML algorithms and show that our designed methodology achieves better results in the given environment. The model is self-training, so no adjustments are needed to detect other manipulative practices. This approach reduces the importance of model selection through forecast combination. An empirical evaluation of the proposed framework demonstrates significant forecasting accuracy in a high-frequency environment. We employ the model to discuss market regulation, particularly financial market surveillance.

Chapter 3

Our study in Chapter 3 contributes to the debate on informed trading by exploiting data of the futures market from MOEX with the indication of the investor type who initiated the trade. The paper addresses a market microstructure and market design question of how different client trading practices affect market outcomes from intraday and daily investment perspectives. By answering this question, we improve our understanding of what drives end-user's demand for futures contracts. We tackle these questions empirically using a dataset covering twenty months in 2022 and 2021 of every trade on the twenty seven most liquid future contracts on MOEX.

We contribute to the discussion about retail investor behaviour and future returns. Firstly, we find a positive relationship between retail flows and future prices, which could be seen from the perspective that retail investors drive future returns. As we run an analysis on the future market that is by definition driven by the price of the underlying asset, we conclude that retail investors are informed about future price movements of the underlying asset (Kaniel et al. 2012, Kelley & Tetlock 2013, Boehmer et al. 2021). Kaniel et al. (2012), Barrot et al. (2016) claim that retail investors are rewarded for providing liquidity to institutional investors. Barber et al. (2008)

show that retail investors' order flows are positively autocorrelated and thus forecast short-term price pressure. The willingness of retail investors to provide liquidity and autocorrelated order flows may contribute to the short-term predictiveness of the retail order flow.

Also, we find that the retail selling volume anticipates negative returns in the next period. Nevertheless, we cannot observe whether the seller-initiated trade was closing the long position or a short sell; our results are most consistent with the information hypothesis that retail short sellers possess and act on unique information beyond that held by other investors. Under this theory, retail short selling predicts negative returns as prices of the underlying assets converge to their fundamental values, just as informed order flow predicts returns in models such as Kyle (1985). Moreover, we find that in a very short-term (30- and 60-second period analysis), retail traders provide liquidity to institutional investors that need to execute their trades immediately, as suggested by Kaniel et al. (2008). Furthermore, our results indicate the liquidity provision role of retail traders to corporate clients.

We also identify how other investors' types differ. Retail traders', institutional investors' and non-residents' order flows are good predictors for intraday returns in commodity futures, while only institutional investors' order flow positively predicts returns in stock futures. Intraday order flows of corporate clients, retail and non-resident traders predict returns on currency futures, while the order flow of institutional investors does not.

We also find that customers systematically trade in opposite directions: corporate clients trade opposite to retail and institutional investors, retail traders trade in the opposite direction to non-residents, and generally in the same direction as institutional traders. However, one should run analysis separately on different time dimensions, as dealers' order flow, for example, has different correlation signs depending on the time-frequency choice. We show empirically in three perspectives (correlation, regression and portfolio analyses) that the order flow signal differs for daily and intraday trading, and the study needs to be built separately for intraday frequencies such as high-frequency trading, algorithmic trading, and daily trading.

Chapter 4

Our empirical study in Chapter 4 contributes to the literature on price discovery in Bitcoin markets and examines the price dynamics of Bitcoin Bitcoin Exchange-Traded Products (ETPs) in relation to spot markets. The process through which new information is efficiently incorporated into asset prices is more complex when trading an asset is fragmented across multiple venues or markets. In such a scenario, it is of interest to identify where price discovery takes place (Hasbrouck 1995).

Crypto spot exchanges have attracted significant interest from both retail and institutional investors. As regulations constrained the ability of traditional funds and banks to participate in these exchanges, an opportunity arose to create a more traditional product, allowing exposure to Bitcoin and other cryptocurrencies. Thus, Bitcoin Exchange-Traded Products allow investors on traditional equity exchanges to gain exposure to the underlying asset without the need to hold Bitcoin.

We apply four popular measures of price discovery to Bitcoin ETP and spot exchange data. The Information Share (IS) (Hasbrouck 1995) estimates the proportion of the efficient price innovation variance explained by innovations stemming from the different markets. Alternatively, the Component Share (CS) approach (Booth et al. 1999, Chu et al. 1999, Harris et al. 2002) adopts the permanent-transitory decomposition technique in Gonzalo & Granger (1995). Yan & Zivot (2010) and Putniņš

(2013) show that a combination of the two measures can remove dependence on noise and liquidity shocks. We use Information Leadership (IL) (Yan & Zivot 2010), and the Information Leadership Share (ILS) (Putniņš 2013) models to make our finding robust.

The study shows that spot markets dominate the price discovery process, suggesting that ETPs tend to lag in terms of informational efficiency. Spot markets dominate this process due to their deeper liquidity, continuous trading hours, and a greater degree of anonymity. Nonetheless, ETPs may play a more significant role in the future as this market matures and complies with regulatory frameworks, thus gaining popularity among institutional investors.

Direction of future research

To increase the impact of the contribution of this thesis, one can apply our algorithm of spoofing identification to another market. Also, a similar methodology could be used to identify other trade-based and order-based manipulations in high-frequency time dimensions, for example, layering, pinging, adjusting the bid, pump-and-dump. Exchanges and regulators may use our ML measure RTSP to forecast spoofing risk to increase market quality. Having our results, researchers may continue improving the detection models while using arising modifications of ML techniques with new adjustments. As we show how to detect market state with high spoofing manipulation risk, theoretical models of rebalancing the order book as a prevention mechanism might be further developed and tested using a simulation approach.

We believe that asset markets behave as social organisms accumulating millions of investors' actions, constantly changing over time as life continues. So, the price discovery question is one of the topics that scholars investigate repeatedly. Our finding addresses the question of information incorporation into prices, which one may further develop regarding the impact of trading volume on price discovery. For example, combining our findings and unique dataset of client order flow with a bias-free generalised approach to measuring information shares (GIS) may be a potential further research area (Hagstromer & Menkveld 2023).

Our findings in Chapter 4 shield the light on price discovery on the new asset class, cryptocurrency, and might be used for further discussion from theoretical price discovery topics on the alternative market and empirical research using similar datasets.

Chapter 1

Spoofing Manipulation and Intraday Market Quality

Parts of this chapter have been presented by Tatiana Franus on:

- December 2020. World Finance and Banking Symposium, Riga, Latvia.
- June 2021. Market Microstructure Summer School, Stockholm Business School, Sweden.

1.1 Introduction

Electronic markets with the automation of trading and high-speed trading have transformed the financial market landscape. By developing trading automation, new possibilities of disruptive practices are introduced, when traders make tremendous profits by artificially affecting market beliefs and negatively affecting other market participants. Significant resources have been invested in automated surveillance systems to detect manipulative behaviours; however, little is known about how introducing these new systems affects market quality. Angel & McCabe (2010) explain that manipulation in attempting to distort market price away from its economic fundamentals is an unethical business practice related to speculative activity.

Our research aims to investigate manipulative practices in financial markets and analyze the relationship between manipulation and market quality. Our research focuses on spoofing manipulation on the financial market Moscow Exchange (MOEX), which has not been examined before, and we find that market quality on MOEX changes due to the presence of spoofing manipulation. Our study suggests that intense spoofing activity is associated with degraded market quality, measured by various proxies, including short-term volatility, quoted and effective spreads.

In 2017, Citigroup was fined the biggest amount ever levied of $25 million in a spoofing case for manipulating the U. S. Treasury futures market. In 2020, U. S. regulators levied a $920 million fine on JP Morgan Chase for eight years of spoofing manipulation in markets for precious metals and treasury bills. In 2019, Tower Research was ordered to pay USD $67.4 million in fines to the CFTC to settle allegations that three former traders at the firm engaged in spoofing. These examples shows that the financial exchange market surveillance has attracted much attention across different exchange markets since the Flash Crash in 2010 and since the Dodd-Frank Act made spoofing illegal (Dodd-Frank 2010). However, the lack of research in effective and efficient detection methods and algorithms in both industry and academia causes challenges for regulators who are required to monitor huge volumes of trading activity in real time.

Market manipulation strategies represent a source of price distortion and the creation of artificial market conditions. They create a significant threat to trust in and integrity of capital markets through mispricing and market imperfections. They harm investors' confidence, resulting in less participation of investors and hence may adversely affect efficiency, liquidity, integrity, and development of the stock market (Guiso et al. 2008, Imisiker & Tas 2013, Punniyamoorthy & Thoppan 2012). These harmful practices have increased over the last years. In 2018, the CFTC created a Task Force on spoofing to target this sort of misconduct, indicating that this type of market manipulation is of high practical importance.

Interestingly, the responsibility for detecting manipulation lies with those who trade. All firms that engage in trading must have monitoring in place to check their activity for signs of market manipulation. This gave rise to a sizable and fast-growing "trade surveillance" industry, that aims to monitor the clients' trades and detect illegal trading activity to improve the quality of financial markets (Cumming et al. (2011), Aitken, Harris & Ji (2015)).

Our research is related to a small but growing body of literature that addresses issues concerning trade-based manipulation.

Allen & Gale (1992) proposed a model for trade-based manipulation and showed that manipulation is theoretically profitable. In trade-based manipulation, the agents use trading activities to create price momentum, such as buying and selling, without using false information or altering the firm value. This price momentum is generated by a large trader that can affect prices by significantly changing the market maker's order flow Jarrow (1992). This manipulation could be profitable only in markets with information asymmetry Allen & Gorton (1992) as the natural asymmetry between liquidity purchases and sales gives rise to profitable trade-based manipulation.

An "equilibrium model" was derived and proved that noise traders' existence made it possible to manipulate the price, although theoretically, no profit should be expected according to the efficient market hypothesis (Allen & Gorton (1992)). A real price manipulation case conducted by large traders was examined and analyzed in the research undertaken by Jarrow (1992). The actual case proved that the manipulation tactic could make a risk-free profit due to the significantly changing order flow. In 2020, Williams & Skrzypacz (2020) showed that spoofing may occur in equilibrium and studied its equilibrium consequences in a formal theoretical model. They present a dynamic trading model and show that equilibrium spoofing slows price discovery, raises bid-ask spreads, and raises return volatility.

Cartea et al. (2020) derived an optimal trading strategy for an investor to improve the rate and price at which she sells shares with limit orders in an order-driven electronic market. They provided a mathematical framework to develop optimal spoofing strategies (Cartea et al. 2020).

Aggarwal & Wu (2006) empirically showed that manipulation is associated with higher stock volatility, greater liquidity, and high returns during the manipulation period. A comprehensive empirical study of spoofing was carried out on data from the Korean Exchange (Lee et al. (2013)). The authors defined a particular type of spoofing strategy and showed that spoofing leads to substantial extra profits. They empirically found that stocks with higher volatility of returns, lower market capitalization, lower price levels, and lower marginal transparency are more prone to spoofing. Another piece of empirical research conducted by Wang (2015) uses historical data from the Taiwan Futures Exchange to analyze spoofing manipulation. The principal findings of this work are that spoofing increases price volatility and trading volume and increases the quoted spread. Aitken, Harris & Ji (2015), in their research on market quality, develop a business ethics framework for assessing security market quality and measure the incidence and impact of prohibited trading behaviour. They question whether a pattern of market integrity violations affects market efficiency. Several empirical studies of the options market confirm the relationship between manipulation and increased volatility and wider spreads (Stoll & Whaley 1987, Chamberlain et al. 1989, Chiou et al. 2007).

Spoofing manipulation involves submitting orders to the market without the intention to execute these orders with the idea to influence other background traders' behaviour. Spoofing orders are often managed by algorithmic trading strategies that help to execute trades in milliseconds. The distinctive feature of the spoofing strategy is often implemented at relatively high frequencies, with positions typically opened and closed intraday (Putniņš 2020). Spoofing manipulation is a worldwide strategy and has been detected in data from the US, UK, EU, Korea, Hong Kong, and other exchanges.

US financial law defines spoofing directly as a criminal and civil offence. Besides, it not only contains regulations that prohibit manipulation but spoofing is specifically forbidden. In contrast, spoofing itself is not considered a criminal offence under British law. Nevertheless, the EU Market Abuse Regulation (MAR) in which spoofing is captured as opposed to general 'market manipulations' still applies to the UK as the procedure for leaving the EU is not yet complete, and therefore spoofing behaviour may lead to civil or administrative prosecution under MAR (Montgomery 2016, Stephens et al. 2019).

The trader's intention on US financial markets is central to law enforcement. The US authorities need proof that the trader's original purpose at the time of the order's placement was to cancel the order in any case. On the other hand, the UK authorities will tend to regard how the market was affected by the order (Stephens et al. 2019). In criminal cases in the United Kingdom, it must also be shown that there was some intention to create an impression in the market, but not that the trader wanted to cancel the order before its execution (Stephens et al. 2019).

In recent years, there has been a trend towards global regulation of financial market participants whose activities are related to algorithmic trading. For manipulation detection, authorities usually use statistical analysis to determine whether a trader's strategy is based on spoofing and examine emails and other correspondence for signs of deceptive intent. Specific requirements have been introduced or declared by legislative acts of individual states to such persons. For example, in May 2013, the legislative act "High Frequency Trading Act" was approved in Germany. This act contains the following requirements for HFT market participants:

- mandatory licensing of all HFT algorithms from the regulator;
- each HFT request must contain the ID of the specific algorithm that it was submitted using;
- HFT must independently monitor the Order-to-Trade Ratio (the ratio of the volume of orders to the transactions volume).

Also, apart from the requirements for algorithmic market participants, there are requirements for licensed investment companies. In particular, according to the MiFID II Directive, which applies to the European Union's financial markets, investment companies that provide direct electronic access to the exchange or trading platform must have defined effective control systems. Exchanges and trading platforms need to ensure correctness in trading and face large flows of information. Thus, they implement specific regulatory measures, such as various types of stress testing. The impact of increased stress is assessed loads on the mechanisms of their functioning and performance. In 2014, the American CME exchanges, CBOT, NYMEX, and COMEX, adopted the regulatory document Rule 575, which defined many trade practices that violate fair trading activity. Some exchanges accept their own technological and economic measures to fight against unfair practices of HFT, for instance, additional commission fees for large quantities.

In Russian legislation, the concepts of HFT and spoofing are not entirely fixed. To discourage spoofing activity regulators make the definition ambiguous. Manipulative activities are regulated on General grounds by Federal law "On the securities market" No. 39-FZ of 22.04.1996 and by Federal law "On countering the misuse of insider information and market manipulation and on making changes in certain legislative acts of the Russian Federation" No. 224-FZ of 27.07.2010 which establish a ban on the activities recognized as manipulation of the market, as well as the misuse of insider information[1]. Article 5 of this Act provides a detailed description of actions defined as manipulation. The integral part of the definition of manipulation is the need for substantial market reaction, which is defined by methodological recommendations of the Central Bank. The law defines the manipulative actions for

[1] Review of the regulation of financial markets, 2016 Central bank of Russia

any type of financial assets traded on the market (stocks, bonds, derivatives, currency, commodity) as mentioned in Table 1.1 [2]. Nevertheless, spoofing as a term is not precisely defined in the law, the misleading actions of the traders with placed orders without the intention to execute them are prohibited and correspond with our definition of spoofing manipulation with the only difference, that in the law the manipulation is a repeated action and should lead to the deviation of the asset price from its fair value (line (7) in Table 1.1).

	Trader's action	**Market reaction**
1	Intentional dissemination of false information	Significant deviation (or maintenance) of the price, demand, supply, or trading volume
2	Actions with preliminary agreement between traders	Significant deviation (or maintenance) of the price, demand, supply, or trading volume
3	Execution of transactions under which the obligations of the parties are performed at the expense or in the interest of one person	Significant deviation (or maintenance) of the price, demand, supply, or trading volume
4	Placing orders at the expense or in the interest of one person with two or more orders of the opposite direction appear simultaneously, and if those orders are executed.	Significant deviation (or maintenance) of the price, demand, supply, or trading volume
5	Repeated transactions on the best bid/ask prices for the purpose of subsequent opposite transactions at the same prices	Significant deviation of the price, demand, supply, or trading volume
6	Repeated transactions to mislead the price	Maintaining a price that is significantly different from a fair price
7	Repeated not executed operations and performing operations without intention to fulfill them	Significant deviation (or maintenance) of the price, demand, supply, or trading volume

TABLE 1.1: Trading action classification. Trading actions relating to market manipulation in Russian Federation according to Federal law "On countering the misuse of insider information and market manipulation and on making changes in certain legislative acts of the Russian Federation" No. 224-FZ of 27.07.2010.

The Moscow Exchange is the largest exchange group in Russia that operates trading markets in equities, bonds, derivatives, the foreign exchange market, money markets, and precious metals. It was established in 2011 by merging the two largest Moscow-based exchanges, the Moscow Interbank Currency Exchange (MICEX) and the Russian Trading System (RTS); MICEX and RTS had been formed in 1992 and 1995. They were the leading Russian exchanges for two decades. The merger created a single entity referred to as MOEX [3].

All stocks traded on MOEX are sorted by the current market risk rates or listing levels. The higher the level of the stock, the less risky the stock is. The table in Appendix 1 shows the main requirements for a stock to be placed in a specific listing level. To understand the differences in stocks from 1^{st}, 2^{nd}, and 3^{rd} listing levels, we conduct a statistical analysis of all stocks' properties from MOEX. We take such stock properties as market capitalization, transaction volume, order volume, bid-ask spread, and return volatility. Capitalization was taken for the date of the 16^{th} of April 2019. We measure other properties using the data for six months of 2019 from January to June. The average bid-ask spread is calculated per minute, while return volatility is calculated for daily returns. According to the data in Table 1.2 there are clear differences between stocks of different listing levels:

[2]Federal law "On countering the misuse of insider information and market manipulation and on making changes in certain legislative acts of the Russian Federation" No. 224-FZ of 27.07.2010 http://www.consultant.ru/document/cons_doc_LAW_103037/

[3]https://www.moex.com/en/

Stock listing levels	1^{st} **level**	2^{nd} **level**	2^{nd} **level***	3^{rd} **level**	3^{rd} **level***
Average capitalization (bn RUB)	692 751	96 479	33 917	44 711	30 082
Average daily transaction volume	8 270	1 176	681	275	238
Average daily order volume	62 863	9 137	5 012	1 875	1 411
Average bid-ask spread, %	0.08	0.49	0.59	1.37	1.41
Average return volatility, %	1.19	1.72	1.81	2.93	2.94

TABLE 1.2: Descriptive statistics of stocks grouped by listing levels on MOEX. 2^{nd} level* shows the group without outliers (SNGS, SNGSP), 3^{rd} level* shows the group without outliers (SIBN).

The average capitalization of companies very clearly shows the differences between the levels. So, for the companies of the 1st listing level, the average capitalization is equal to 692.75 bn rubles, while for the companies of the 2^{nd} and 3^{rd} level the average market caps are 33.92 bn rubles and 30.08 bn rubles (excluding outliers). 75% of 1^{st} list level companies have a capitalization from 100 billion to 4.3 trillion rubles (the largest being Gazprom), and only a few are less than 50 bn RUB. Among such companies, RSTIP stands out with a capitalization of only 3 bn rubles. 75% of 2^{nd} level companies have a capitalization below 50 billion rubles. The outlier here is SNGS, with a capitalization of 1.2 trillion rubles. 3^{rd} level companies are characterized by a capitalization of fewer than 50 bn rubles (85% of companies), while 35% of 3^{rd} list level companies have a capitalization of fewer than 1 bn rubles. An obvious outlier in the 3^{rd} listing level is SIBN, with a capitalization of about 1.5 trillion rubles.

The average daily transaction volume also differs dramatically among levels: the 1^{st} listing level stocks have a figure almost 7 times that of the 2^{nd} listing level stocks and 35 times that of the 3^{rd} list level companies. The average daily order volume shows that companies of the 1^{st} listing level have an average of 62 863 orders per day, which is seven times higher than the 2^{nd} listing level companies and almost 33 times higher than 3^{rd} listing level companies.

Transaction and order volumes clearly show the differences between the levels. On average, the more orders and transactions per day, the higher the company's listing level. The average daily bid-ask spread is inversely proportional to the listing level of the company. The spread is smallest for the 1^{st} list level companies with an average of 0.08%; however, it rises dramatically to 1.4% for the 3^{rd} level companies. The average return volatility increases from 1.19% to 2.93% from the 1^{st} to the 3^{rd} listing level companies, which is reasonable due to the lack of trading for less liquid stocks that prevents the stock price from moving smoothly intraday.

Overall, after analyzing the stocks from different listing levels, we can conclude that 1^{st} and 2^{nd} listing level stocks have properties that allow the market participants to trade on an intraday basis, which is essential for our research. Stocks from the 3^{rd} listing level stocks do not have enough trading and order volume and suffer from high volatility.

For our research, we use historical information on all orders placed on the stock market for 91 of the most traded and liquid stocks on MOEX from January to June 2019. We adapt the algorithm of Lee et al. (2013) to identify spoofing orders in our empirical analysis. The algorithm relies on, amongst other things, the following components: the number of ticks the order is away from the best price, the lifetime of the order and the minimum order size. Different combinations of parameters give us 18 versions of the spoofing order identification algorithm. We then relate spoofing intensity to well-investigated market quality measures such as volatility and liquidity. Volatility is measured as a one-minute standard deviation of trade prices normalized by midquote, while quoted and effective spreads measure liquidity. To understand the effect that spoofing has on market quality, we check market quality measures before and after the spoofing manipulation. We run a series of panel regressions that control for factors associated with changes in market quality. Variables are constructed for 10-, 30- and 60-minute time windows. Our finding shows that the ratio of cancelled orders to all placed on MOEX is, on average, 83.88%. Our measure of possible spoofing manipulation activity is defined as the "spoofing ratio" or SP, which is a ratio of identified possible spoofing orders by the algorithm to all placed orders for the specific stock. SP ratio varies from 0.84% to 4.6% for different algorithms. Overall, 93% of all identified spoofing orders belong to stocks from the most liquid stocks, while 90% of spoofing orders for these stocks lie before the 45$^{\text{th}}$ order book level and 70% of spoofing orders – before the 23rd level.

Our key results are as follows. Our measure of spoofing manipulation, SP, tends to be associated with a decrease in the change in quoted and effective spreads and short-term volatility just before spoofing manipulation periods. However, higher SP tends to be associated with a significant rise in the change in quoted and effective spreads and short-term volatility after spoofing manipulation periods. More spoofing leads to rising spreads and volatility in the next period. Our findings are in line with those of Wang (2015) on data from the Taiwan Futures Exchange and of Brogaard et al. (2022) on data from the Toronto Stock Exchange. Both papers show that spoofing orders lead to greater spreads and higher market volatility, while Brogaard et al. (2022) also show that spoofing slows price discovery. However, in contrast to the abovementioned papers, we use intraday time dimensions according to the nature of the spoofing strategy and contribute to the literature with broader results of market quality instability around the high intensity of spoofing events on the stock market.

Relationships in all panel regressions run in the research show similar results. This finding makes economic sense as manipulation creates a fake or temporary perception in the market and thus can subsequently make traders less confident about the level of the 'true' price. This effect is robust to different specifications and several variations of SP. Our results hold after controlling for volatility, day trading volume, and periods of intensive trading during the day.

The remainder of this paper proceeds as follows. In Section 1.2 we discuss the data we employ and in Section 1.3 we define our spoofing detection algorithm and report the empirical results and statistical analysis. In Section 1.4, we show how we build the SP measure and our market quality measures; in Section 1.5, we present the methodology used in our main empirical work, and in Section 1.6, we present the results, including the robustness and endogeneity tests. Section 1.6.1 looks at possible economic explanations for the effect of SP and its economic significance. We conclude in Section 1.7 and collect tables in Appendices 1.8.

1.2 Data

In the research, we use market data from six consecutive months from January until June 2019 from MOEX. We use historical information on all orders placed on the stock market. The data include the following information for each order: record number, security code, type of the order (sell or buy), time (in the format of microseconds), order number, type of the action (placed, cancelled, executed), price, volume, trade number (if executed), and trade price (if executed).

Before selecting the list of companies for the analysis, we remove stocks that had at least one day with no transactions. We select the companies for our analysis based on three parameters: capitalization, average daily transaction volume, and average daily order volume.

The selection of companies is based on a ranking of these measures. The main characteristic of the final selection was the overall ranking, which is calculated as follows: $OVERALL\_RANK = 1.5(NormRank\_of\_capitalization)+$
$+1.25(NormRank\_of\_transaction\_volume) + (NormRank\_of\_order\_volume)$.

The ranks were normalized by the MinMax scaling method. So, capitalization is considered the most important measure, followed by the ranks of the transaction volume, and the least important is the rank of the order volume. However, changing ranking coefficients does not lead to a change of the stock list. We don't include stocks with low trading and order volume where spoofing manipulation is irrelevant by definition.

We run our further analysis on the first 80% of stocks from MOEX based on the ranking presented above. The full list of 91 stocks used in the research is presented in Appendix 2. $OVERALL\_RANK$ allowed us to include all stocks from the 1^{st} listing level and some stocks from the 2^{nd} and 3^{rd} listing levels.

1.3 Definition and identification of spoofing

Spoofing manipulation is one of the possible order-based manipulation practices in financial markets. It is an illegal trading strategy. The Dodd-Frank Act describes spoofing as "bidding or offering with the intent to cancel the bid or offer before execution" (Dodd-Frank (2010)). Spoofing manipulation involves submitting orders to the market without the intention to execute these orders with the idea to influence other traders' behaviour. A simple example is as follows. A trader wishes to buy a stock, but at a price below the current best offer. She, therefore, places a very large spoofing limit sell order near the best offer and, simultaneously, she places a small limit buy order at the best bid. It is hoped that the large limit sell (i) does not execute and (ii) due to its size causes others to revise their valuation of the stock downwards. Effect (ii) causes other traders to hit her order at the best bid and when it happens, the trader removes the large limit sell order. Ultimately, she bought the stock more cheaply than she would have done without spoofing through the use of a limit sell order that she never intended to execute. Of course, a spoofing order may well be implemented by an algorithm rather than a human trader.

The precise definition of spoofing varies in the literature. Lee et al. (2013) defined "a spoofing order as a bid or an ask with a size at least twice the previous day's average order size and with an order price at least 6 ticks away from the market price, followed by an order on the opposite side of the market, and subsequently followed by the withdrawal of the first order". Cartea et al. (2020), in the article "spoofing and price manipulation in Order Driven Markets," define spoofing as a combination of placing sell limit orders at the best ask price and a spoof buy limit orders at the best bid price with a large volume. According to the author, the size of spoof buy limit orders depend on the volume required to tilt the limit order book into a buy-heavy regime. Brogaard et al. (2022) define spoofing similarly as Lee et al. (2013), but the manipulative order is placed inside the spread to mimic buying or selling pressure

The first group of scholars defines spoofing manipulation as only one spoofing order from an account with a significant volume placed on the market to mislead other market participants with the wrong demand or supply. Other traders who use order book information in their trading strategies to enter the market see the big ask order several points higher than the best ask in the order book. They understand that a big seller comes to the market, and if his order is executed, then the asset's price will go down. In that case, the safe strategy for noise traders will either not enter the market or place an ask order. All these actions could lead the market price to go down. When the market went down by several points, the manipulator enters the market with a small buy order. He cancels his big ask order, and the market simultaneously rebounds back to the before-spoofing price levels. This strategy is a low-risk strategy to earn small profits. However, if done many times on different assets, this strategy can lead to tremendous profits for the spoofer. The bid spoofing order strategy could be done with mirror actions.

Other scholars define spoofing more broadly as a category of four types of manipulation, often but not always implemented via a computer algorithm, including layering, advancing bid/offer, quote stuffing, and pinging or phishing (Putniņš 2020). According to this classification, the "layering" manipulation type is similar to the definition above of Lee et al. (2013), with the only difference being that Putniņš (2020) says that orders could be placed on one or several price steps. Layering as a type of order-based market manipulation occurs when a trader places multiple orders at different price levels close to each other for the same financial asset to create heavy buying or selling interest. The trader then subsequently cancels the orders. Manipulative orders generate a layer of significant volume from one side and mislead the market with fraudulent supply or demand information. Layering orders could be placed from one account or several accounts. The main feature of these orders is that most of them are intended not to be executed, leading to their cancellation. It is considered a more sophisticated form of spoofing. A typical layering cycle is as follows: (i) place a small sell order at or near the best ask price, (ii) layer the bid side of the order book until the market moves up and the small sell order executes, (iii) cancel the layering bid orders and repeat the above steps in the opposite direction (Putniņš 2020).

Other manipulation strategies described by Putniņš (2020) include "Advancing the bid/offer", which is an aggressive layering strategy with orders placed as a new best bid or best offer without intention to execute them, and is highly related to the definition of spoofing by Cartea et al. (2020). "Quote stuffing" is a strategy of submitting an enormous number of order submissions, amendments, and cancellation messages in a short period. This definition coincides with the one given by Egginton, Ness & Ness (2016): "quote stuffing is a practice where a large number of orders to buy and sell securities are placed and then cancelled almost immediately. "Pinging", or "phishing", which is the fourth type of spoofing manipulation strategy, involves submitting small probing orders to detect hidden or latent liquidity (Putniņš 2020). Such techniques are unlikely to be manipulative as they don't mislead other participants and don't generate a supply and demand imbalance; however, Canadian and European regulators have expressed views that such trading is considered market

FIGURE 1.1: Sell spoofing order

manipulation[4]. Wang & Wellman (2017) in their agent-based model research define spoofing manipulation in the same manner as Lee et al. (2013).

The literature shows that all definitions have main futures in common. First, spoofing manipulation is an order-based manipulation with no intention of manipulating orders to be executed but with the idea to mislead other market participants. Secondly, the main future of the spoofing order is its cancellation in the vast majority of cases. Thirdly, spoofing manipulation is done via placing large orders in the market so that traders or trading algorithms identify orders with bigger than average volume. The size of the spoofing order volume varies in the different research. Finally, spoofing manipulation is followed by an order execution on the opposite side of the market.

The schematic ask spoofing order identification is shown in Fig. 1.1, while the bid spoofing order identification has a mirror representation.

For example, a spoofer submits a large ask order of 100 000 shares at 230 Rub (P_{sp}), while the best ask is 227 RUB (P_2), and the market price is 226.5 RUB (P_0). The market reaction of a big seller's appearance is to close the holding position or sell the stock. The market price starts to move down to 224 RUB (P_1), where the spoofer placed its small buy order for 200 units of stock. When the spoofer executes his small order, he cancels the big sell order, and the market rebounds back to the level of 226.5 RUB, where the spoofer closes his long position and earns a low-risk profit.

The spoofing manipulation strategy is created to be a low-risk strategy, however, it is not risk-free. The manipulator bears the risk that an incoming buy from another market participant hits his spoofing sell limit order against the manipulator's goal. In the example above, the spoofing sell order price must be substantially above the current ask price to avoid risking execution; however, it should be visible in the order book so as to influence other market participants. The closer the spoofing limit order to the best ask price, the higher the risk of being executed. The

[4]See IIROC Rules Notice Guidance Note 13-0053 and ESMA Final Report 2015/224.

spoofing order's lifetime is the second risky aspect for the manipulator. The quicker the spoofing order is cancelled after the small buy order execution, the less likely it is undesirably hit by the market price. A similar logic is applicable to buy spoofing orders.

The algorithm for identifying spoofing orders in the current research is based on the definition of spoofing used in the empirical analysis by Lee et al. (2013) in their article on the Korean Stock Exchange. They defined "a spoofing order as a bid/ask with a size at least twice the previous day's average order size and with an order price at least 6 ticks away from the market price, followed by an order on the opposite side of the market, and subsequently followed by the withdrawal of the first order" (Lee et al. 2013).

We define a spoofing order as a bid/ask with a size at least twice or Y-times the average order size of the five previous trading days and with an order price at least one tick away from the market price and within a specified number of ticks from the best price, followed by an execution of an order on the opposite side of the market during the lifetime of the first order, and, finally, the withdrawal of the first order. The maximum number of ticks away from the market price are parameters that we vary below.

While our spoofing definition is very close to that used in the paper by Lee et al. (2013), we stress that our data does not include any trader identification, therefore does not allow us to match, by the trader, spoofing orders with the orders that execute on the other side of the market. This was possible with the Korean data from KRX that contained information on the customer account number of the person who placed the order used by Lee et al. (2013). We define a spoofing order outside the spread in contrast to Brogaard et al. (2022) definition, as their definition describes not a general spoofing order manipulation, but an aggressive and highly risky strategy, which is called "advancing the bid" by other scholars (Putniņš 2020) and regulators[5]. However, we believe, that the lack of trading id information is not essential, as traders may hide their manipulative strategy by executing them in groups with different id numbers.

Our spoofing order identification includes three main parameters:

- number of ticks away from the market price ("order book level");
- a maximum lifetime of the order;
- minimum size of the order (volume).

We experimented with these three parameters to yield nine different algorithms. Type 0 algorithms use a maximum possible order book depth from the 1^{st} till the 500^{th} ticks away from the market price. Types 1 and 2 algorithms use the order book level that is taken based on the analysis of cumulative order numbers placed on each order book level for every stock. Algorithms' modifications A, B, and C differ in the lifetime of spoofing order depending on the stock's trading activity and its listing level. We present the details of the algorithms in Appendix 3.

In our research, we define nine algorithms $Z \in \{0-A,\ 0-B,\ 0-C,\ 1-A,\ 1-B,\ 1-C,\ 2-A,\ 2-B,\ 2-C\}$ calculated within time windows of $X \in \{10,\ 30,\ 60\}$ minutes with the requirement of spoofing order to have a minimum size of $Y \in \{2,\ 4\}$ mean of the average volume for prevailing five consecutive days, which gives us 54 modifications of the algorithms. "Mean of the average volume" is calculated by the average order volume of the previous five trading days for the specific stock.

[5] COMMISSION DELEGATED REGULATION (EU) 2016/522

The overall results are not significantly qualitatively different regardless of what algorithm we use[6]. Therefore, in the rest of the paper, we present the results using the algorithm $1 - C$ with a minimum order size of $Y \in \{2\}$ mean volume while showing other algorithms' outcomes in the Appendices.

Aggarwal & Wu (2006) demonstrate that small illiquid stocks traded on markets with little regulation are more likely to be manipulated. Lee et al. (2013) find that spoofing orders were more frequently observed in stocks with higher return volatility and lower market capitalization. Following similar logic, we aim to answer the same question: which kind of stock is more likely to be manipulated on the Moscow Exchange?

According to the data available from MOEX for 91 stocks for the first six months of 2019, the ratio of cancelled orders to all placed orders varies from 47.84% for CBOM (Credit Bank of Moscow) to 95.18% for DVEC (Far-Eastern Energy Company). On average, 83.88% of orders placed on the market were cancelled. Table 1.3 below shows the average ratios depending on the listing level.

	Orders placed	Orders cancelled	Cancelled / placed orders
1^{st} listing level	340 357 721	285 702 770	83.94%
2^{nd} listing level	24 011 239	19 814 904	82.52%
3^{rd} listing level	14 947 500	12 653 210	84.65%
Total	379 316 460	318 170 884	83.88%

TABLE 1.3: Ratio of cancelled to all placed orders on MOEX

We define "spoofing ratio" henceforth SP as a ratio of identified possible spoofing orders by the algorithm to all placed orders for the specific stock.

Table 1.4 below shows the average ratio of spoofing orders (SP ratio) to all placed orders identified by different algorithms.

	2 mean	*3 mean*	*4 mean*
Algo 0-A	4.62%	2.45%	1.32%
Algo 0-B	4.46%	2.35%	1.25%
Algo 0-C	4.58%	2.43%	1.31%
Algo 1-A	3.34%	1.65%	0.86%
Algo 1-B	3.28%	1.62%	0.84%
Algo 1-C	3.34%	1.65%	0.86%
Algo 2-A	4.40%	2.31%	1.24%
Algo 2-B	4.28%	2.25%	1.20%
Algo 2-C	4.38%	2.30%	1.24%

TABLE 1.4: SP ratio by different algorithms. Table represents SP ratio of spoofing order to all placed orders on MOEX identified by different algorithms among 91 stocks for the period of January-June 2019, grouped by the assumption for spoofing order to have minimum volume of $Y \in \{2, 3, 4\}$ times the mean volume for five previous trading days.

According to Lee et al. (2013), the shares covered by the spoofing-buy orders represented 0.81% of the total shares across all buy orders on the Korean Exchange in 2002. Our results show, on average, that the fraction of spoofing orders is three times higher (Table 1.4). However, Lee et al. showed that the ratio of executed orders

[6]Statistical analysis in Section 1.3 quantitatively shows the absence of significant differences among algorithms.

to all placed orders is 69.78%, which leaves 30.22% cancelled orders. This figure is much smaller than the 83.88% on MOEX in 2019 (Table 1.4), which is compatible with the rise of the SP ratio by the same amount. As we analyze different markets and data with 17 years gap, financial markets all over the world have developed dramatically, primarily due to high-frequency trading technologies.

Lee et al. (2013) present empirical evidence that the average size of day traders' spoofing-buy orders is approximately 4.1 times larger than that of non-spoofing buy orders by day traders in general. Based on the mentioned statistics, we include in our further analysis the 4 *mean* modification of the algorithm with the assumption of the minimum volume of the spoofing order being 4 mean order volumes for five previous trading days.

Our analysis reveals that the spoofing ratio does not vary significantly with a change of the maximum lifetime parameter in the algorithm, as 90% of spoofing orders have a lifetime of up to 5 minutes and 98% have a lifetime not greater than 30 minutes. The distribution of spoofing orders by lifetime stays consistent among algorithms: with around 45% of spoofing orders having a lifetime less than 5 seconds, approximately 25% living between 5 and 30 seconds (Appendix 4).

Lee et al. (2013) present empirical evidence that spoofing order duration is about 79 minutes; however, they argue that the low probability of execution and a desire to minimize potential regulatory penalties lead to delayed spoofing order cancellation.

Regarding the ratio of buying and selling spoofing order types, each algorithm has about a 50:50 split with slightly more selling orders (Appendix 5).

Results show that more spoofing manipulation cases took place in April, May, and June (around 20% per each month), while 10-11% in January and 15% in February and March. There are no significant differences in these numbers across algorithms (Appendix 6).

Spoofing order distribution among different stock listing levels shown in Appendix 7 indicates that 92-93% of all identified orders belong to stocks from the 1^{st} listing level, around 4.4% are orders for 2^{nd} listing level stocks and only 2.8% for 3^{rd} listing level stocks. There are no significant differences in these numbers across algorithms.

While ranking stock by the highest spoofing ratio using algo $1 - C$ with $Y = 2$ and $Y = 4$, eleven stocks[7] are the same in the top twenty, showing that some stocks are prone for being manipulated then others. We do not observe a significant difference in SP ratio between various time windows, as 90% of spoofing orders live up to 5 minutes.

The main difference in algorithms arises from the parameter of the maximum number of ticks away from the market price for the spoofing order to occur. For the stocks from the 1^{st} listing level, 90% of spoofing orders occur below the 45^{th} tick[8] away from the best price, while 70% of spoofing orders - below the 23^{rd} tick away from the best price. The data are available in Table 1.5 below.

[7] VTBR, AFLT, BANEP, RNFT, POLY, MSNG, GAZP, PIKK, RUAL, SBER, SFIN

[8] By tick we mean the minimum price increment of the stock.

1^{st} listing level stocks									
Algo	0-A	0-B	0-C	1-A	1-B	1-C	2-A	2-B	2-C
70%	23	22	23	16	17	17	22	22	22
90%	44	43	45	21	22	22	38	33	39
95%	65	63	67	23	23	23	48	47	57
2^{nd} listing level stocks									
Algo	0-A	0-B	0-C	1-A	1-B	1-C	2-A	2-B	2-C
70%	23	20	20	15	15	15	18	19	32
90%	75	50	49	21	22	22	39	42	60
95%	147	87	85	26	26	26	57	62	78
3^{rd} listing level stocks									
Algo	0-A	0-B	0-C	1-A	1-B	1-C	2-A	2-B	2-C
70%	49	38	48	25	25	25	32	33	22
90%	131	99	131	33	32	33	60	62	38
95%	220	129	224	35	36	36	78	80	49

TABLE 1.5: Spoofing order placement inside the order book. The table shows the number of ticks away from the best price for the spoofing order to occur, depending on the algorithms and stock listing level. For example, in the first line and 1^{st} column we observe 23, which shows that 70% of spoofing orders for the 1^{st} listing level stocks using Algo $0-A$ placed below 23 ticks away from the best price.

Based on the data from Table 1.5 we observe, that type 0 algorithms ($0-A$, $0-B$ and $0-C$) represent the maximum number of ticks the spoofing orders may occur. The results show that 95% of spoofing orders are placed below the 67^{th} tick away from the best price for the 1^{st} listing level stocks, while 92-93% of all identified spoofing orders belong to stocks from the 1^{st} listing level. That gives us a reason to analyse those potential spoofing orders that occur below the 75^{th} tick away from the best price rather then 500^{th}. For the 2^{nd} listing level stocks, 93% to 98% of spoofing orders depending on the lifetime modification of spoofing orders lie below 100^{th} ticks away from the best price. For the 3rd listing level stocks, 85% to 90% of spoofing orders also lie below 100^{th} tick away from the best price. Our findings do not contradict ones by Lee et al. (2013), who present empirical evidence that the vast majority (79.70%) of spoofing-buy orders submitted by day traders were more than 10 ticks away from the current best bid.

Fig. 1.2 shows an average value of SP ratio for quintiles of our 91 stocks grouped by ruble volume, from largest to smallest using traded ruble volume (January – June 2019), for the nine algorithms. Group 1 has the largest traded ruble volume, while Group 5 – has the lowest. Groups 1, 2, and 3 mainly consist of stocks from the 1^{st} listing level, while Group 4 is a mix from the 2^{nd} and 3^{rd} listing level and Group 5 is mainly 3^{rd} listing level stocks.

The figure shows that Group 1, which includes stocks with a higher traded ruble volume, has a higher average SP. The lower the ruble traded volume, the lower the average SP. It is clear from the data that there is more spoofing in large firms, which is the opposite result to that in Lee et al. (2013) and Aggarwal & Wu (2006).

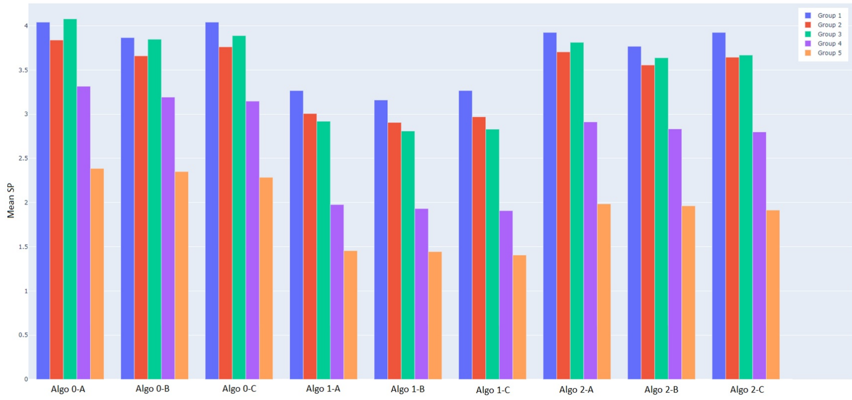


FIGURE 1.2: Average SP (marked as "Mean SP" on the figure) for groups ranked using traded ruble volume. Group 1 is the group with firms that registered the highest ruble traded volume (first bar for each algo), and Group 5 is those with the lowest (last bar in each Algo).

Conducted statistical analysis allows us to present the results using the algorithm $1-C$ in the rest of the paper while showing other algorithms' outcomes in the Appendices. We choose algorithm $1-C$ based on the assumption that a spoofing order is more likely to be placed closer to the market price while having wider boards for the order's lifetime. Algorithm $1-C$ includes only those spoofing orders that occur below 25$^{\text{th}}$, 30$^{\text{th}}$, and 40$^{\text{th}}$ ticks away from the best price for 1^{st}, 2^{nd}, and 3^{rd} listing level stock accordingly. In contrast, other types of algorithms include orders that occur deeper in the order book. We follow the logic that the closer the spoofing order to the best price, the more noticeable it is for market participants. Also, we choose the medium lifetime parameter of the algorithms. Modification C uses the following maximum lifetimes for the spoofing order: 60, 180 minutes, and all day for the 1^{st}, 2^{nd}, and 3^{rd} listing level stocks accordingly. Modification C lifetime is slightly longer than Modification B, but much shorter than Modification A. We do not identify the $1-C$ algorithm as the best for spoofing order identification, rather we choose it for presentation purposes. Appendices contain the results from analysis of all 18 algorithms[9].

[9]9 algorithms $0-A$ to $2-C$ with two modifications of the minimum order volume of 2 mean and 4 mean average order volume give us 18 algorithm modifications.

1.4 Measure of spoofing manipulation and market quality

Our measure of spoofing manipulation activity on the asset i is defined as the ratio of the number of cancelled orders identified by the algorithm to all placed orders. We call this measure SP and construct for 10-, 30- and 60-minute time windows.

Market quality measures as dependent variables used in the research are:

- $QSX_{i,t}$. Quoted spread for asset i is the time-weighted (by minute) average, over the period, of $(a_{t'} - b_{t'})/m_{t'}$, where $a_{t'}$ is the best ask, $b_{t'}$ - the best bid, $m_{t'}$ – the midquote, and t' indexes observations within $X \in \{10, 30, 60\}$ minutes (Aitken & Frino 1996).

- $ESX_{i,t}$. Effective (half) spread for asset i is the time-weighted (by minute) average of $(p_{t'} - m_{t'})/m_{t'}$, where $p_{t'}$ is the trade price and $m_{t'}$ is the prevailing midquote (prior to execution), and t' indexes observations within $X \in \{10, 30, 60\}$ minutes (Lee 1993, Blume & Goldstein 1992).

- $VOLX_{i,t}$. A measure of short-term volatility for asset i is the time-weighted (by minute) average of $(stdev_{t'})/m_{t'}$, where $stdev_{t'}$ is the standard deviation of trade prices, $m_{t'}$ – the midquote, and t' indexes observations within $X \in \{10, 30, 60\}$ minutes[10].

Brogaard et al. (2018), in their research, examine the activity of high-frequency traders (HFT) around extreme price movements and use the net imbalance metrics of HFT-NET. In their research, they look at differences in variables. They also use this metric in period t and focus on event windows that span 20 seconds algorithm after the extreme price movement interval. In our research, we follow the same logic and use differences in variables in the periods before and after the spoofing manipulation event.

As our research aims to understand the effect that SP has on market quality, we check the change in market quality measures before and after the spoofing manipulation. Figure 1.3 shows the logic of our measures on the timeline.

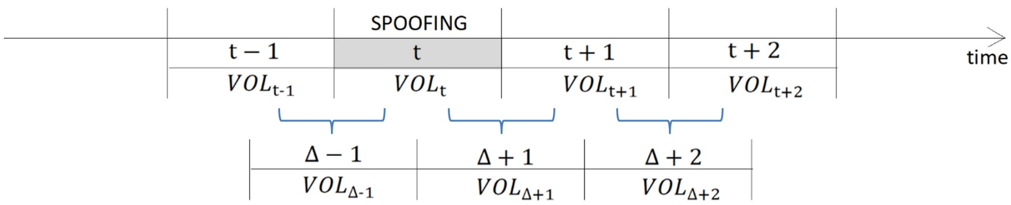


FIGURE 1.3: Time window scheme. Time windows on the timeline with spoofing manipulation activity at time t and short-term volatility (VOL_t) as an example of the market quality measure. Change in market quality measure from time $(t-1)$ and (t) is defined as time period $(\Delta - 1)$ with corresponding short-term volatility change $VOL_{\Delta-1}$.

The time window t identified in Figure 1.3 represents the time window of 10, 30, or 60 minutes where the spoofing manipulative order was in place. Time window

[10]Egginton, Ness & Ness (2016) use a short-term volatility measure Voltil as the one-minute standard deviation of trade prices. We use the same logic for our measure, but normalize the scale by midquote.

$(t-1)$ is previous to spoofing manipulative orders time window, while $(t+1)$ and $(t+2)$ are the subsequent and one lag time windows to the one with spoofing manipulative orders. $(\Delta - 1)$ as an index shows the change in variables between time t and $(t-1)$; $(\Delta+1)$ – between $(t+1)$ and t, while $(\Delta+1)$ – between $(t+2)$ and $(t+1)$. Our variables used in the analysis are following:

- $SPX_{i,t}^{Ym(Z)}$. Our measure of spoofing manipulation defined as a ratio of spoofing orders identified by the algorithm to all placed orders and subsequently cancelled within $X \in \{10, 30, 60\}$ minutes, and with the requirement of spoofing order to have a minimum volume of $Y \in \{2, 4\}$ times the mean volume observed across the 5 previous trading days for a stock i. Z is defined as a detection algorithm, $Z \in \{0-A,\ 0-B,\ 0-C,\ 1-A,\ 1-B,\ 1-C,\ 2-A,\ 2-B,\ 2-C\}$. For example, $SP10_{i,t}^{2m(0-A)}$ is a spoofing ratio for asset i identified using algorithm $0-A$ with minimum volume for the spoofing order of 2 times mean order volume and calculated within a 10-minute time window.

- The changes in market quality measures around spoofing are identified as follows:

$$\begin{aligned} VOLX_{i,\Delta-1} &= VOLX_{i,t} - VOLX_{i,t-1} \\ VOLX_{i,\Delta+1} &= VOLX_{i,t+1} - VOLX_{i,t} \\ VOLX_{i,\Delta+2} &= VOLX_{i,t+2} - VOLX_{i,t+1} \end{aligned} \tag{1.1}$$

  The other market quality measures' changes are formed the same way as in Equation 1.1.

- $VolumeX_{i,\Delta}$. Our measure of change in the trading volume for asset i within $X \in \{10, 30, 60\}$ minutes. As a control variable we use the change in trading volume ($Volume$) which is identified as follows:

$$VOLX_{i,\Delta-1} = \frac{Volume_{i,t} - Volume_{i,t-1}}{Volume_{i,t} + Volume_{i,t-1}} \tag{1.2}$$

The value of the variable defined in Equation 1.2 lies between -1 and +1, which gives us the normalized scale where the sign of the change indicates if the volume drops or rises while keeping comparability across securities. The change in trading volume for periods $(\Delta+1)$ and $(\Delta+2)$ are formed the same way as in Equation 1.2.

Table 1.6 shows means and standard deviations for the variables used in the analysis. The upper panel shows the market quality measures in levels using different time windows, while the last three rows of the table present the mean and the standard deviation of our primary variable SP.

	Time window	VOL	QS	ES	VOL_Δ	QS_Δ	ES_Δ	$Volume_\Delta$
Mean	10 min	0.035	0.140	0.078	-0.001	-0.002	-0.002	-0.077
S.D.	10 min	0.068	0.260	0.152	0.055	0.108	0.084	0.516
Mean	30 min	0.041	0.188	0.103	-0.002	-0.004	-0.003	-0.083
S.D.	30 min	0.075	0.313	0.176	0.064	0.140	0.101	0.488
Mean	60 min	0.046	0.217	0.118	-0.003	-0.011	-0.006	-0.063
S.D.	60 min	0.080	0.339	0.187	0.069	0.153	0.106	0.466

Algo $1-C$	$SP10^{2m(Z)}_{i,t}$	$SP10^{4m(Z)}_{i,t}$	$SP30^{2m(Z)}_{i,t}$	$SP30^{4m(Z)}_{i,t}$	$SP60^{2m(Z)}_{i,t}$	$SP60^{4m(Z)}_{i,t}$
Mean	2.83	0.74	2.64	0.69	2.50	0.65
S.D.	3.64	1.59	3.22	1.34	3.02	1.22

TABLE 1.6: Summary statistics of market quality variables. The table shows the mean and standard deviation for the variables in our analysis by time windows ($X \in \{10, 30, 60\}$ minutes): short-term volatility (VOL), quoted spread (QS), effective spread (ES), change in short-term volatility (VOL_Δ), change in quoted spread (QS_Δ), change in effective spread (ES_Δ), the measure of trading volume ($Volume_\Delta$), $SPX^{Ym(Z)}_{i,t}$ is a ratio of spoofing orders identified by the algorithms to all placed orders and subsequently cancelled within $X \in \{10, 30, 60\}$ minutes, and the requirement of spoofing order to have a minimum volume of $Y \in \{2, 4\}$ times the mean volume for 5 previous trading days. The statistics for all algorithms are presented in the table in Appendix 9.

Statistical analysis shows that spreads vary enormously across stocks. Aggregating the 91 most liquid stocks on MOEX, quoted spreads vary from 0.02% for the most liquid stock such as GAZP and SBER to 1.33% for USBN and even 3.01% for VJGZ as the least liquid stocks. Effective spreads vary from 0.01% to 1.51%. Aitken, Harris & Ji (2015) in their research mentioned that relative liquidity really matters to trade-based manipulation incidence and detection. The least liquid stocks are much the same everywhere and are traded under much the same conditions. The most liquid stocks are less homogeneous. Our sample shows similar results. Quoted and effective spreads are relatively stable for the stocks from the 3^{rd} listing level. However, they vary greatly for the stocks from the 1^{st} listing level. Detailed statistical data for stocks grouped by listing levels are presented in the tables in Appendix 9.

Figure 1.4 shows the means of the market quality measures grouped by ruble volume, from largest to smallest using traded ruble volume (January – June 2019). Group 1 has the largest traded ruble volume, while Group 5 – the lowest. Groups 1, 2, and 3 mainly consist of stocks from the 1^{st} listing level, while Group 4 is a mix from the 2^{nd} and 3^{rd} listing level, Group 5 – mainly 3^{rd} listing level stocks. The figure reflects the relationship between the trading volume and market quality: the higher the stock's trading volume, the better the market quality measure.

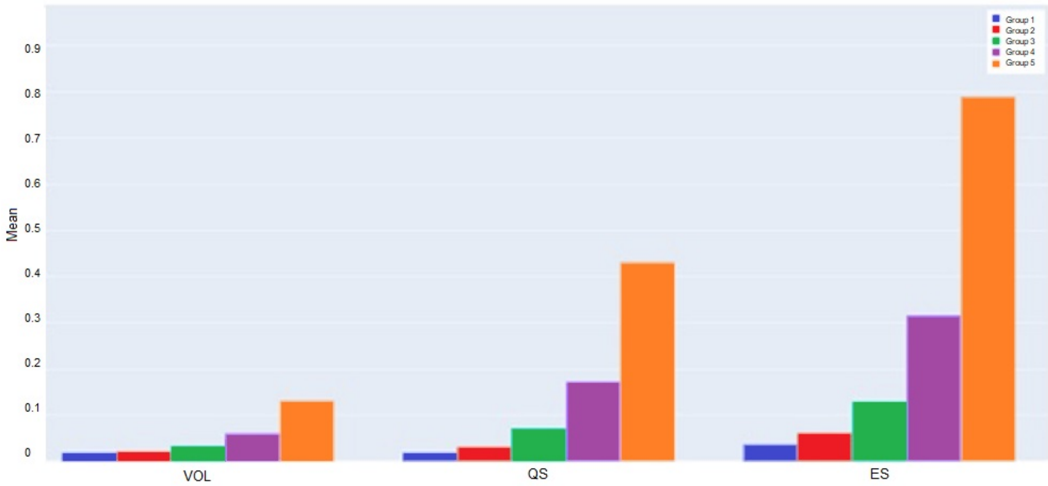


FIGURE 1.4: Mean of market quality variables for groups ranked using traded ruble volume. Group 1 is the group with firms that registered the highest ruble traded volume (first bar for each variable), and Group 5 those with the lowest (last bar in each variable).

We test dependent variables at time t for serial correlation. Firstly, we checked for the correlation between variables and their previous values. Table 1.7 shows statistical results with strong dependence.

Time window	Dependent variable	T-statistic	P-value
10 min	$VOLX_{i,\Delta}$	-224.5090	0.0000
30 min	$VOLX_{i,\Delta}$	-159.2210	0.0000
60 min	$VOLX_{i,\Delta}$	-119.8770	0.0000
10 min	$QSX_{i,\Delta}$	-153.9700	0.0000
30 min	$QSX_{i,\Delta}$	-106.8390	0.0000
60 min	$QSX_{i,\Delta}$	-86.9930	0.0000
10 min	$ESX_{i,\Delta}$	-166.6500	0.0000
30 min	$ESX_{i,\Delta}$	-127.9850	0.0000
60 min	$ESX_{i,\Delta}$	-98.4230	0.0000

TABLE 1.7: The mean-comparison test (t-test) between variables and their previous values

Also, in Figure 1.5 we plot autocorrelation of the dependent variables to check the dependence with the first ten lags, where "lag 0" has 100% correlation as it has the same values. We observe a strong correlation with the first and the second lag values for $QSX_{i,\Delta}$ and $ESX_{i,\Delta}$, while only with the first lag for variable $VOLX_{i,\Delta}$.

Moreover, we run a Breusch-Godfrey test for serial correlation in the panel models. The higher the test statistic is, the more serial correlation is left; if the test statistic is 0, there is no serial correlation. The test statistics results are presented in Figure 1.6, which shows that adding a variable with one lag decreases serial correlation significantly while adding more variables does not improve the model.

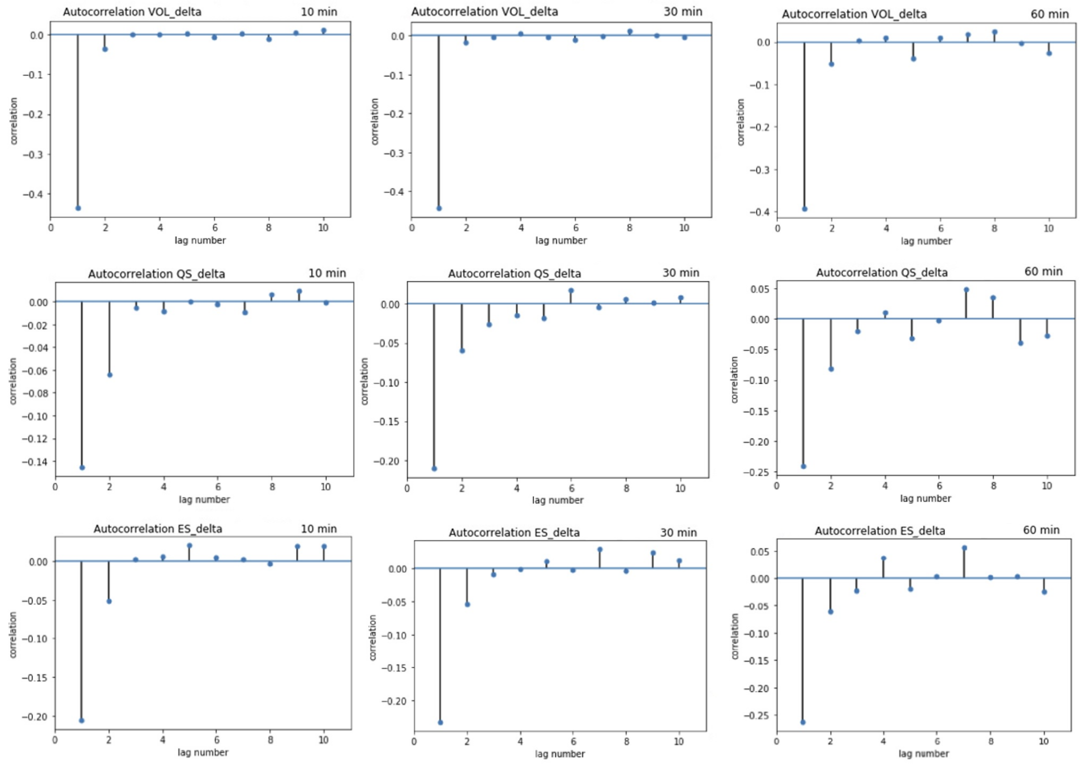


FIGURE 1.5: Autocorrelation of the dependent variables

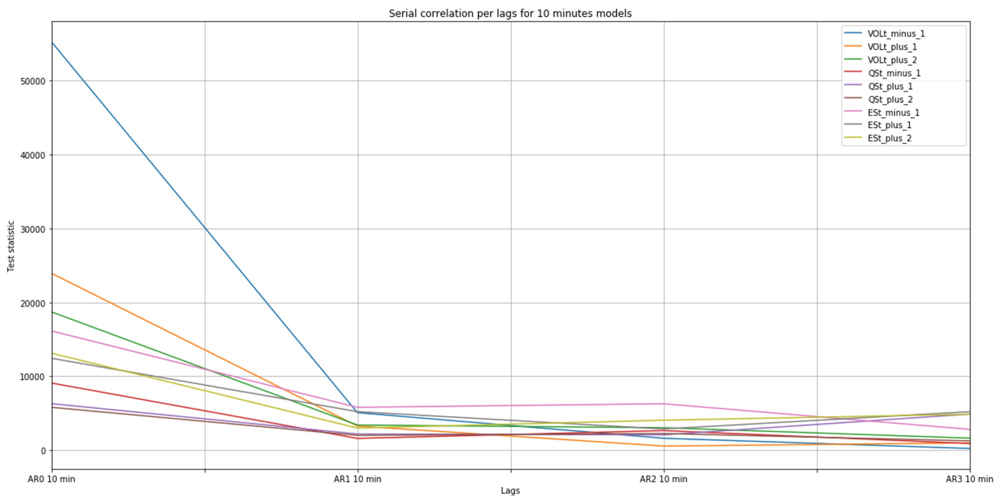


FIGURE 1.6: Breusch-Godfrey test for serial correlation in the panel models (10-minute window models). For 30- and 60-minute windows models results are presented in Appendix 15.

1.5 Methodology

The research aims to understand the effect that spoofing manipulation activities have on market quality. To explore spoofing activities and market quality, we estimate several models controlling for factors associated with market quality. Each regression uses data from the time windows before and after the spoofing manipulation period. For each algorithm, we run three types of regression for time windows of X minutes, where $X \in \{10, 30, 60\}$.

We run a dynamic panel regression with fixed effects[11], as shown below, where data are pooled across 91 stocks traded on MOEX using data for 6 months from January to June 2019.

$$VOLX_{i,\Delta-1} = \beta_0 + \beta_1 SPX^{Ym(Z)}_{i,t} + \beta_2 Aft_{i,t} + \beta_3 Ev_{i,t} + \beta_4 VolumeX_{i,\Delta-1} + \beta_5 VOLX_{i,\Delta-2} + \epsilon_{i,t}, \tag{1.3}$$

$$QSX_{i,\Delta-1} = \beta_0 + \beta_1 SPX^{Ym(Z)}_{i,t} + \beta_2 Aft_{i,t} + \beta_3 Ev_{i,t} + \beta_4 VolumeX_{i,\Delta-1} + \beta_5 VOLX_{i,\Delta-1} + \beta_6 QSX_{i,\Delta-2} + \epsilon_{i,t}, \tag{1.4}$$

$$ESX_{i,\Delta-1} = \beta_0 + \beta_1 SPX^{Ym(Z)}_{i,t} + \beta_2 Aft_{i,t} + \beta_3 Ev_{i,t} + \beta_4 VolumeX_{i,\Delta-1} + \beta_5 VOLX_{i,\Delta-1} + \beta_6 ESX_{i,\Delta-2} + \epsilon_{i,t}, \tag{1.5}$$

where $VOLX_{i,\Delta}$ is a change in short-term one-minute volatility for asset i; $QSX_{i,\Delta}$ is a change in a quoted spread for stock i; $ESX_{i,\Delta}$ is a change in an effective (half) spread for stock i; $SPX^{Ym(Z)}_{i,t}$ is a spoofing ratio calculated within the previous period. Similarly, we run regressions for the change in market quality measures for periods $(\Delta + 1)$ and $(\Delta + 2)$ after spoofing manipulation. Based on the memory tests conducted for the dependent variables in Section 1.4, we include the first lags of dependent variables to the model to make it dynamic. We include right-hand side control variables to account for stock or stock-day-specific conditions: (i) $Aft_{i,t}$ and $Ev_{i,t}$ is a set of dummy variables for three periods during the day, based on trading activity by volume (Appendix 8): afternoon (13:00 till 16:00), evening (16:00 till 18:45) respectively, or morning (10:00 till 13:00) otherwise; (ii) $VolumeX_{i,\Delta}$ is a change in trading volume within the time-window X minutes for stock i, calculated as presented in Equation 1.2; and (iii) one lag value of the dependent variable. All control variables are time-varying.

We estimate panel regressions for all algorithms, which gives us 54 panel regression results corresponding to different algorithms for every market quality measure.

Table 1.8 below shows the estimation results of the panel regression model for all 91 stocks for the 6 months from January to June 2019. For the $1 - C$ algorithm and 10- minute time window, the data include the estimated coefficient on the variable of interest ($SP10^{2m(1-C)}_{i,t}$) as well as those for the control variables. We run regressions for the change of three different dependent variables (short-term volatility, quoted spread, and effective spread), and the results are in different columns of the table. The independent variable coefficient indicates how much the dependent variable changes over time, on average per stock, when the independent variable increases by one unit.

Tables in Appendix 10 show the results of the panel regression with fixed effects for all algorithms. We present $SPX^{Ym(Z)}_{i,t}$ coefficient for all algorithms using time windows $X \in \{10, 30, 60\}$ and with the minimum volume assumptions for the spoofing order to be $Y \in \{2, 4\}$ mean of the average volume for prevailing five consecutive days. Full regression results with all coefficients are available on request.

[11] To check the choice of the fixed effects models we run the Hausman test. If p-value of test ≤ 0.05, then the fixed effects model is preferred. Tables in Appendix 1.8.14 show the results of the test.

Variables	$VOL10_{\Delta-1}$	$QS10_{\Delta-1}$	$ES10_{\Delta-1}$	$VOL10_{\Delta+1}$	$QS10_{\Delta+1}$	$ES10_{\Delta+1}$	$VOL10_{\Delta+2}$	$QS10_{\Delta+2}$	$ES10_{\Delta+2}$
$VOL10_{\Delta}$		0.384^{***}	0.405^{***}		0.341^{***}	0.340^{***}		0.372^{***}	0.389^{***}
		(0.003)	(0.002)		(0.003)	(0.002)		(0.003)	(0.002)
$SP10_{i,t}^{2m(1-C)}$	**-0.0002*****	**-0.0004*****	**-0.001*****	**0.0002*****	**0.001*****	**0.0004*****	**0.0001*****	**0.0004*****	**0.0003*****
	(0.00005)	**(0.0001)**	**(0.0001)**	**(0.00004)**	**(0.0001)**	**(0.00001)**	**(0.00004)**	**(0.0001)**	**(0.0001)**
Afternoon	0.0004	0.007^{***}	-0.00004	0.003^{***}	0.012^{***}	0.005^{***}	0.002^{***}	0.005^{***}	0.003^{***}
	(0.0002)	(0.001)	(0.0004)	(0.0003)	(0.002)	(0.001)	(0.0002)	(0.001)	(0.0005)
Evening	0.001^{***}	0.008^{***}	0.0002	0.004^{***}	0.013^{***}	0.005^{***}	0.004^{***}	0.013^{***}	0.005^{***}
	(0.002)	(0.001)	(0.0004)	(0.003)	(0.002)	(0.001)	(0.003)	(0.001)	(0.001)
$Volume10_{\Delta}$	0.020^{***}	-0.001	0.002^{***}	0.016^{***}	0.001	-0.0002	0.011^{***}	-0.008^{***}	-0.003^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.013)	(0.001)	(0.001)	(0.001)	(0.0005)
Lag variable	-0.436^{***}	-0.242^{***}	-0.276^{***}	-0.290^{***}	-0.155^{***}	-0.169^{***}	-0.252^{***}	-0.098^{***}	-0.196^{***}
	(0.008)	(0.018)	(0.015)	(0.010)	(0.013)	(0.020)	(0.013)	(0.013)	(0.020)
Observations	363,621	363,621	363,621	363,621	363,621	363,621	363,621	363,621	363,621
Adjusted R2	0.141	0.110	0.162	0.175	0.074	0.103	0.095	0.054	0.113
F Statistic	$11,937.920^{***}$	$7,515.449^{***}$	$11,772.690^{***}$	$15,408.870^{***}$	$4,832.157^{***}$	$6,970.292^{***}$	$7,620.080^{***}$	$3,505.171^{***}$	$7,738.014^{***}$
	(df = 5; 363525)	(df = 6; 363524)	(df = 6; 363524)	(df = 5; 363525)	(df = 6; 363524)	(df = 6; 363524)	(df = 5; 363525)	(df = 6; 363524)	(df = 6; 363524)
note:									$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

TABLE 1.8: The effect of SP on market quality: baseline results. Regression coefficients for the panel regressions (Equation 1.3, 1.4, 1.5) on the change in short-term volatility ($VOL10_{\Delta}$), quoted spread ($QS10_{\Delta}$) and effective spread ($ES10_{\Delta}$). $SP10_{i,t}^{2m(1-C)}$ represents a ratio of spoofing orders for the asset i identified by the algorithm $1 - C$ described in Section 1.3 to all placed orders and subsequently cancelled within 10-minute time window, and with *2 mean* of the average minimum volume for prevailing five consecutive days. As control variables we include dummies for three periods during the day: morning, afternoon, and evening; and $Volume10_{\Delta}$ that represents the change in trading volume described by Eq.(2), and one lag value of the dependent variable. All variables are standardized, and the panel estimation clusters errors by asset id and time (day-time window). Below each coefficient we show the number of observation, the standard errors in parenthesis, and the regression's adjusted R2. 'F statistic' row shows the F-test results whether the model is significant or not (it should be significant if it explains dependent variable), stars at the right of F-statistic mean that the model is significant; 'df' in parenthesis shows degrees of freedom in the model. Significance levels are denoted by $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

1.6 Results

1.6.1 Main results and economic significance

The key results show the negative relationship between spoofing activity and the change in market quality at the current manipulation period while demonstrating a positive relationship after the manipulation period. Here we observe that spoofing causes a short-term misleading increase in market quality, followed by the more extended period of market quality decrease after spoofing order cancellation.

Our results show that the increase of our measure of spoofing manipulation SP tends to be associated with the rise in the change in quoted and effective spreads and short-term volatility (significant at 1% level, p-value < 0.01) for stocks traded on Moscow Exchange after spoofing manipulation periods. More spoofing leads to faster-growing spreads or increases in a change in spreads after the cancellation of spoofing orders. We can interpret it the other way: after high spoofing activity, spreads tend to increase quicker in the next period. When the volatility is growing, spoofing is more likely, or higher spoofing tends to increase the volatility in the period just after spoofing. Moreover, this persists for two periods, subsequently after the spoofing.

For illustration, when someone is spoofing, she may place a large ask order on one side, which she wants to execute. She also may put smaller ask orders and moves them in front of each other to show that many traders are competing. She then puts smaller bid orders on the other side. Other players will also join the market and place more competitive ask orders, thus resulting in a tightening of the spread. Our results show that spoofing leads to a short-term decrease in spreads and volatility in the current period. If the spread is tight, the price does not have too much room to move, so volatility falls during the spoofing compared with the pre-spoofing period.

The negative consequences occur right after the spoofing or if the large order on the ask side is suddenly removed. While the price was compressed during the spoofing between the small bid order and the large spoofing order, the market quality showed a temporary deceptive better state. Suddenly, the large order is removed, and the price continues upward move with higher volatility and wider spread, which destabilizes the market and worsens its quality.

We run an additional set of regressions as shown below to check the cumulative effect on market quality across the period from (t) to $(t+2)$ including the change inside those periods to follow the rise and fall of market quality measures around the spoofing manipulation event:

$$VOLX\_T_i = \beta_0 + \beta_1 SPX_{i,t}^{Ym(Z)} + \beta_2 Aft_{i,t} + \beta_3 Ev_{i,t} + \beta_4 VolumeX_{i,\Delta-1} + \beta_5 VOLX_{i,\Delta-2} + \epsilon_{i,t}, \tag{1.6}$$

$$QSX\_T_i = \beta_0 + \beta_1 SPX_{i,t}^{Ym(Z)} + \beta_2 Aft_{i,t} + \beta_3 Ev_{i,t} + \beta_4 VolumeX_{i,\Delta-1} + \beta_5 VOLX_{i,\Delta-1} + \beta_6 QSX_{i,\Delta-2} + \epsilon_{i,t}, \tag{1.7}$$

$$ESX\_T_i = \beta_0 + \beta_1 SPX_{i,t}^{Ym(Z)} + \beta_2 Aft_{i,t} + \beta_3 Ev_{i,t} + \beta_4 VolumeX_{i,\Delta-1} + \beta_5 VOLX_{i,\Delta-1} + \beta_6 ESX_{i,\Delta-2} + \epsilon_{i,t}, \tag{1.8}$$

where $VOLX\_T_i = VOLX\_T_{i,t+2} - VOLX\_T_{i,t}$. We use the same logic for QS and ES. The results shown in Appendix 19 give us positive coefficients of the main variable of our interest SP and tell us that spoofing manipulation increases all market quality measures making the market more volatile with wider spreads.

To illustrate the economic significance of the obtained results, we take five stocks from each listing level and calculate the SP ratio using algorithms $1-C$ with 10-, 30- and 60-minute windows and $Y = 2$ ($SP10_{i,t}^{2m(1-C)}$, $SP30_{i,t}^{2m(1-C)}$, $SP60_{i,t}^{2m(1-C)}$).

As 93% of all identified spoofing orders belong to stocks from the 1^{st} listing level and 90% of these orders live a maximum of 5 minutes, the choice of algorithm $1-C$ with 10 minutes windows is reasonable for analyzing the economic significance effect for the 1^{st} listing level stocks. However, we include 30- and 60-minute time windows modification of the algorithm for comparison and check for the 2^{nd} and 3^{rd} listing level stocks.

To analyze the change of variables before and after the spoofing manipulation period, we assume that spoofing manipulation occurs at time t and SP ratio equals its mean; VOL, QS and ES at the time $(t-1)$ also equal their mean values. Then we compare the value of variables if SP ratio is one standard deviation above the mean. The obtained results are shown in Table 1.9 and Appendix 11.

We are interested in the time window after the manipulation as it shows us the effect spoofing has on the market after the cancellation of manipulative orders. The derived examples show the significant change in the market quality after the spoofing manipulation period. Table 1.9 for SBER (Sberbank) shows if SP value is one standard deviation above the mean, quoted and effective spreads tend to rise 5.64% and 4.44% accordingly in their value in the period just after the spoofing manipulation. For example, quoted spread for SBER was 0.02% during the period $(t+1)$. If SP ratio rises by one standard deviation, the quoted spread for SBER take its new value of 0.02%*(1+5.64%)=0.021%. Here we measure the percentage change in the variables measured in percentage points.

We observe a weaker relationship between spoofing and volatility. The results show that quoted and effective spreads have an economically significant change in their values in the next period after manipulation with the rise of spoofing manipulation. For the 1^{st} listing levels stocks QS and ES rise on average by 7% if SP rises by one standard deviation above its mean in the ten-minute window algorithms. The same happens two periods after the spoofing: although the relationship is not so strong, it remains positive, which shows that even two periods after the spoofing order cancellation, the volatility and spreads tend to rise on average by 2.5%.

The economic significance is slightly smaller for less liquid stocks, however not uniform, e.g. OGKB and BANEP stocks have similar results as 1^{st} listing level stocks, e. g., 3% and 5% respectively for $QS_{\Delta+1}$. For the 30- and 60-minute time windows modifications of the algorithm, we obtain similar results with higher economic impact for the 1^{st} listing level stocks and lower for the others. Results are presented in Appendix 11. For the 3^{rd} listing level stocks except for BANEP, the numbers show less than 1% impact, which is comparatively low. In showing examples of economic significance, we also demonstrate different spoofing manipulation effects on market quality for various stock groups.

10 min time window	1st listing level stocks					2nd listing level stocks					3rd listing level stocks				
	SBER	GMKN	LKOH	MGNT	MOEX	OGKB	APTK	DVEC	KMAZ	MRKU	BANEP	MOBB	BLNG	LSNG	LSNGP
$SP30^{2m(1-C)}_{(i,t)}$ mean	4.33	3.95	2.53	3.08	3.24	3.41	5.29	1.10	0.85	0.47	4.05	2.74	1.32	0.35	2.51
$SP30^{2m(1-C)}_{(i,t)}$ SD	1.26	2.69	1.96	2.21	2.82	3.80	5.81	2.18	2.03	1.72	3.75	4.68	2.47	1.42	3.04
Change in $QS_{\Delta-1}$	-3.10%	-3.45%	-3.40%	-2.73%	-4.78%	-1.26%	-0.82%	-0.08%	-0.12%	-0.12%	-2.33%	-0.21%	-0.12%	-0.11%	-0.89%
Change in $QS_{\Delta+1}$	5.64%	7.32%	7.37%	6.03%	10.03%	3.03%	1.99%	0.20%	0.31%	0.31%	5.36%	0.52%	0.30%	0.27%	2.17%
Change in $QS_{\Delta+2}$	2.55%	3.13%	3.13%	2.54%	4.31%	1.23%	0.80%	0.08%	0.12%	0.12%	2.22%	0.21%	0.12%	0.11%	0.88%
Total change across periods (t) to ($t+2$)	8.34%	10.69%	10.73%	8.72%	14.77%	4.30%	2.81%	0.28%	0.44%	0.44%	7.71%	0.73%	0.42%	0.37%	3.07%
Change in $ES_{\Delta-1}$	-23.40%	-20.63%	-20.18%	-15.45%	-27.93%	-6.09%	-3.75%	-0.36%	-0.53%	-0.54%	-12.33%	-0.93%	-0.52%	-0.44%	-3.90%
Change in $ES_{\Delta+1}$	4.40%	5.79%	5.92%	4.75%	7.71%	2.26%	1.43%	0.14%	0.26%	0.21%	4.16%	0.37%	0.21%	0.18%	1.49%
Change in $ES_{\Delta+2}$	3.43%	4.44%	4.52%	3.62%	5.91%	1.71%	1.08%	0.11%	0.10%	0.16%	3.15%	0.28%	0.16%	0.13%	1.12%
Total change across periods (t) to ($t+2$)	7.98%	10.49%	10.71%	8.55%	14.08%	4.01%	2.52%	0.25%	0.37%	0.38%	7.44%	0.64%	0.37%	0.31%	2.63%
Change in $VOL_{\Delta-1}$	-1.31%	-3.19%	-2.56%	-2.11%	-3.66%	-2.38%	-1.78%	-0.34%	-0.16%	-0.38%	-4.41%	-0.60%	-0.34%	-0.29%	-1.97%
Change in $VOL_{\Delta+1}$	1.20%	2.92%	2.40%	2.00%	3.37%	2.29%	1.72%	0.33%	0.32%	0.38%	4.03%	0.60%	0.33%	0.29%	1.91%
Change in $VOL_{\Delta+2}$	0.61%	1.49%	1.22%	1.01%	1.72%	1.16%	0.87%	0.17%	0.16%	0.19%	2.06%	0.30%	0.17%	0.14%	0.96%
Total change across periods (t) to ($t+2$)	1.82%	4.45%	3.65%	3.03%	5.15%	3.47%	2.60%	0.50%	0.48%	0.56%	6.17%	0.90%	0.50%	0.43%	2.89%

TABLE 1.9: Example of economic significance of the change in volatility, quoted and effective spreads with the change in spoofing ratio for 15 stocks from different listing levels. An example of SP ratio using algorithm $1-C$ with 10-minute windows and $Y = 2\ (SP10^{2m(1-C)}_{(i,t)})$. Examples with 30- and 60-minute windows are presented in Appendix 11.

1.6.2 Regression tests

Multicollinearity

We use Variance Inflation Factor (VIF) to test for collinearity. VIF measures linear dependencies between independent variables used in the model. To estimate, we take the following steps:

1. We have regression in form $y = \alpha_0 + \alpha_1 * x_1 + \alpha_2 * x_2 + \epsilon$. To find VIF for each variable, we build linear regression for said variable with another variable being used to predict the first one. $x_1 = \alpha_0 + \alpha_2 * x_2 + \epsilon$ and $x_2 = \alpha_0 + \alpha_1 * x_1 + \epsilon$.
2. We derive VIF for variables using the following formula VIF= $\dfrac{1}{1 - R^2}$ with R-squared derived from each regression in step 1.
3. If all VIF coefficients for all variables are less than 5, there is no multicollinearity in the model.

The results show that all 54 algorithms have VIF values below 5, which indicates no collinearity in the models.

Heteroskedasticity

We use the Goldfeld-Quandt test for heteroskedasticity with a 5% significance level. The null hypothesis of the test is homoskedasticity. The results of the test are presented in Appendix 16 and show that regressions for almost all algorithm modifications do not have heteroskedasticity in the models. Only models for $VOL60_{\Delta-1}$ with a 60-minute time window are heteroskedastic as p-values are less than 5%.

Unit roots/stationarity

Covariance stationarity requires that the unconditional first two moments of a stochastic process (mean and variance) do not change over time. If the data generating process is non-stationary, and we ignore that fact, we could have a spurious regression that will imply that our estimated coefficients are inconsistent. So, when the panel's time dimension is not small, it is essential to verify its stationarity. We test our dependent variables in the models for stationarity using the Pesaran-Shin (IPS) test with the null hypothesis that all panels contain unit roots. The results of tests conducted for all models show that the data generating processes are stationary. We report the p-values in Appendix 18.

1.6.3 Robustness

As in the current research we have analysed not only one model, but 18 modifications of the model depending on the algorithm's parameters, such as minimum spoofing order size, maximum price level in the order book for the spoofing order to occur and different values for the lifetime of the spoofing order. We run 54 panel regressions, and they show no significant difference in the results and overall conclusions. These algorithm modifications allow us to check whether our results are robust to our particular definition of the SP measure as a metric of spoofing manipulation without any substantive difference. In none of these cases, the results change in any economically significant way.

1.6.4 Endogeneity

In our research, we check the endogeneity of the SP variable to understand how spoofing drives market quality. In our statistical model, SP could be changed or determined by its relationship with other variables. At the same time, we determine spoofing as an intraday activity that could be subject to common shocks. So there is a possibility that market quality affects SP and/or SP may be driving market quality. To address this question, if endogeneity is distorting our inference in a significant way, we identify an instrumental variable for SP, which is $MeanSP$, described below.

Correlation analysis

First, we run a correlation analysis on the SP ratio to determine whether there is a linear correlation between the SP ratio in one stock with the SP ratio in another stock. We get matrices of 91×91 for each of the nine algorithms with three different time windows (all matrices are available on request). The results show some positive correlation found for 1^{st} listing level stocks for 10-minute time window algorithms, but not very obvious. For 30-minute time window algorithms, the test shows some positive correlation between SP ratio for less liquid stocks form 3^{rd} listing level (Appendix 12). For a 60-minute time window, the is some positive correlation between the ten least liquid stocks from the 3^{rd} listing level; however, the average correlation coefficient for them is about 0.4 (Appendix 12). Overall, our correlation analysis shows that there is very little evidence of the correlation of spoofing manipulation activities among 91 stocks on MOEX. Furthermore, no correlation was found between SP ratios for 1^{st} listing level stocks and those for 3^{rd} listing level stocks in any models.

Instrumental variable

Secondly, we construct an instrumental variable following the approach of Hasbrouck & Saar (2013). We follow the same logic: spoofing manipulation activity that may be correlated across stocks as showed in correlation analysis above, but that the effect of SP on the market quality of one particular stock should be unaffected by the incidence of SP in another stock. Therefore, we instrument SP in a particular stock using SP in unrelated stocks. To construct the instrumental variable (IV) in a particular stock i we exclude this stock from the sample and average contemporaneous SP for all other stocks.

As a condition for a good instrument is that the instrument is correlated with the variables being instrumented, we run a series of the following regressions:

$$SPX_{i,t}^{Ym(Z)} = \beta_0 + \beta_1 SPX_{i,t}^{Ym(Z)\sim} + \beta_2 Aft_{i,t} + \beta_3 Ev_{i,t} + \beta_4 VolumeX_{i,\Delta-1} + \beta_5 VOLX_{i,\Delta-2} + \epsilon_{i,t}, \tag{1.9}$$

$$SPX_{i,t}^{Ym(Z)} = \beta_0 + \beta_1 SPX_{i,t}^{Ym(Z)\sim} + \beta_2 Aft_{i,t} + \beta_3 Ev_{i,t} + \beta_4 VolumeX_{i,\Delta-1} + \beta_5 VOLX_{i,\Delta-1} + \beta_6 QSX_{i,\Delta-2} + \epsilon_{i,t}, \tag{1.10}$$

$$SPX_{i,t}^{Ym(Z)} = \beta_0 + \beta_1 SPX_{i,t}^{Ym(Z)\sim} + \beta_2 Aft_{i,t} + \beta_3 Ev_{i,t} + \beta_4 VolumeX_{i,\Delta-1} + \beta_5 VOLX_{i,\Delta-1} + \beta_6 ESX_{i,\Delta-2} + \epsilon_{i,t}, \tag{1.11}$$

where $SPX_{i,t}^{Ym(Z)\sim}$ represents an average SP (or MeanSP) in other stocks. Similarly, we run regressions for periods $(\Delta + 1)$ and $(\Delta + 2)$ after spoofing manipulation. The regression results show that $MeanSP$ is correlated with SP as p-values of the regression coefficients are significant for all models and time windows (Appendix 20), meaning that the choice of the IV is appropriate.

Tables in Appendix 13 show the results obtained using the IV for six months in 2019. We find that the IV results are stronger than the benchmark result. All coefficients in the IV are significant; the IV has the same sign and is larger in magnitude than those in the benchmark analysis. These findings support our hypothesis of a negative relationship between SP and market quality after spoofing manipulation.

We also run exogeneity tests for SP using Sargan-Hansen J-statistic. We build a model with SP treated as an endogenous variable and then as an exogenous variable. The Null hypothesis: SP is exogenous, so p-value < 0.05 is required to reject it. In general, models with $SPX_{i,t}^{Ym(Z)\sim}$ ($MeanSP$) being used as instrumental variable show valuable results as SP is always significant, has correct signs, and is exogenous in most cases. Figures marked grey in Appendix 17 show models with exogenous variables. So, we find the evidence for the need to instrument SP in the benchmark analysis. Using models with $MeanSP$ as an instrumental variable gives us better results than benchmark models as coefficients are always significant, higher in value, and mostly exogenous.

1.7 Conclusion

This paper analyzes a microstructure-based manipulation strategy, spoofing, its determination and its effect on market quality in stocks on the Moscow Exchange (MOEX).

Our finding shows that the ratio of cancelled orders to all orders placed on MOEX is, on average, 83.88%. Our measure of possible spoofing manipulation activity is defined as the "spoofing ratio" or SP, which is the ratio of identified possible spoofing orders to all placed orders for the specific stock. SP ratio varies from 0.84% to 4.6% for different spoofing identification algorithms.

Our research investigates the relationship between the liquidity of the security and spoofing manipulation in Section 1.3, and shows that more liquid stocks are more prone to be manipulated, which contradicts previous research by Huang & Cheng (2015), who argue that more illiquid stocks are more likely to be manipulated. Also, Aitken et al. (2009) show that ramping manipulation is most likely in moderate liquidity stocks where surveillance detection is less likely than in illiquid stocks, and yet capital requirements for a successful manipulation are somewhat reduced relative to the most liquid stocks. To compare results, we must consider the differences in the surveillance systems, trading volumes, and market liquidity characteristics on NYSE and MOEX exchanges.

Our key results show the negative relationship between spoofing activity and the change in market quality for the period just before the manipulation while demonstrating a positive relationship after the manipulation. So, our measure of spoofing manipulation SP tends to be associated with the decrease in the change in quoted and effective spreads and short-term volatility just before spoofing manipulation periods. Here we observe that spoofing causes a short-term increase in market quality, however, followed by a period of market quality distortion after spoofing order cancellation.

The results show that the increase in our measure of spoofing manipulation tends to be associated with a rise in the change in quoted and effective spreads and short-term volatility (significant at 1% level, p-value < 0.01) for stocks traded on Moscow Exchange after spoofing manipulation periods. More spoofing leads to faster-growing spreads just after the spoofing order cancellation. To put it another way, spreads tend to increase more quickly in the period after spoofing. Moreover, this effect persists for two periods subsequent to the spoofing.

Relationships in all 54 panel regressions show a similar result. The effect identified is robust to different specifications and several modifications of the spoofing ratio. Our results hold after controlling for volatility, day trading volume, and periods of intensive trading during the day.

We contribute to the existing literature in several layers. First, we develop a broad spoofing identification algorithm using intraday time dimensions according to the nature of the spoofing strategy. We widen the spoofing definition to allow manipulative orders to happen outside the spread and introduce spoofing intensity measure SP. Our approach could be further applied to different asset classes and exchanges.

Secondly, we are the first to show a rebound effect of spoofing manipulation on market quality. Specifically, our measure of spoofing manipulation intensity, SP, tends to be associated with a decrease in the change in quoted and effective spreads and short-term volatility just before spoofing manipulation periods. However, higher SP tends to be associated with a significant rise in market quality measures after spoofing manipulation. We find that manipulation creates a fake or temporary perception in the market and can subsequently make traders less confident about the level of the actual asset price. The change in market quality conditions leads to an unstable trading environment, and we find that spoofing orders have a destabilizing effect on market quality. More spoofing leads to rising spreads and volatility in the next period. This effect is economically significant and robust to different specifications, endogeneity tests, and alternative modifications of SP. Our results hold after controlling for volatility, day trading volume, and periods of intensive trading during the day. So, our study suggests that intense spoofing activity is associated with degraded market quality.

The limitation of our study could be seen from the perspective of the Russian stock market data specificity. We address the question by increasing the variety of identification variables in spoofing identification methodology and then showing the robustness of the results. We show how to identify spoofing depending on stock liquidity characteristics, which could be further used as an approach to analyse the market rather than a straightforward determination of variables. This is why we do not limit spoofing identification by keeping only the high-frequency orders that live less than a minute, as we aim to show the robustness of our key market quality finding for less liquid stocks, where spoofing orders might leave longer.

The research approach used in the paper can be employed in other markets and with other data to analyze spoofing and to compare whether it has the same effect in other stock markets.

1.8 Appendices

1.8.1 Appendix 1. Listing level requirements on MOEX

№	Requirement	1st list level	2nd list level	3rd list
1	Number of stocks in free float (FFs) and their total market value (FFC)	If the market capitalization is $>$ 60 billion rubles, then for common stocks and for preferred stocks FFs $\geq$ 10% If the market capitalization is $\leq$ 60 billion rubles, then for common stocks and for preferred stocks (preferred stocks of a certain type) FFs $\geq$ FF – calculated according to the formula where $FF = (0.25789 - 0.00263 * Cap) * 100\%$, Cap – issuer's market capitalization in billion rubles Common stocks FFC $\geq$ 3 billion rubles, from all issued common stocks Preferred stocks (preferred stocks of a certain type) FFC $\geq$ 1 billion rubles, from all issued preferred stocks (preferred stocks of a certain type)	Common stocks when stocks are included in the Second level or are transferred from the Third level to the Second level (except for cases when they are included in the Growth Sector) FFC $\geq$ 1 billion rubles, FFs $\geq$ 10% of all issued common stocks. when stocks are transferred to the Second level from the First level FFs $\geq$ 4% of all issued common stocks Preferred stocks when stocks are included in the Second level or are transferred from the Third level to the Second level (except for cases when they are included in the Growth Sector) FFC $\geq$ 500 million rubles, FFs $\geq$ 10% of all issued preferred stocks (preferred stocks of a certain type) when stocks are transferred to the Second level from the First level FFs $\geq$ 4% of all issued preferred stocks (preferred stocks of a certain type)	None
2	Issuer's period of existence	Not less than 3 years	At least 1 year, or at least 1 month, if the issuer has control over the company (subsidiary), which period of existence is at least 1 year, provided that the share of the business (businesses) of such a company according to the consolidated financial statements is at least 50% of the total business of the group of which the issuer is a member	None
3	Preparation and disclosure (publication) of financial statements in accordance with IFRS or other internationally recognized standards	For 3 completed years preceding the date of the stock's inclusion in the First level.	For 1 completed year preceding the date of the stock's inclusion in the Second level.	None
4	Information disclosure	The issuer has made a commitment to disclose throughout the time the stocks are on the quotation list information in the manner and amount established by the rules (requirements) approved by the Exchange.	No conditions	

Continued on next page

Maintenance requirements				
№	Requirement	1st list level	2nd list level	3rd list
5	Daily median transaction volume for each calendar quarter	Not less than 3 million rubles. and the number of trading days in which transactions were made is at least 70% of the number of all trading days in the corresponding quarter, or: 1) the daily median volume of transactions for each calendar quarter is not less than 1 million rubles, while the number of trading days in which transactions made must be - at least 70% of the number of all trading days in the corresponding quarter; 2) agreements are concluded, stipulated by clause 11 of the table and the obligations of the market-maker with respect to securities are fulfilled	At least 500 thousand rubles. and the number of trading days in which transactions were made is at least 70% of the number of all trading days in the corresponding quarter. The requirement for the daily median volume does not apply if contracts are entered into as provided for in clause 11 of the table and the obligations of the market maker regarding securities are fulfilled	None
6	Contract for the provision of services of market-maker	The presence of 2 agreements concluded between the Organization, market-makers and the Exchange and the fulfillment of market-maker obligations with respect to securities. The presence of 2 contracts is not required if the daily median volume of transactions for each calendar quarter is at least 3 million rubles. and the condition for the number of trading days on which transactions were made, indicated in paragraph 10 of the table, is observed.	The presence of 2 agreements concluded between the Organization, market-makers and the Exchange and the fulfillment of market maker obligations with respect to securities. The presence of 2 contracts is not required if the daily median volume of transactions for each calendar quarter is not less than 500 thousand rubles. and the condition is observed for the number of trading days in which transactions were made, indicated in paragraph 10 of the table	None

TABLE 1.10: Listing level requirements on MOEX

1.8.2 Appendix 2. The list of chosen stocks

1st listing level stocks: 'SBER', 'GAZP', 'GMKN', 'AFLT', 'AFKS', 'ALRS', 'BSPB', 'CBOM', 'CHMF', 'DSKY', 'ENRU', 'FEES', 'HYDR', 'IRAO', 'LKOH', 'MAGN', 'MGNT', 'MOEX', 'MSNG', 'MTLR', 'MTLRP', 'MTSS', 'MVID', 'NLMK', 'NVTK', 'PIKK', 'PLZL', 'POLY', 'ROSN', 'RSTI', 'RSTIP', 'RTKM', 'RTKMP', 'RUAL', 'SBERP', 'TATN', 'TATNP', 'TGKA', 'TRMK', 'UPRO', 'VTBR', 'YNDX', 'AKRN', 'LSRG', 'PHOR', 'RNFT', 'TRNFP'.

2nd listing level stocks: 'RASP', 'OGKB', 'APTK', 'DVEC', 'MRKC', 'MRKP', 'MRKS', 'MRKV', 'MRKZ', 'MSRS', 'SNGS', 'SNGSP', 'VSMO', 'MSTT', 'KMAZ', 'FTRE', 'MRKU'.

3rd listing level stocks: 'BANEP', 'MOBB', 'BLNG', 'LSNG', 'LSNGP', 'MRKY', 'NMTP', 'RGSS', 'SIBN', 'UNAC', 'SFIN', 'BANE', 'FESH', 'GCHE', 'IRGZ', 'IRKT', 'KBTK', 'MFON', 'NKNC', 'NKNCP', 'URKA', 'UWGN', 'VJGZ', 'PRTK', 'TGKD', 'TRCN', 'USBN'.

1.8.3 Appendix 3. Modifications of spoofing identification algorithm

"Order book level" types

Type 0 of the order book level uses the maximum possible order book depth from the 1^{st} till the 500^{th} ticks away from the market price. Best bid or best ask orders are placed on level "0" and excluded from the analysis.

Types 1 and 2 of the order book level are taken based on the analysis of cumulative volumes (CV), distributed from the 1^{st} till the 500^{th} order book levels.

Table 1.11 shows that 70% of the cumulative volume of the placed orders lies around 25^{th}, 30^{th} and 40^{th} ticks away from the market price for the 1^{th}, 2^{nd} and 3^{rd} listing level stocks respectively. We call it Type 1 "Order book level". 90% of cumulative order volume lies around 75^{th} and 100^{th} levels (Type 2).

		1^{st} listing level	2^{nd} listing level	3^{rd} listing level
Type 0:	maximum	500^{th} level	500^{th} level	500^{th} level
Type 1:	70% of CV	25^{th} level	30^{th} level	40^{th} level
Type 2:	90% of CV	75^{th} level	100^{th} level	100^{th} level

TABLE 1.11: Type of spoofing algorithms depending on the order book level. Table shows the typology of the spoofing identification algorithms according to the maximum level in the order book, where the spoofing order is placed.

"Lifetime" modifications

Modification A is represented in Table 1.12, starting from 30 minutes as a maximum lifetime of spoofing orders for the most liquid stocks.

Modification B is represented in Table 1.12 as a stock-specific lifetime, based on the same number of orders to be placed during the life of the spoofing order. 2000 orders were chosen to be set during the lifetime for the first 54 out of 91 stocks ranked by an average number of orders placed in a minute, and 1500 orders were chosen the rest 37 stocks, as they have less than 10 orders placed in a minute on average.

Company	Lifetime Modification A (min)	Lifetime Modification B (min)	Company	Lifetime Modification A (min)	Lifetime Modification B (min)	Company	Lifetime Modification A (min)	Lifetime Modification B (min)
SBER	30	5	FEES	180	32	IRKT	all day	209
GAZP	30	5	TATNP	300	35	AKRN	all day	218
SBERP	30	7	FTRE	300	36	MRKV	all day	225
LKOH	30	8	TRMK	300	42	BSPB	all day	225
GMKN	30	8	SFIN	300	47	MSTT	all day	236
POLY	30	10	RSTI	300	48	NKNC	all day	277
MOEX	30	10	PHOR	300	56	MRKZ	all day	281
MGNT	30	10	UPRO	300	56	RTKMP	all day	288
ALRS	60	13	TRNFP	300	63	MRKC	all day	294
ROSN	60	13	MVID	300	65	RSTIP	all day	332
YNDX	60	13	CBOM	all day	89	DVEC	all day	338
TATN	60	14	OGKB	all day	94	BLNG	all day	344
VTBR	60	14	MTLRP	all day	94	MOBB	all day	359
PLZL	60	15	MRKS	all day	102	UNAC	all day	377
MTLR	60	15	LSRG	all day	122	KMAZ	all day	407
NLMK	60	16	URKA	all day	129	LSNG	all day	426
SNGS	60	18	DSKY	all day	143	NMTP	all day	428
NVTK	60	18	BANEP	all day	161	FESH	all day	488
MAGN	60	20	MFON	all day	169	IRGZ	all day	492
MTSS	60	20	GCHE	all day	190	VJGZ	all day	all day
CHMF	60	21	MRKP	all day	194	MSRS	all day	all day
HYDR	60	26	ENRU	all day	203	MRKU	all day	all day
AFLT	180	26	MSNG	all day	203	TGKD	all day	all day
RASP	180	27	NKNCP	all day	174	KBTK	all day	all day
SIBN	180	27	MRKY	all day	175	PRTK	all day	all day
IRAO	180	28	UWGN	all day	186	VSMO	all day	all day
AFKS	180	30	LSNGP	all day	191	USBN	all day	all day
RTKM	180	30	PIKK	all day	193	RGSS	all day	all day
SNGSP	180	31	RNFT	all day	195	TRCN	all day	all day
RUAL	180	31	BANE	all day	197			
TGKA	180	32	APTK	all day	199			

TABLE 1.12: Modifications of spoofing order identification algorithm according to orders' lifetime

Modification C uses the following maximum lifetimes: 60 minutes for the 1st listing level stocks, 180 minutes for the 2nd listing level stocks, and the whole trading day for the 3rd listing level stocks. Different combinations of the algorithm's main components for spoofing order identification give us nine algorithms, presented in Table 1.13.

	Lifetime modifications		
Order book level types	**Modification A**	**Modification B**	**Modification C**
Type 0	Algo 0-A	Algo 0-B	Algo 0-C
Type 1	Algo 1-A	Algo 1-B	Algo 1-C
Type 2	Algo 2-A	Algo 2-B	Algo 2-C

TABLE 1.13: Spoofing identification algorithms

1.8.4 Appendix 4. Spoofing orders' lifetime distribution

Algo	**Lifetime**	**0-5 sec**	**5-30 sec**	**30-60 sec**	**1-5 min**	**5-10 min**	**10-30 min**	**30-60 min**	**1-3 hours**	**3-6 hours**	**6 hrs-all day**
0-A	2 mean	44.20	26.30	8.11	12.70	3.21	3.23	0.87	0.50	0.20	0.68
	3 mean	45.69	25.15	7.50	12.17	3.33	3.42	0.94	0.63	0.26	0.90
	4 mean	39.71	24.92	8.23	14.57	4.30	4.51	1.24	0.87	0.37	1.27
0-B	2 mean	45.84	27.28	8.41	13.18	2.77	1.80	0.30	0.21	0.07	0.13
	3 mean	47.61	26.20	7.82	12.68	2.89	1.92	0.34	0.28	0.10	0.17
	4 mean	41.99	26.35	8.70	15.41	3.77	2.55	0.45	0.39	0.14	0.25
0-C	2 mean	44.55	26.51	8.18	12.80	3.23	3.25	0.88	0.23	0.07	0.29
	3 mean	46.17	25.41	7.58	12.29	3.36	3.46	0.95	0.29	0.10	0.39
	4 mean	40.29	25.28	8.35	14.78	4.36	4.57	1.26	0.41	0.14	0.57
1-A	2 mean	45.63	26.93	8.35	12.54	2.88	2.59	0.67	0.29	0.07	0.06
	3 mean	46.27	25.83	8.06	12.71	3.11	2.76	0.72	0.36	0.09	0.08
	4 mean	40.30	25.81	8.95	15.36	4.12	3.73	0.97	0.50	0.14	0.13
1-B	2 mean	46.54	27.47	8.52	12.79	2.62	1.66	0.23	0.13	0.03	0.02
	3 mean	47.19	26.34	8.22	12.96	2.92	1.86	0.27	0.17	0.04	0.03
	4 mean	41.36	26.49	9.18	15.76	3.90	2.58	0.36	0.25	0.06	0.05
1-C	2 mean	45.73	26.99	8.37	12.56	2.88	2.59	0.67	0.14	0.03	0.04
	3 mean	46.40	25.90	8.08	12.74	3.12	2.77	0.72	0.17	0.04	0.05
	4 mean	40.46	25.91	8.98	15.42	4.14	3.75	0.97	0.25	0.05	0.07
2-A	2 mean	45.18	26.54	8.22	12.69	3.06	2.88	0.76	0.36	0.11	0.19
	3 mean	47.14	25.43	7.60	12.09	3.14	2.98	0.79	0.44	0.14	0.25
	4 mean	41.45	25.49	8.40	14.5	4.06	3.92	1.03	0.59	0.19	0.37
2-B	2 mean	46.38	27.25	8.44	13.03	2.69	1.68	0.26	0.16	0.04	0.07
	3 mean	48.46	26.14	7.81	12.43	2.77	1.76	0.28	0.20	0.05	0.09
	4 mean	43.00	26.44	8.72	15.04	3.6	2.34	0.37	0.28	0.08	0.14
2-C	2 mean	45.34	26.64	8.25	12.74	3.08	2.89	0.76	0.17	0.04	0.10
	3 mean	47.35	25.54	7.63	12.14	3.15	3.00	0.79	0.21	0.05	0.13
	4 mean	41.70	25.65	8.45	14.59	4.08	3.95	1.03	0.29	0.07	0.19

TABLE 1.14: Distribution of spoofing orders' lifetime in percentages of all orders spoofing orders identified by the algorithm

Algo	Lifetime	0-5 sec	5-30 sec	30-60 sec	1-5 min	5-10 min	10-30 min	30-60 min	1-3 hours	3-6 hours	6hrs- all day
0-A	2 mean	44.20	70.50	78.61	91.31	94.52	97.75	98.62	99.12	99.32	100.00
	3 mean	45.69	70.84	78.34	90.51	93.84	97.26	98.20	98.83	99.09	99.99
	4 mean	39.71	64.63	72.86	87.43	91.73	96.24	97.48	98.35	98.72	99.99
0-B	2 mean	45.84	73.12	81.53	94.71	97.48	99.28	99.58	99.79	99.86	99.99
	3 mean	47.61	73.81	81.63	94.31	97.20	99.12	99.46	99.74	99.84	100.01
	4 mean	41.99	68.34	77.04	92.45	96.22	98.77	99.22	99.61	99.75	100.00
0-C	2 mean	44.55	71.06	79.24	92.04	95.27	98.52	99.40	99.63	99.70	99.99
	3 mean	46.17	71.58	79.16	91.45	94.81	98.27	99.22	99.51	99.61	100.00
	4 mean	40.29	65.57	73.92	88.70	93.06	97.63	98.89	99.30	99.44	100.01
1-A	2 mean	45.63	72.56	80.91	93.45	96.33	98.92	99.59	99.88	99.95	100.01
	3 mean	46.27	72.10	80.16	92.87	95.98	98.74	99.46	99.82	99.91	99.99
	4 mean	40.30	66.11	75.06	90.42	94.54	98.27	99.24	99.74	99.88	100.01
1-B	2 mean	46.54	74.01	82.53	95.32	97.94	99.60	99.83	99.96	99.99	100.01
	3 mean	47.19	73.53	81.75	94.71	97.63	99.49	99.76	99.93	99.97	100.00
	4 mean	41.36	67.85	77.03	92.79	96.69	99.27	99.63	99.88	99.94	99.99
1-C	2 mean	45.73	72.72	81.09	93.65	96.53	99.12	99.79	99.93	99.96	100.00
	3 mean	46.40	72.30	80.38	93.12	96.24	99.01	99.73	99.90	99.94	99.99
	4 mean	40.46	66.37	75.35	90.77	94.91	98.66	99.63	99.88	99.93	100.00
2-A	2 mean	45.18	71.72	79.94	92.63	95.69	98.57	99.33	99.69	99.80	99.99
	3 mean	47.14	72.57	80.17	92.26	95.40	98.38	99.17	99.61	99.75	100.00
	4 mean	41.45	66.94	75.34	89.84	93.90	97.82	98.85	99.44	99.63	100.00
2-B	2 mean	46.38	73.63	82.07	95.10	97.79	99.47	99.73	99.89	99.93	100.00
	3 mean	48.46	74.60	82.41	94.84	97.61	99.37	99.65	99.85	99.90	99.99
	4 mean	43.00	69.44	78.16	93.20	96.80	99.14	99.51	99.79	99.87	100.01
2-C	2 mean	45.34	71.98	80.23	92.97	96.05	98.94	99.70	99.87	99.91	100.01
	3 mean	47.35	72.89	80.52	92.66	95.81	98.81	99.60	99.81	99.86	99.99
	4 mean	41.70	67.35	75.80	90.39	94.47	98.42	99.45	99.74	99.81	100.00

TABLE 1.15: Cumulative distribution of spoofing orders' lifetime in percentages of all orders spoofing orders identified by the algorithm

1.8.5 Appendix 5. Ratio of buy and sell spoofing orders by algorithms

Algo	Buy spoofing orders (%)			Sell spoofing orders (%)		
	2 mean	3 mean	4 mean	2 mean	3 mean	4 mean
0-A	49.89%	49.70%	49.63%	50.11%	50.30%	50.37%
0-B	49.84%	49.65%	49.58%	50.16%	50.35%	50.42%
0-C	49.87%	49.69%	49.63%	50.13%	50.31%	50.37%
1-A	49.18%	49.36%	49.47%	50.82%	50.64%	50.53%
1-B	49.14%	49.32%	49.41%	50.86%	50.68%	50.59%
1-C	49.17%	49.35%	49.46%	50.83%	50.65%	50.54%
2-A	49.82%	49.54%	49.53%	50.18%	50.46%	50.47%
2-B	49.79%	49.49%	49.46%	50.21%	50.51%	50.54%
2-C	49.81%	49.53%	49.52%	50.19%	50.47%	50.48%

TABLE 1.16: Ratio of buy and sell spoofing orders by algorithms

1.8.6 Appendix 6. Quantity of spoofing orders by algorithms

Algo	Volume	January	February	March	April	May	June	6 months
0-A	2 means	1 843 967	2 472 777	2 543 699	3 294 386	3 713 816	3 659 377	17 528 022
	3 means	814 625	1 251 318	1 249 345	1 993 749	1 923 576	2 070 002	9 302 615
	4 means	490 130	724 085	729 994	1 069 499	872 116	1 135 581	5 021 405
	2 means	11%	14%	15%	19%	21%	21%	100%
	3 means	9%	13%	13%	21%	21%	22%	100%
	4 means	10%	14%	15%	21%	17%	23%	100%
0-B	2 means	1 753 833	2 357 338	2 434 652	3 179 107	3 618 071	3 557 139	16 900 140
	3 means	763 006	1 180 739	1 184 015	1 925 845	1 866 294	2 008 476	8 928 375
	4 means	451 433	673 761	682 747	1 021 174	829 576	1 090 787	4 749 478
	2 means	10%	14%	14%	19%	21%	21%	100%
	3 means	9%	13%	13%	22%	21%	22%	100%
	4 means	10%	14%	14%	22%	17%	23%	100%
0-C	2 means	1 826 807	2 448 466	2 517 992	3 270 768	3 691 660	3 634 202	17 389 895
	3 means	802 826	1 234 295	1 231 615	1 977 711	1 908 385	2 052 223	9 207 055
	4 means	481 217	710 691	717 425	1 058 023	860 561	1 122 229	4 950 146
	2 means	11%	14%	14%	19%	21%	21%	100%
	3 means	9%	13%	13%	21%	21%	22%	100%
	4 means	10%	14%	14%	21%	17%	23%	100%
1-A	2 means	1 263 193	1 777 201	1 970 386	2 420 619	2 689 918	2 560 763	12 682 080
	3 means	525 768	877 045	938 866	1 348 766	1 266 306	1 308 329	6 265 080
	4 means	304 981	491 652	533 341	685 782	548 317	701 320	3 265 393
	2 means	10%	14%	16%	19%	21%	20%	100%
	3 means	8%	14%	15%	22%	20%	21%	100%
	4 means	9%	15%	16%	21%	17%	21%	100%
1-B	2 means	1 226 930	1 731 141	1 923 783	2 374 654	2 654 134	2 523 473	12 434 115
	3 means	509 145	852 781	915 379	1 326 659	1 248 967	1 290 260	6 143 191
	4 means	293 196	475 467	517 054	670 722	535 997	688 960	3 181 396
	2 means	10%	14%	15%	19%	21%	20%	100%
	3 means	8%	14%	15%	22%	20%	21%	100%
	4 means	9%	15%	16%	21%	17%	22%	100%
1-C	2 means	1 259 433	1 772 615	1 964 624	2 415 733	2 685 821	2 556 431	12 654 657
	3 means	523 322	873 853	935 265	1 345 722	1 263 874	1 305 540	6 247 576
	4 means	303 173	489 064	531 084	683 801	546 445	699 167	3 252 734
	2 means	10%	14%	16%	19%	21%	20%	100%
	3 means	8%	14%	15%	22%	20%	21%	100%
	4 means	9%	15%	16%	21%	17%	21%	100%
2-A	2 means	1 730 957	2 332 555	2 437 503	3 150 122	3 509 490	3 512 195	16 672 822
	3 means	739 519	1 166 379	1 180 616	1 894 788	1 810 239	1 981 669	8 773 210
	4 means	442 809	671 638	678 930	1 009 281	828 362	1 085 511	4 716 531
	2 means	10%	14%	15%	19%	21%	21%	100%
	3 means	8%	13%	13%	22%	21%	23%	100%
	4 means	9%	14%	14%	21%	18%	23%	100%
2-B	2 means	1 666 739	2 251 328	2 358 889	3 072 261	3 445 836	3 444 940	16 239 993
	3 means	706 011	1 119 400	1 136 678	1 852 800	1 775 231	1 944 646	8 534 766
	4 means	418 471	639 373	647 813	979 574	802 443	1 058 963	4 546 637
	2 means	10%	14%	15%	19%	21%	21%	100%
	3 means	8%	13%	13%	22%	21%	23%	100%
	4 means	9%	14%	14%	22%	18%	23%	100%
2-C	2 means	1 723 070	2 322 439	2 425 806	3 140 299	3 500 481	3 502 121	16 614 216
	3 means	734 345	1 159 362	1 172 968	1 888 400	1 804 497	1 974 846	8 734 418
	4 means	439 004	666 125	673 768	1 004 865	823 969	1 080 344	4 688 075
	2 means	10%	14%	15%	19%	21%	21%	100%
	3 means	8%	13%	13%	22%	21%	23%	100%
	4 means	9%	14%	14%	21%	18%	23%	100%

TABLE 1.17: Quantity of spoofing orders by different algorithms; nominal and in percentages of all spoofing orders identified by the algorithm

1.8.7 Appendix 7. Spoofing orders distribution among listing levels

Algo	1^{st} listing level	2^{nd} listing level	3^{rd} listing level
0-A	92.48%	4.53%	2.99%
0-B	92.85%	4.34%	2.81%
0-C	92.69%	4.30%	3.01%
1-A	92.61%	4.69%	2.70%
1-B	92.48%	4.53%	2.99%
1-C	92.63%	4.66%	2.71%
2-A	92.96%	4.40%	2.64%
2-B	93.06%	4.32%	2.62%
2-C	93.04%	4.31%	2.65%
Avarage	92.76%	4.44%	2.80%

TABLE 1.18: Spoofing orders distribution among stock listing levels. The table shows the distribution of spoofing orders identified by the algorithms with a minimum order volume of two means ($Y = 2$) of the average order volume over the prevailing five consecutive trading days.

1.8.8 Appendix 8. Intraday distribution of the trading volume

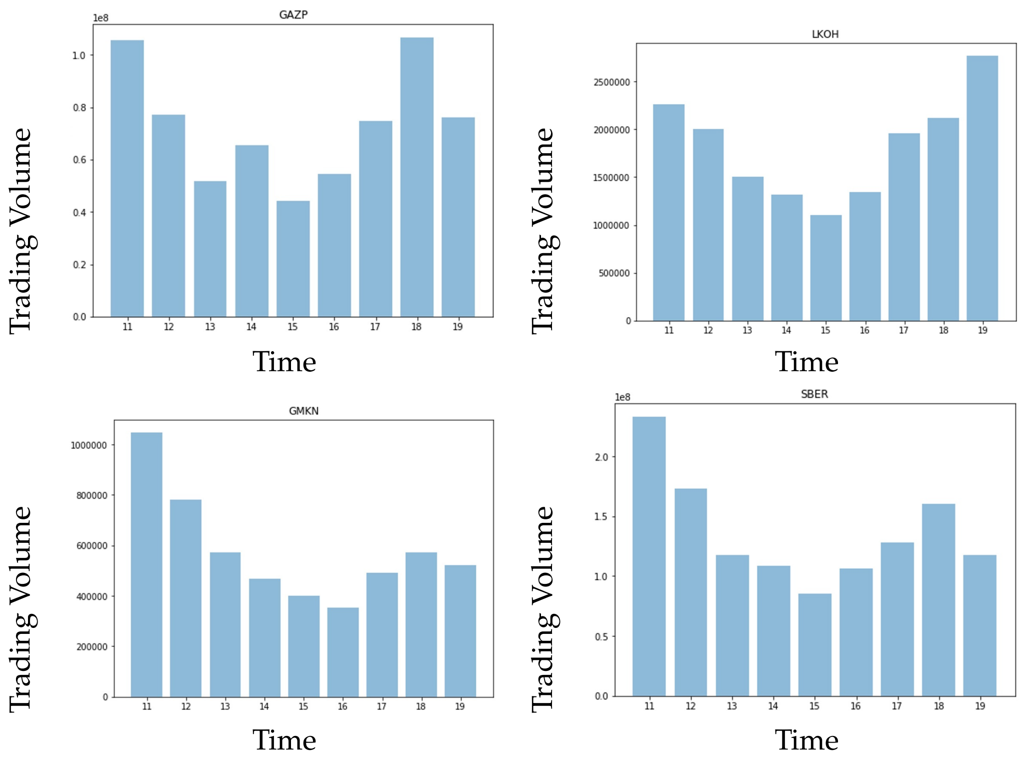


TABLE 1.19: Intraday distribution of the trading volume

1.8.9 Appendix 9. Summary statistics of market quality variables

	Time window	VOL	QS	ES	VOL_{Δ}	QS_{Δ}	ES_{Δ}	$Volume_{\Delta}$
Mean	10 min	0.02	0.07	0.03	-0.03	-0.0007	-0.0014	-0.0009
S.D.	10 min	0.02	0.07	0.04	0.47	0.02	0.04	0.03
Min	10 min	0.00	0.0047	0.0025	-1.00	-1.55	-2.31	-1.59
Max	10 min	1.69	3.08	2.11	1.00	1.67	1.34	1.45
Skewness	10 min	11.46	7.11	8.81	0.00	-2.86	-6.88	-3.69
Kurtosis	10 min	435.31	123.31	193.40	-0.69	611.93	363.81	342.14
Mean	30 min	0.05	0.24	0.14	-0.12	-0.0019	-0.0029	-0.0028
S.D.	30 min	0.09	0.32	0.19	0.58	0.08	0.15	0.11
Min	30 min	0.00	0.01	0.0038	-1.00	-1.60	-4.57	-2.85
Max	30 min	1.80	6.38	5.61	1.00	1.66	2.19	2.02
Skewness	30 min	5.11	3.68	4.07	0.12	-0.70	-2.39	-1.32
Kurtosis	30 min	45.18	23.04	35.63	-1.11	59.84	71.25	52.10
Mean	60 min	0.07	0.35	0.20	-0.22	-0.0031	-0.0054	-0.0047
S.D.	60 min	0.13	0.47	0.28	0.60	0.11	0.22	0.17
Min	60 min	0.00	0.00	0.00	-1.00	-2.49	-6.21	-3.79
Max	60 min	3.81	12.83	5.44	1.00	2.75	4.07	2.90
Skewness	60 min	6.33	4.72	4.53	0.30	-1.22	-1.01	-0.81
Kurtosis	60 min	74.81	45.33	38.00	-1.13	87.59	49.48	53.29

TABLE 1.20: Summary statistics of market quality variables for the 1^{st} listing level stocks. The table shows the mean, standard deviation, minimum and maximum values of the variables with skewness and kurtosis measured separately for different time windows ($X \in \{10, 30, 60\}$ minutes): short-term volatility (VOL), quoted spread (QS), effective spread (ES), change in short-term volatility (VOL_{Δ}), change in quoted spread (QS_{Δ}), change in effective spread (ES_{Δ}), a measure of trading volume ($Volume_{\Delta}$).

	Time window	VOL	QS	ES	VOL_Δ	QS_Δ	ES_Δ	$Volume_\Delta$
Mean	10 min	0.02	0.07	0.04	-0.04	-0.001	0.0003	-0.0011
S.D.	10 min	0.02	0.07	0.04	0.42	0.02	0.05	0.03
Min	10 min	0.00	0.01	0.0038	-1.00	-0.91	-1.15	-1.82
Max	10 min	0.94	1.40	1.92	1.00	0.91	1.09	0.70
Skewness	10 min	7.40	5.03	6.56	0.07	-0.31	-0.80	-4.64
Kurtosis	10 min	146.42	42.36	99.34	-0.46	175.31	75.34	277.58
Mean	30 min	0.06	0.30	0.16	-0.12	-0.0026	-0.007	-0.0038
S.D.	30 min	0.09	0.35	0.20	0.54	0.08	0.17	0.12
Min	30 min	0.00	0.013	0.006	-1.00	-1.60	-3.01	-3.59
Max	30 min	1.66	4.85	4.25	1.00	1.65	1.88	1.62
Skewness	30 min	4.25	2.90	3.11	0.12	-0.54	-1.18	-1.49
Kurtosis	30 min	31.63	13.18	19.55	-0.98	49.07	27.67	51.53
Mean	60 min	0.08	0.42	0.23	-0.18	-0.0045	-0.01	-0.01
S.D.	60 min	0.12	0.48	0.27	0.58	0.11	0.24	0.18
Min	60 min	0.00	0.014	0.0072	-1.00	-3.41	-3.06	-3.21
Max	60 min	3.84	8.63	4.62	1.00	2.68	3.85	2.90
Skewness	60 min	5.88	3.85	3.78	0.26	-1.08	-0.61	-0.19
Kurtosis	60 min	72.06	28.62	27.20	-1.09	74.38	20.51	28.46

TABLE 1.21: Summary statistics of market quality variables for the 2^{nd} listing level stocks. The table shows the mean, standard deviation, minimum and maximum values of the variables with skewness and kurtosis measured separately for different time windows ($X \in \{10, 30, 60\}$ minutes): short-term volatility (VOL), quoted spread (QS), effective spread (ES), change in short-term volatility (VOL_Δ), change in quoted spread (QS_Δ), change in effective spread (ES_Δ), a measure of trading volume ($Volume_\Delta$).

	Time window	VOL	QS	ES	VOL_{Δ}	QS_{Δ}	ES_{Δ}	$Volume_{\Delta}$
Mean	10 min	0.02	0.07	0.04	-0.02	-0.0016	-0.0014	-0.0018
S.D.	10 min	0.02	0.073	0.041	0.39	0.0197	0.0435	0.0272
Min	10 min	0.00	0.009	0.005	-1.00	-0.81	-0.68	-0.96
Max	10 min	0.89	1.14	1.00	1.00	0.58	1.25	0.75
Skewness	10 min	7.80	4.20	4.89	0.06	-1.14	-0.03	-1.06
Kurtosis	10 min	161.90	27.11	41.49	-0.23	169.17	56.70	101.05
Mean	30 min	0.06	0.32	0.18	-0.09	-0.0042	-0.0151	-0.0075
S.D.	30 min	0.08	0.36	0.19	0.51	0.08	0.18	0.12
Min	30 min	0.00	0.015	0.007	-1.00	-1.58	-2.68	-1.87
Max	30 min	1.62	3.75	2.45	1.00	1.60	1.90	1.10
Skewness	30 min	3.98	2.51	2.44	0.09	-0.03	-1.34	-0.65
Kurtosis	30 min	28.60	9.08	9.14	-0.85	47.14	22.44	18.51
Mean	60 min	0.08	0.45	0.25	-0.13	-0.0065	-0.0287	-0.0125
S.D.	60 min	0.13	0.49	0.27	0.56	0.11	0.25	0.18
Min	60 min	0.00	0.014	0.007	-1.00	-3.42	-3.47	-1.88
Max	60 min	3.84	7.78	3.74	1.00	3.42	2.91	2.95
Skewness	60 min	6.11	3.53	3.44	0.20	-0.63	-0.96	0.22
Kurtosis	60 min	83.37	24.07	22.00	-1.04	105.59	14.31	21.53

TABLE 1.22: Summary statistics of market quality variables for the 3^{rd} listing level stocks. The table shows the mean, standard deviation, minimum and maximum values of the variables with skewness and kurtosis measured separately for different time windows ($X \in \{10, 30, 60\}$ minutes): short-term volatility (VOL), quoted spread (QS), effective spread (ES), change in short-term volatility (VOL_{Δ}), change in quoted spread (QS_{Δ}), change in effective spread (ES_{Δ}), a measure of trading volume ($Volume_{\Delta}$).

	Algo(Z)	$SP10_{i,t}^{2m(Z)}$	$SP10_{i,t}^{4m(Z)}$	$SP30_{i,t}^{2m(Z)}$	$SP30_{i,t}^{4m(Z)}$	$SP60_{i,t}^{2m(Z)}$	$SP60_{i,t}^{4m(Z)}$
Mean	0-A	3.91	1.09	3.68	1.03	3.51	0.99
S.D.		4.50	2.06	3.95	1.70	3.71	1.55
Mean	0-B	3.73	1.00	3.51	0.96	3.37	0.92
S.D.		4.39	1.96	3.88	1.66	3.65	1.52
Mean	0-C	3.80	1.04	3.57	0.98	3.41	0.94
S.D.		4.41	1.97	3.90	1.65	3.68	1.52
Mean	1-A	2.88	0.77	2.68	0.72	2.55	0.68
S.D.		3.68	1.62	3.24	1.37	3.04	1.24
Mean	1-B	2.79	0.73	2.60	0.68	2.47	0.65
S.D.		3.62	1.59	3.20	1.35	3.00	1.23
Mean	1-C	2.83	0.74	2.64	0.69	2.50	0.65
S.D.		3.64	1.59	3.22	1.34	3.02	1.22
Mean	2-A	3.67	0.99	3.44	0.93	3.28	0.88
S.D.		4.33	1.89	3.83	1.59	3.60	1.45
Mean	2-B	3.53	0.92	3.31	0.87	3.16	0.83
S.D.		4.25	1.84	3.77	1.56	3.55	1.43
Mean	2-C	3.60	0.95	3.36	0.89	3.20	0.85
S.D.		4.27	1.85	3.79	1.56	3.58	1.42

TABLE 1.23: Summary statistics of the SP variable. The table shows the mean and standard deviation for the primary variable across 91 stocks measured by time windows and detection algorithms: $SPX_{i,t}^{Ym(Z)}$ is a ratio of spoofing orders identified by the algorithms to all placed orders and subsequently cancelled within $X \in \{10, 30, 60\}$ minutes, and the requirement of the spoofing order to have a minimum volume of $Y \in \{2, 4\}$ times the mean volume for 5 previous trading days.

1.8.10 Appendix 10. The effect of SP on market quality

Tables below show the effect of SP on market quality presenting the regression coefficients for the panel regressions (Equation 1.3, 1.4, 1.5) on the change in short-term volatility ($VOL10_\Delta$), quoted spread ($QS10_\Delta$) and effective spread ($ES10_\Delta$). $SPX_{i,t}^{Ym(Z)}$ represents a ratio of spoofing orders for the asset i identified by algorithms Z described in Part 4.2 to all placed orders and subsequently cancelled within $X \in \{10, 30, 60\}$ minute time window, and with the minimum volume requirement of spoofing order to be $Y \in \{2, 4\}$ mean of the average volume for prevailing five consecutive trading days. As control variables we also include dummies for three periods during the day: $Aft_{i,t}$ is for the afternoon (13:00 till 16:00), $Ev_{i,t}$ is for the evening (16:00 till 18:45), or morning (10:00 till 13:00) otherwise; $VolumeX_\Delta$, that represents the change in trading volume, is described by Eq. (2), and one lag value of the dependent variable. All these variables are standardized, and the panel estimation clusters errors by asset id and time (day-time window). Below each coefficient, we show the robust standard errors in parenthesis, R2 and the regression's adjusted R2. Significance levels are denoted by *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

$SP10_{i,t}^{2m(Z)}$	$VOL10_{\Delta-1}$	$VOL10_{\Delta+1}$	$VOL10_{\Delta+2}$	$QS10_{\Delta-1}$	$QS10_{\Delta+1}$	$QS10_{\Delta+2}$	$ES10_{\Delta-1}$	$ES10_{\Delta+1}$	$ES10_{\Delta+2}$
0-A	-0.0001***	0.0002***	0.0001***	-0.0003***	0.001***	0.0004***	-0.0004***	0.0003***	0.0002***
	(0.0001)	(0.00003)	(0.00003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
0-B	-0.0001***	0.0001***	0.00004*	-0.0004***	0.001***	0.0004***	-0.0005***	0.0003***	0.0002***
	(0.0001)	(0.00003)	(0.00003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
0-C	-0.0001***	0.0002***	0.00005	-0.0004***	0.001***	0.0004***	-0.0005***	0.0003***	0.0002***
	(0.0001)	(0.00003)	(0.00003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
1-A	-0.0002***	0.0002***	0.0001**	-0.0004***	0.001***	0.0004***	-0.001***	0.0004***	0.0003***
	(0.00005)	(0.00004)	(0.00004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
1-B	-0.0002***	0.0002***	0.0001***	-0.0004***	0.001***	0.0004***	-0.001***	0.0004***	0.0003***
	(0.00005)	(0.00004)	(0.00004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
1-C	-0.0002***	0.0002***	0.0001***	-0.0004***	0.001***	0.0004***	-0.001***	0.0004***	0.0003***
	(0.00005)	(0.00004)	(0.00004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
2-A	-0.0002***	0.0002***	0.0001***	-0.0003***	0.001***	0.0003***	-0.0005***	0.0003***	0.0002***
	(0.0001)	(0.00003)	(0.00003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
2-B	-0.0002***	0.0002***	0.0001***	-0.0004***	0.001***	0.0003***	-0.0005***	0.0003***	0.0002***
	(0.0001)	(0.00003)	(0.00003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001
2-C	-0.0002***	0.0002***	0.0001***	-0.0004***	0.001***	0.0003***	-0.0005***	0.0003***	0.0002***
	(0.0001)	(0.00003)	(0.00003)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

TABLE 1.24: The effect of SP on market quality using $SP10_{i,t}^{2m(Z)}$ algorithm group for 91 stocks, totally 9 algorithms using 10 minutes time windows and 2-mean minimum volume requirement. Below each coefficient, we show the robust standard errors in parenthesis, R2 and the regression's adjusted R2. Significance levels are denoted by $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

$SP30_{i,t}^{2m(Z)}$	$VOL30_{\Delta-1}$	$VOL30_{\Delta+1}$	$VOL30_{\Delta+2}$	$QS30_{\Delta-1}$	$QS30_{\Delta+1}$	$QS30_{\Delta+2}$	$ES30_{\Delta-1}$	$ES30_{\Delta+1}$	$ES30_{\Delta+2}$
0-A	-0.0004***	0.0001***	0.0001	-0.001***	0.001***	0.001***	-0.001***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0003)	(0.0002)	(0.0001)	(0.0002)
0-B	-0.0004***	0.0001***	0.00003	-0.002***	0.001***	0.0005***	-0.001***	0.0005***	0.0005***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
0-C	-0.0004***	0.0001***	0.0001	-0.002***	0.001***	0.001***	-0.001***	0.001***	0.0005***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0003)	(0.0002)	(0.0001)	(0.0002)
1-A	-0.001***	0.0003***	0.0003***	-0.002***	0.001***	0.001***	-0.001***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)
1-B	-0.001***	0.0003***	0.0002***	-0.002***	0.001***	0.001***	-0.001***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)
1-C	-0.001***	0.0003***	0.0002***	-0.002***	0.001***	0.001***	-0.001***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)
2-A	-0.0005***	0.0002***	0.0001***	-0.002***	0.001***	0.001***	-0.001***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
2-B	-0.0004***	0.0002***	0.0001**	-0.002***	0.001***	0.001***	-0.001***	0.0005***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
2-C	-0.0005***	0.0002***	0.0001**	-0.002***	0.001***	0.001***	-0.001***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)

TABLE 1.25: The effect of SP on market quality using $SP30_{i,t}^{2m(Z)}$ algorithm group for 91 stocks, totally 9 algorithms using 30 minutes time windows and 2-mean minimum volume requirement. Below each coefficient, we show the robust standard errors in parenthesis, R2 and the regression's adjusted R2. Significance levels are denoted by $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

$SP60^{2m(Z)}_{i,t}$	$VOL_{\Delta-1}$	$VOL_{\Delta+1}$	$VOL_{\Delta+2}$	$QS_{\Delta-1}$	$QS_{\Delta+1}$	$QS_{\Delta+2}$	$ES_{\Delta-1}$	$ES_{\Delta+1}$	$ES_{\Delta+2}$
0-A	-0.001 ***	0.0001*	0.0003 ***	-0.003 ***	0.001 ***	0.001 ***	-0.001 ***	0.0004 ***	0.001 ***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0002)	(0.0003)	(0.0003)	(0.0001)	(0.0002)
0-B	-0.001 ***	0.0001	0.0003 ***	-0.003 ***	0.001 ***	0.001 ***	-0.001 ***	0.0004 ***	0.001 ***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0002)	(0.0003)	(0.0003)	(0.0001)	(0.0001)
0-C	-0.001 ***	0.0001*	0.0003 ***	-0.003 ***	0.001 ***	0.001 ***	-0.001 ***	0.0004 ***	0.001 ***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0002)	(0.0003)	(0.0003)	(0.0001)	(0.0001)
1-A	-0.001 ***	0.0002 ***	0.001 ***	-0.003 ***	0.001 ***	0.002 ***	-0.002 ***	0.0004 ***	0.001 ***
	(0.0001)	(0.0001)	(0.0001)	(0.001)	(0.0002)	(0.0003)	(0.0003)	(0.0001)	(0.0002)
1-B	-0.001 ***	0.0002**	0.0005 ***	-0.003 ***	0.001 ***	0.001 ***	-0.002 ***	0.0004 ***	0.001 ***
	(0.0001)	(0.0001)	(0.0001)	(0.001)	(0.0003)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
1-C	-0.001 ***	0.0002**	0.001 ***	-0.003 ***	0.001 ***	0.001 ***	-0.002 ***	0.0004 ***	0.001 ***
	(0.0001)	(0.0001)	(0.0001)	(0.001)	(0.0003)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
2-A	-0.001 ***	0.0002**	0.0004 ***	-0.003 ***	0.001 ***	0.001 ***	-0.001 ***	0.0004 ***	0.001 ***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0002)	(0.0003)	(0.0003)	(0.0001)	(0.0001)
2-B	-0.001 ***	0.0002*	0.0003 ***	-0.003 ***	0.001 ***	0.001 ***	-0.001 ***	0.0004 ***	0.001 ***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0002)	(0.0002)	(0.0003)	(0.0001)	(0.0001)
2-C	-0.001 ***	0.0002**	0.0003 ***	-0.003 ***	0.001 ***	0.001 ***	-0.001 ***	0.0004 ***	0.001 ***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0002)	(0.0003)	(0.0003)	(0.0001)	(0.0001)

TABLE 1.26: The effect of SP on market quality using $SP60^{2m(Z)}_{i,t}$ algorithm group for 91 stocks, totally 9 algorithms using 60 minutes time windows and 2-mean minimum volume requirement. Below each coefficient, we show the robust standard errors in parenthesis, R2 and the regression's adjusted R2. Significance levels are denoted by $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

$SP10^{4m(Z)}_{i,t}$	$VOL_{\Delta-1}$	$VOL_{\Delta+1}$	$VOL_{\Delta+2}$	$QS_{\Delta-1}$	$QS_{\Delta+1}$	$QS_{\Delta+2}$	$ES_{\Delta-1}$	$ES_{\Delta+1}$	$ES_{\Delta+2}$
0-A	-0.0002***	0.0001***	0.0001	-0.001***	0.001***	0.001***	-0.001***	0.0004***	0.0002***
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
0-B	-0.0001**	0.0001*	0.00004	-0.001***	0.001***	0.001***	-0.001***	0.0003***	0.0002***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
0-C	-0.0002***	0.0001*	0.00004	-0.001***	0.001***	0.001***	-0.001***	0.0004***	0.0002***
	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
1-A	-0.0001*	0.0002***	0.0002**	-0.001***	0.001***	0.001***	-0.001***	0.0005***	0.0004***
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
1-B	-0.0001	0.0002***	0.0002***	-0.001***	0.001***	0.001***	-0.001***	0.0005***	0.0003***
	(0.0001)	(0.00005)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
1-C	-0.0001*	0.0002***	0.0002***	-0.001***	0.001***	0.001***	-0.001***	0.0005***	0.0004***
	(0.0001)	(0.00005)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
2-A	-0.0002***	0.0001***	0.0001**	-0.001***	0.001***	0.001***	-0.001***	0.0003***	0.0002***
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0001)	(0.0001)
2-B	-0.0002***	0.0001**	0.0001	-0.001***	0.001***	0.001***	-0.001***	0.0003***	0.0002***
	(0.0001)	(0.00005)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
2-C	-0.0002***	0.0001**	0.0001*	-0.001***	0.001***	0.001***	-0.001***	0.0004***	0.0002***
	(0.0001)	(0.0001)	(0.00005)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

TABLE 1.27: The effect of SP on market quality using $SP10^{4m(Z)}_{i,t}$ algorithm group for 91 stocks, totally 9 algorithms using 10 minutes time windows and 4-mean minimum volume requirement. Below each coefficient, we show the robust standard errors in parenthesis, R2 and the regression's adjusted R2. Significance levels are denoted by $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

$SP30^{4m(Z)}_{i,t}$	$VOL_{\Delta-1}$	$VOL_{\Delta+1}$	$VOL_{\Delta+2}$	$QS_{\Delta-1}$	$QS_{\Delta+1}$	$QS_{\Delta+2}$	$ES_{\Delta-1}$	$ES_{\Delta+1}$	$ES_{\Delta+2}$
0-A	-0.001***	0.0001	0.0001	-0.002***	0.002***	0.001***	-0.001***	0.0005***	0.001***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0004)	(0.001)	(0.0003)	(0.0002)	(0.0003)
0-B	-0.001***	0.00001	0.00001	-0.002***	0.002***	0.001***	-0.001***	0.0004***	0.001***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0004)	(0.001)	(0.0004)	(0.0002)	(0.0003)
0-C	-0.001***	0.00002	0.00004	-0.002***	0.002***	0.001***	-0.001***	0.0005***	0.001***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0004)	(0.001)	(0.0004)	(0.0002)	(0.0003)
1-A	-0.001***	0.0003**	0.0004**	-0.002***	0.002***	0.001**	-0.002***	0.001***	0.001***
	(0.0002)	(0.0001)	(0.0002)	(0.001)	(0.0005)	(0.001)	(0.0004)	(0.0002)	(0.0003)
1-B	-0.001***	0.0003**	0.0004***	-0.002***	0.002***	0.001***	-0.002***	0.001***	0.001***
	(0.0002)	(0.0001)	(0.0002)	(0.001)	(0.0005)	(0.0005)	(0.0005)	(0.0002)	(0.0004)
1-C	-0.001***	0.0003**	0.0003***	-0.002***	0.002***	0.001***	-0.002***	0.001***	0.001***
	(0.0002)	(0.0001)	(0.0002)	(0.001)	(0.0005)	(0.0005)	(0.0005)	(0.0002)	(0.0004)
2-A	-0.001***	0.0002*	0.0002*	-0.002***	0.002***	0.001***	-0.001***	0.0005***	0.001***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0003)	(0.0005)	(0.0004)	(0.0002)	(0.0003)
2-B	-0.001***	0.0002	0.0001	-0.002***	0.002***	0.001***	-0.001***	0.0004**	0.001***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0004)	(0.0005)	(0.0004)	(0.0002)	(0.0003)
2-C	-0.001***	0.0002	0.0002	-0.002***	0.002***	0.001***	-0.001***	0.0004**	0.001***
	(0.0002)	(0.0001)	(0.0001)	(0.001)	(0.0004)	(0.0005)	(0.0004)	(0.0002)	(0.0003)

TABLE 1.28: The effect of SP on market quality using $SP30^{4m(Z)}_{i,t}$ algorithm group for 91 stocks, totally 9 algorithms using 30 minutes time windows and 4-mean minimum volume requirement. Below each coefficient, we show the robust standard errors in parenthesis, R2 and the regression's adjusted R2. Significance levels are denoted by $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

$SP60^{4m(Z)}_{i,t}$	$VOL_{\Delta-1}$	$VOL_{\Delta+1}$	$VOL_{\Delta+2}$	$QS_{\Delta-1}$	$QS_{\Delta+1}$	$QS_{\Delta+2}$	$ES_{\Delta-1}$	$ES_{\Delta+1}$	$ES_{\Delta+2}$
0-A	-0.001**	0.00000	0.0005*	-0.004***	0.001*	0.001***	-0.002***	0.0002	0.001***
	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0003)	(0.0004)
0-B	-0.001***	-0.00002	0.0003	-0.004***	0.001**	0.001**	-0.002***	0.0002	0.001***
	(0.0004)	(0.0002)	(0.0002)	(0.001)	(0.001)	(0.0005)	(0.001)	(0.0003)	(0.0004)
0-C	-0.001***	-0.00001	0.0004**	-0.004***	0.001**	0.001***	-0.002***	0.0002	0.001***
	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0003)	(0.0004)
1-A	-0.001***	0.0003*	0.001***	-0.004***	0.001*	0.002***	-0.002***	0.001*	0.001***
	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0004)	(0.0004)
1-B	-0.001***	0.0003	0.001**	-0.004***	0.001**	0.002***	-0.002***	0.001*	0.001***
	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0004)	(0.0004)
1-C	-0.001***	0.0003	0.001***	-0.004***	0.001**	0.002***	-0.002***	0.001	0.001***
	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0004)	(0.0004)
2-A	-0.001***	0.0001	0.0004**	-0.004***	0.001**	0.001***	-0.002***	0.0003	0.001***
	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0003)	(0.0003)
2-B	-0.001***	0.00004	0.0003	-0.004***	0.001*	0.001**	-0.002***	0.0002	0.001***
	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.001)	(0.0005)	(0.001)	(0.0003)	(0.0003)
2-C	-0.001***	0.00004	0.0004*	-0.004***	0.001*	0.001**	-0.002***	0.0003	0.001***
	(0.0003)	(0.0001)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0003)	(0.0003)

TABLE 1.29: The effect of SP on market quality using $SP60^{4m(Z)}_{i,t}$ algorithm group for 91 stocks, totally 9 algorithms using 60 minutes time windows and 4-mean minimum volume requirement. Below each coefficient, we show the robust standard errors in parenthesis, R2 and the regression's adjusted R2. Significance levels are denoted by $*p < 0.1$; $**p < 0.05$; $***p < 0.01$.

1.8.11 Appendix 11. Example of the economic significance

30 min time window	1^{st} listing level stocks					2^{nd} listing level stocks					3^{rd} listing level stocks				
	SBER	GMKN	LKOH	MGNT	MOEX	OGKB	APTK	DVEC	KMAZ	MRKU	BANEP	MOBB	BLNG	LSNG	LSNGP
$SP30_{(i,t)}^{(2m(1-C))}$ mean	4.33	3.98	2.53	3.14	3.31	3.31	5.40	0.95	0.69	0.39	4.11	2.59	1.25	0.30	2.40
$SP30_{(i,t)}^{(2m(1-C))}$ SD	1.14	2.54	1.91	2.07	2.69	2.69	4.65	1.62	1.31	1.07	2.87	3.42	1.70	0.85	2.28
Change in $QS_{\Delta-1}$	-24.37%	-20.51%	-20.06%	-15.14%	-29.45%	-29.45%	-3.51%	-0.30%	-0.41%	-0.39%	-9.96%	-0.80%	-0.42%	-0.33%	-3.49%
Change in $QS_{\Delta+1}$	5.09%	6.92%	7.16%	5.63%	9.55%	9.55%	1.65%	0.15%	0.21%	0.20%	4.10%	0.40%	0.21%	0.16%	1.66%
Change in $QS_{\Delta+2}$	5.09%	6.92%	7.16%	5.63%	9.55%	9.55%	1.65%	0.15%	0.21%	0.20%	4.10%	0.40%	0.21%	0.16%	1.66%
Change in $ES_{\Delta-1}$	-21.09%	-19.56%	-19.62%	-14.52%	-26.83%	-26.83%	-3.15%	-0.29%	-0.36%	-0.35%	-9.48%	-0.71%	-0.37%	-0.28%	-3.01%
Change in $ES_{\Delta+1}$	8.09%	12.12%	12.90%	10.08%	16.17%	16.17%	2.94%	0.29%	0.36%	0.35%	7.45%	0.71%	0.37%	0.28%	2.83%
Change in $ES_{\Delta+2}$	8.09%	12.12%	12.90%	10.08%	16.17%	16.17%	2.94%	0.29%	0.36%	0.35%	7.45%	0.71%	0.37%	0.28%	2.83%
Change in $VOL_{\Delta-1}$	-7.19%	-18.60%	-14.33%	-11.24%	-21.07%	-21.07%	-8.15%	-1.34%	-1.10%	-1.19%	-20.86%	-2.43%	-1.22%	-0.87%	-8.07%
Change in $VOL_{\Delta+1}$	1.59%	4.04%	3.45%	2.76%	4.73%	4.73%	2.18%	0.40%	0.33%	0.36%	4.51%	0.71%	0.36%	0.26%	2.18%
Change in $VOL_{\Delta+2}$	1.08%	2.75%	2.33%	1.87%	3.22%	3.22%	1.46%	0.26%	0.22%	0.24%	3.07%	0.48%	0.24%	0.17%	1.46%

TABLE 1.30: Example of the economic significance of the change in volatility, quoted and effective spreads with the change in spoofing ratio for 15 stocks from different listing levels. An example of SP ratio using algorithm $1-C$ with 30 minutes time windows and $Y = 2\left(SP30_{i,t}^{2m(1-C)}\right)$.

60 min time window	1^{st} listing level stocks					2^{nd} listing level stocks					3^{rd} listing level stocks				
	SBER	GMKN	LKOH	MGNT	MOEX	OGKB	APTK	DVEC	KMAZ	MRKU	BANEP	MOBB	BLNG	LSNG	LSNGP
mean	4.33	4.01	2.51	3.16	3.35	3.35	5.14	0.81	0.57	0.34	4.06	2.26	1.12	0.25	2.32
SD	1.07	2.47	1.89	2.01	2.66	3.22	4.22	1.22	0.92	0.74	2.56	2.64	1.37	0.56	2.04
Change in	-62.48%	-35.50%	-33.92%	-24.78%	-52.89%	-8.65%	-4.92%	-0.35%	-0.44%	-0.41%	-14.28%	-0.95%	-0.53%	-0.34%	-4.80%
Change in	4.76%	6.70%	7.06%	5.44%	9.34%	2.57%	1.52%	0.12%	0.15%	0.14%	3.66%	0.31%	0.17%	0.11%	1.49%
Change in	4.76%	6.70%	7.06%	5.44%	9.34%	2.57%	1.52%	0.12%	0.15%	0.14%	3.66%	0.31%	0.17%	0.11%	1.49%
Change in	-18.97%	-54.71%	-51.63%	-35.97%	-78.37%	-10.95%	-5.98%	-0.45%	-0.53%	-0.49%	-19.48%	-1.14%	-0.63%	-0.39%	-5.58%
Change in	3.71%	5.30%	5.66%	4.29%	7.17%	1.93%	1.10%	0.09%	0.11%	0.10%	2.84%	0.23%	0.12%	0.08%	1.04%
Change in	7.57%	11.74%	12.72%	9.74%	15.80%	4.68%	2.70%	0.23%	0.27%	0.24%	6.65%	0.56%	0.31%	0.19%	2.55%
Change in	-6.72%	-18.08%	-14.16%	-10.92%	-20.76%	-10.91%	-7.47%	-1.02%	-0.76%	-0.81%	-18.31%	-1.91%	-0.99%	-0.59%	-7.08%
Change in	1.01%	2.68%	2.31%	1.81%	3.16%	1.92%	1.35%	0.20%	0.15%	0.16%	2.71%	0.37%	0.20%	0.12%	1.29%
Change in	4.35%	11.40%	10.29%	8.13%	13.63%	8.89%	6.32%	1.01%	0.75%	0.80%	11.58%	1.85%	0.98%	0.59%	6.10%

TABLE 1.31: Example of the economic significance of the change in volatility, quoted and effective spreads with the change in spoofing ratio for 15 stocks from different listing levels. An example of SP ratio using algorithm $1 - C$ with 60 minutes time windows and $Y = 2\left(SP60_{(i,t)}^{(2m(1-C))}\right)$.

1.8.12 Appendix 12. Correlation matrices between SP ratio in one stock with SP ratio in another stock

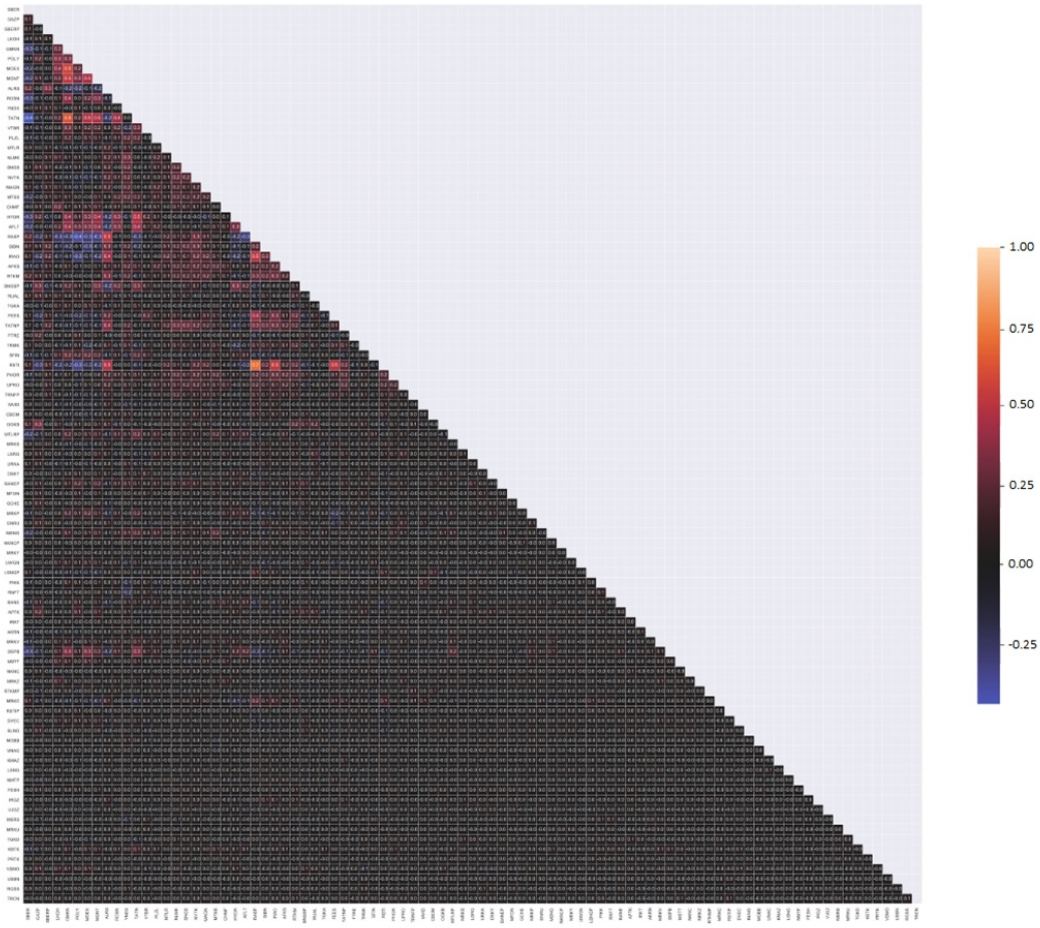


FIGURE 1.7: Correlation matrix of $SP10_{i,t}^{2m(1-C)}$: 91 stocks, 10 minutes, 2-mean minimum volume, algorithm $(1 - C)$. Stocks are presented from most liquid (left top) to less liquid (right bottom). Black colour is used for 0 correlation, light orange colour for correlation of 1 and light blue for negative correlation of -0.5.

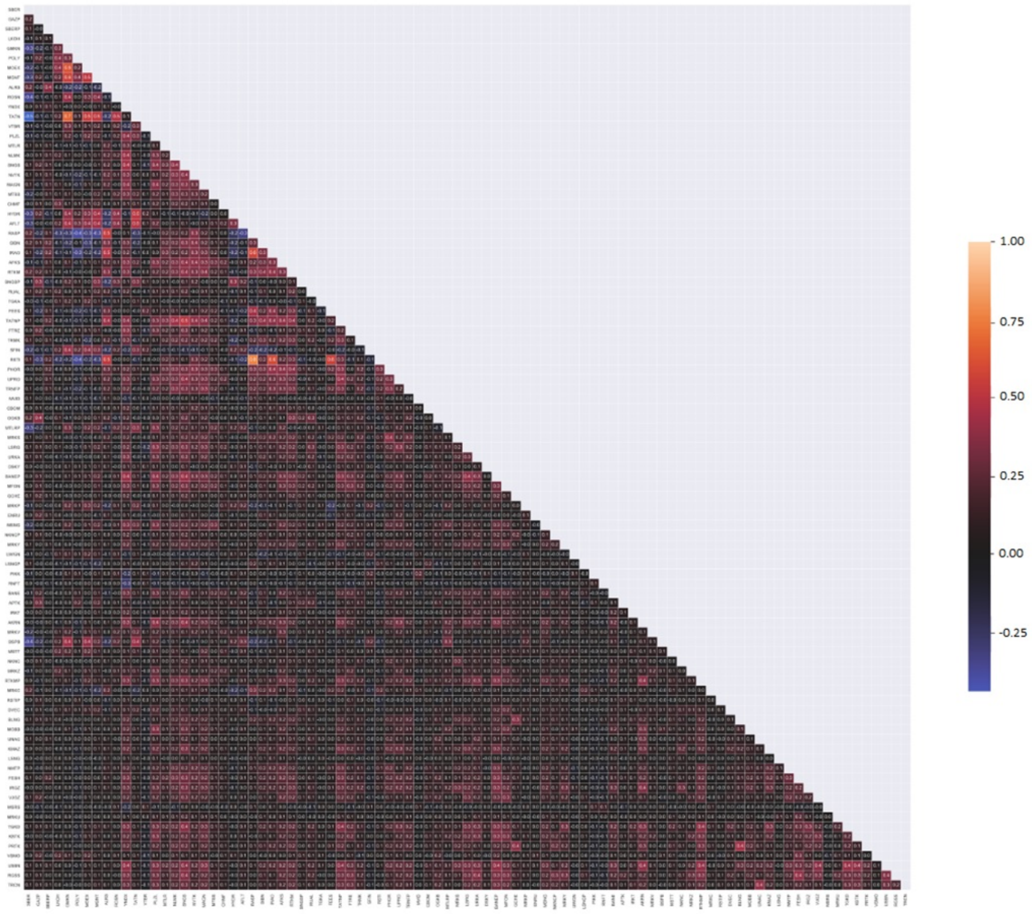


FIGURE 1.8: Correlation matrix of $SP30_{i,t}^{2m(1-C)}$: 91 stocks, 30 minutes, 2-mean minimum volume, algorithm $(1 - C)$. Stocks are presented from most liquid (left top) to less liquid (right bottom). Black colour is used for 0 correlation, light orange colour for correlation of 1 and light blue for negative correlation of -0.5.

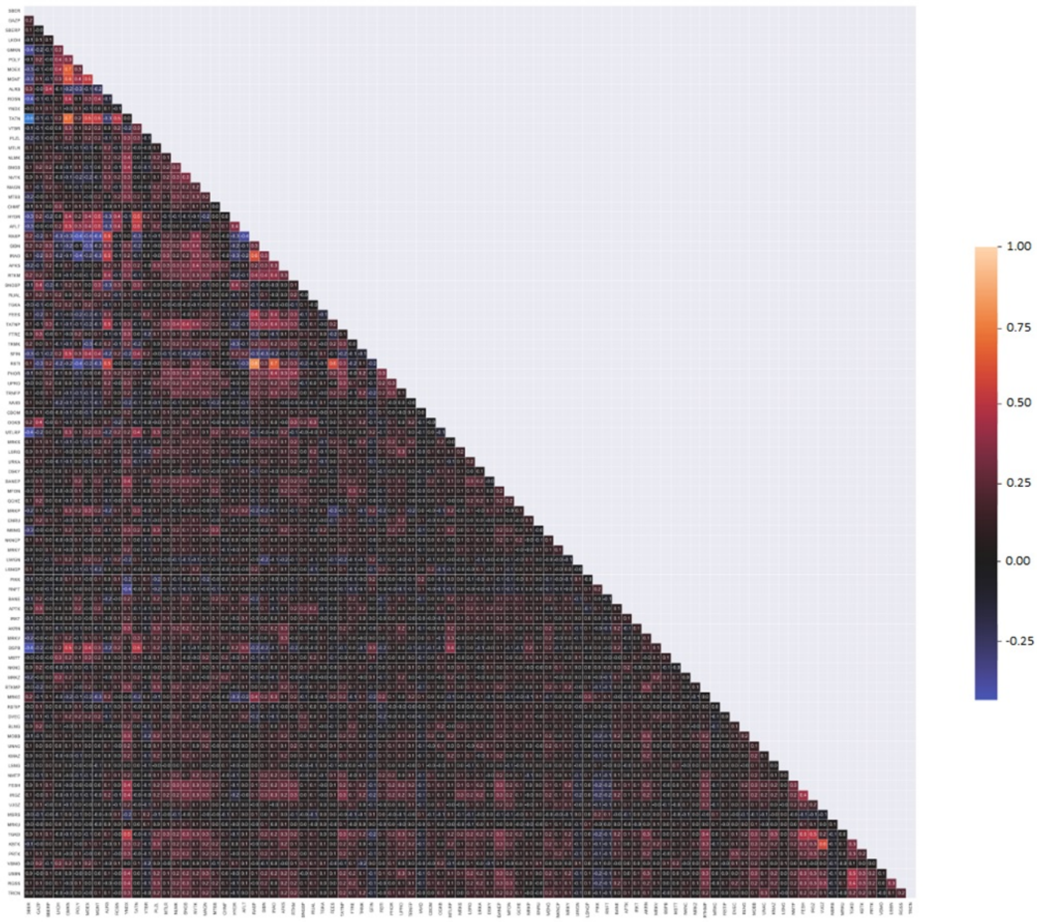


FIGURE 1.9: Correlation matrix of $SP60^{2m(1-C)}_{i,t}$: 91 stocks, 60 minutes, 2-mean minimum volume, algorithm $(1-C)$. Stocks are presented from most liquid (left top) to less liquid (right bottom). Black colour is used for 0 correlation, light orange colour for correlation of 1 and light blue for negative correlation of -0.5.

1.8.13 Appendix 13. The effect of *MeanSP* on market quality

	$\mathbf{VOL}_{\Delta-1}$	$\mathbf{VOL}_{\Delta+1}$	$\mathbf{VOL}_{\Delta+2}$	$\mathbf{QS}_{\Delta-1}$	$\mathbf{QS}_{\Delta+1}$	$\mathbf{QS}_{\Delta+2}$	$\mathbf{ES}_{\Delta-1}$	$\mathbf{ES}_{\Delta+1}$	$\mathbf{ES}_{\Delta+2}$
$SP10^{2m(1-C)\sim}_{i,t}$	-0.006***	0.005***	0.004***	-0.008***	0.022***	0.010***	-0.011***	0.008***	0.006***
	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)
Observations	363,621	363,621	363,621	363,621	363,621	363,621	363,621	363,621	363,621
R2	0.099	0.126	0.062	0.086	0.016	0.021	0.103	0.063	0.08
Adjusted R2	0.099	0.126	0.062	0.086	0.015	0.021	0.103	0.063	0.079

	$\mathbf{VOL}_{\Delta-1}$	$\mathbf{VOL}_{\Delta+1}$	$\mathbf{VOL}_{\Delta+2}$	$\mathbf{QS}_{\Delta-1}$	$\mathbf{QS}_{\Delta+1}$	$\mathbf{QS}_{\Delta+2}$	$\mathbf{ES}_{\Delta-1}$	$\mathbf{ES}_{\Delta+1}$	$\mathbf{ES}_{\Delta+2}$
$SP30^{2m(1-C)\sim}_{i,t}$	-0.013***	0.004***	0.008***	-0.044***	0.032***	0.020***	-0.021***	0.007***	0.012***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	147,276	147,276	147,276	147,276	147,276	147,276	147,276	147,276	147,276
R2	0.088	0.16	0.053	0.03	0.032	0.022	0.094	0.107	0.092
Adjusted R2	0.088	0.16	0.052	0.029	0.031	0.022	0.093	0.106	0.091

	$\mathbf{VOL}_{\Delta-1}$	$\mathbf{VOL}_{\Delta+1}$	$\mathbf{VOL}_{\Delta+2}$	$\mathbf{QS}_{\Delta-1}$	$\mathbf{QS}_{\Delta+1}$	$\mathbf{QS}_{\Delta+2}$	$\mathbf{ES}_{\Delta-1}$	$\mathbf{ES}_{\Delta+1}$	$\mathbf{ES}_{\Delta+2}$
$SP60^{2m(1-C)\sim}_{i,t}$	-0.020***	-0.0004	0.018***	-0.078***	0.013***	0.051***	-0.028***	0.001	0.022***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)
Observations	75,421	75,421	75,421	75,421	75,421	75,421	75,421	75,421	75,421
R2	0.079	0.209	0.042	0.028	0.111	0.01	0.093	0.154	0.062
Adjusted R2	0.078	0.208	0.041	0.026	0.11	0.009	0.092	0.153	0.061

TABLE 1.32: The effect of SP on market quality using *MeanSP* as instrumental variable (IV). This table shows the coefficient of the IV of SP: baseline results. $SPX^{Ym(1-C)\sim}_{i,t}$ represents a *MeanSP* ratio for asset i identified by the algorithm $1 - C$ to all placed orders and subsequently cancelled within $X \in \{10, 30, 60\}$ minute time window and with the requirement of minimum volume for the spoofing order to be two means ($Y = 2$) of the average volume for prevailing five consecutive days. Below each coefficient, we show the standard errors. Significance levels are denoted by $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

2-mean Algo $0-A$	$\mathbf{VOLX_{\Delta-1}}$	$\mathbf{VOLX_{\Delta+1}}$	$\mathbf{VOLX_{\Delta+2}}$	$\mathbf{QSX_{\Delta-1}}$	$\mathbf{QSX_{\Delta+1}}$	$\mathbf{QSX_{\Delta+2}}$	$\mathbf{ESX_{\Delta-1}}$	$\mathbf{ESX_{\Delta+1}}$	$\mathbf{ESX_{\Delta+2}}$
$SP10_{i,t}^{2m(0-A)\sim}$	-0.005***	0.004***	0.003***	-0.007***	0.019***	0.009***	-0.009***	0.007***	0.005***
	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0004)
$SP30_{i,t}^{2m(0-A)\sim}$	-0.011***	0.003***	0.007***	-0.037***	0.026***	0.016***	-0.017***	0.006***	0.009***
	(0.001)	(0.0004)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$SP60_{i,t}^{2m(0-A)\sim}$	-0.017***	-0.001	0.016***	-0.066***	0.010***	0.044***	-0.023***	0.0002	0.019***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
2-mean Algo $0-B$									
$SP10_{i,t}^{2m(0-B)\sim}$	-0.006***	0.005***	0.003***	-0.008***	0.021***	0.009***	-0.010***	0.007***	0.005***
	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)
$SP30_{i,t}^{2m(0-B)\sim}$	-0.012***	0.003***	0.006***	-0.039***	0.026***	0.013***	-0.018***	0.006***	0.008***
	(0.001)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$SP60_{i,t}^{2m(0-B)\sim}$	-0.018***	-0.001	0.016***	-0.068***	0.010***	0.045***	-0.024***	-0.00001	0.019***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)
2-mean Algo $0-C$									
$SP10_{i,t}^{2m(0-C)\sim}$	-0.006***	0.004***	0.003***	-0.008***	0.020***	0.009***	-0.010***	0.007***	0.005***
	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0004)
$SP30_{i,t}^{2m(0-C)\sim}$	-0.012***	0.003***	0.006***	-0.039***	0.026***	0.014***	-0.018***	0.006***	0.009***
	(0.001)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$SP60_{i,t}^{2m(0-C)\sim}$	-0.018***	-0.001	0.016***	-0.068***	0.010***	0.045***	-0.024***	0.0001	0.019***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)
2-mean Algo $1-A$									
$SP10_{i,t}^{2m(1-A)\sim}$	-0.006***	0.005***	0.004***	-0.007***	0.022***	0.010***	-0.010***	0.008***	0.006***
	(0.0004)	(0.0004)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)
$SP30_{i,t}^{2m(1-A)\sim}$	-0.013***	0.004***	0.008***	-0.042***	0.031***	0.021***	-0.020***	0.007***	0.012***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$SP60_{i,t}^{2m(1-A)\sim}$	-0.020***	-0.0004	0.018***	-0.077***	0.013***	0.051***	-0.027***	0.001	0.022***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
2-mean Algo $1-B$									
$SP10_{i,t}^{2m(1-B)\sim}$	-0.006***	0.005***	0.004***	-0.008***	0.023***	0.010***	-0.011***	0.008***	0.006***
	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)
$SP30_{i,t}^{2m(1-B)\sim}$	-0.013***	0.004***	0.008***	-0.044***	0.031***	0.020***	-0.021***	0.007***	0.012***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$SP60_{i,t}^{2m(1-B)\sim}$	-0.020***	-0.0004	0.018***	-0.078***	0.013***	0.051***	-0.028***	0.001	0.023***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)
2-mean Algo $1-C$									
$SP10_{i,t}^{2m(1-C)\sim}$	-0.006***	0.005***	0.004***	-0.008***	0.022***	0.010***	-0.011***	0.008***	0.006***
	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)
$SP30_{i,t}^{2m(1-C)\sim}$	-0.013***	0.004***	0.008***	-0.044***	0.032***	0.020***	-0.021***	0.007***	0.012***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$SP60_{i,t}^{2m(1-C)\sim}$	-0.020***	-0.0004	0.018***	-0.078***	0.013***	0.051***	-0.028***	0.001	0.022***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)

Continued on next page

2-mean Algo $2-A$	$\mathbf{VOLX_{\Delta-1}}$	$\mathbf{VOLX_{\Delta+1}}$	$\mathbf{VOLX_{\Delta+2}}$	$\mathbf{QSX_{\Delta-1}}$	$\mathbf{QSX_{\Delta+1}}$	$\mathbf{QSX_{\Delta+2}}$	$\mathbf{ESX_{\Delta-1}}$	$\mathbf{ESX_{\Delta+1}}$	$\mathbf{ESX_{\Delta+2}}$
$SP10^{2m(2-A)\sim}_{i,t}$	-0.005***	0.004***	0.003***	-0.007***	0.019***	0.008***	-0.009***	0.006***	0.005***
	(0.0003)	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0004)	(0.0004)
$SP30^{2m(2-A)\sim}_{i,t}$	-0.011***	0.003***	0.006***	-0.036***	0.025***	0.015***	-0.017***	0.006***	0.009***
	(0.001)	(0.0004)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$SP60^{2m(2-A)\sim}_{i,t}$	-0.017***	-0.001	0.015***	-0.065***	0.010***	0.043***	-0.023***	0.0003	0.019***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
2-mean Algo $2-B$									
$SP10^{2m(2-B)\sim}_{i,t}$	-0.006***	0.004***	0.003***	-0.007***	0.020***	0.009***	-0.010***	0.007***	0.005***
	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0004)
$SP30^{2m(2-B)\sim}_{i,t}$	-0.012***	0.003***	0.006***	-0.038***	0.026***	0.014***	-0.018***	0.006***	0.008***
	(0.001)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$SP60^{2m(2-B)\sim}_{i,t}$	-0.017***	-0.001	0.016***	-0.068***	0.010***	0.044***	-0.024***	0.0002	0.019***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
2-mean Algo $2-C$									
$SP10^{2m(2-C)\sim}_{i,t}$	-0.005***	0.004***	0.003***	-0.007***	0.020***	0.009***	-0.009***	0.007***	0.005***
	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0004)	(0.0004)
$SP30^{2m(2-C)\sim}_{i,t}$	-0.011***	0.003***	0.006***	-0.038***	0.026***	0.014***	-0.018***	0.006***	0.009***
	(0.001)	(0.0004)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$SP60^{2m(2-C)\sim}_{i,t}$	-0.017***	-0.001	0.015***	-0.067***	0.010***	0.043***	-0.023***	0.0004	0.019***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
4-mean Algo $0-A$									
$SP10^{4m(0-A)\sim}_{i,t}$	-0.015***	0.012***	0.010***	-0.023***	0.055***	0.037***	-0.026***	0.019***	0.016***
	(0.002)	(0.001)	(0.001)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
$SP30^{4m(0-A)\sim}_{i,t}$	-0.034***	0.015***	0.029***	-0.105***	0.100***	0.069***	-0.053***	0.025***	0.041***
	(0.003)	(0.002)	(0.003)	(0.007)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)
$SP60^{4m(0-A)\sim}_{i,t}$	-0.060***	-0.002	0.054***	-0.216***	0.029***	0.139***	-0.080***	0.004	0.059***
	(0.004)	(0.003)	(0.005)	(0.013)	(0.008)	(0.013)	(0.006)	(0.005)	(0.006)
4-mean Algo $0-B$									
$SP10^{4m(0-B)\sim}_{i,t}$	-0.018***	0.013***	0.011***	-0.031***	0.063***	0.041***	-0.032***	0.021***	0.018***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.005)	(0.004)	(0.003)	(0.003)	(0.003)
$SP30^{4m(0-B)\sim}_{i,t}$	-0.037***	0.015***	0.025***	-0.117***	0.102***	0.056***	-0.058***	0.026***	0.037***
	(0.003)	(0.003)	(0.003)	(0.009)	(0.008)	(0.007)	(0.005)	(0.004)	(0.005)
$SP60^{4m(0-B)\sim}_{i,t}$	-0.062***	-0.002	0.056***	-0.226***	0.024***	0.134***	-0.082***	0.002	0.058***
	(0.005)	(0.003)	(0.006)	(0.015)	(0.008)	(0.016)	(0.007)	(0.005)	(0.008)
4-mean Algo $0-C$									
$SP10^{4m(0-C)\sim}_{i,t}$	-0.017***	0.013***	0.010***	-0.029***	0.058***	0.040***	-0.030***	0.019***	0.018***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)
$SP30^{4m(0-C)\sim}_{i,t}$	-0.036***	0.015***	0.027***	-0.115***	0.104***	0.062***	-0.056***	0.025***	0.039***
	(0.003)	(0.002)	(0.003)	(0.008)	(0.008)	(0.007)	(0.005)	(0.004)	(0.004)
$SP60^{4m(0-C)\sim}_{i,t}$	-0.062***	-0.001	0.053***	-0.223***	0.028***	0.126***	-0.082***	0.002	0.055***
	(0.005)	(0.003)	(0.006)	(0.015)	(0.009)	(0.014)	(0.007)	(0.006)	(0.007)

Continued on next page

4-mean Algo $1-A$	$\mathbf{VOLX_{\Delta-1}}$	$\mathbf{VOLX_{\Delta+1}}$	$\mathbf{VOLX_{\Delta+2}}$	$\mathbf{QSX_{\Delta-1}}$	$\mathbf{QSX_{\Delta+1}}$	$\mathbf{QSX_{\Delta+2}}$	$\mathbf{ESX_{\Delta-1}}$	$\mathbf{ESX_{\Delta+1}}$	$\mathbf{ESX_{\Delta+2}}$
$SP10_{i,t}^{4m(1-A)\sim}$	-0.014***	0.012***	0.010***	-0.023***	0.056***	0.035***	-0.027***	0.021***	0.014***
	(0.002)	(0.001)	(0.001)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
$SP30_{i,t}^{4m(1-A)\sim}$	-0.035***	0.017***	0.031***	-0.105***	0.103***	0.075***	-0.056***	0.028***	0.044***
	(0.003)	(0.002)	(0.003)	(0.007)	(0.007)	(0.006)	(0.004)	(0.004)	(0.004)
$SP60_{i,t}^{4m(1-A)\sim}$	-0.067***	-0.0003	0.057***	-0.237***	0.024***	0.141***	-0.089***	0.001	0.064***
	(0.005)	(0.004)	(0.004)	(0.013)	(0.008)	(0.011)	(0.006)	(0.006)	(0.006)
4-mean Algo $1-B$									
$SP10_{i,t}^{4m(1-B)\sim}$	-0.017***	0.013***	0.011***	-0.028***	0.062***	0.039***	-0.031***	0.023***	0.015***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)
$SP30_{i,t}^{4m(1-B)\sim}$	-0.037***	0.017***	0.031***	-0.115***	0.110***	0.075***	-0.061***	0.030***	0.045***
	(0.003)	(0.003)	(0.003)	(0.008)	(0.008)	(0.007)	(0.005)	(0.004)	(0.004)
$SP60_{i,t}^{4m(1-B)\sim}$	-0.070***	-0.0002	0.060***	-0.251***	0.025***	0.146***	-0.093***	0.001	0.067***
	(0.005)	(0.004)	(0.005)	(0.015)	(0.0109)	(0.013)	(0.007)	(0.006)	(0.007)
4-mean Algo $1-C$									
$SP10_{i,t}^{4m(1-C)\sim}$	-0.015***	0.013***	0.010***	-0.025***	0.058***	0.038***	-0.029***	0.021***	0.016***
	(0.002)	(0.001)	(0.001)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
$SP30_{i,t}^{4m(1-C)\sim}$	-0.036***	0.017***	0.032***	-0.112***	0.110***	0.076***	-0.058***	0.030***	0.044***
	(0.003)	(0.002)	(0.003)	(0.007)	(0.007)	(0.007)	(0.004)	(0.004)	(0.004)
$SP60_{i,t}^{4m(1-C)\sim}$	-0.069***	0.001	0.059***	-0.247***	0.029***	0.137***	-0.092***	0.002	0.063***
	(0.005)	(0.004)	(0.005)	(0.014)	(0.009)	(0.012)	(0.007)	(0.006)	(0.007)
4-mean Algo $2-A$									
$SP10_{i,t}^{4m(2-A)\sim}$	-0.013***	0.011***	0.010***	-0.020***	0.051***	0.033***	-0.025***	0.018***	0.014***
	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
$SP30_{i,t}^{4m(2-A)\sim}$	-0.030***	0.014***	0.026***	-0.093***	0.088***	0.060***	-0.048***	0.024***	0.036***
	(0.003)	(0.002)	(0.002)	(0.006)	(0.006)	(0.005)	(0.004)	(0.004)	(0.003)
$SP60_{i,t}^{4m(2-A)\sim}$	-0.055***	-0.001	0.045***	-0.196***	0.019***	0.116***	-0.072***	0.002	0.050***
	(0.004)	(0.003)	(0.004)	(0.011)	(0.007)	(0.009)	(0.005)	(0.005)	(0.005)
4-mean Algo $2-B$									
$SP10_{i,t}^{4m(2-B)\sim}$	-0.015***	0.012***	0.010***	-0.025***	0.055***	0.035***	-0.029***	0.020***	0.015***
	(0.002)	(0.001)	(0.001)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
$SP30_{i,t}^{4m(2-B)\sim}$	-0.033***	0.015***	0.024***	-0.101***	0.089***	0.055***	-0.052***	0.024***	0.034***
	(0.003)	(0.003)	(0.002)	(0.007)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)
$SP60_{i,t}^{4m(2-B)\sim}$	-0.057***	-0.001	0.046***	-0.204***	0.017**	0.112***	-0.075***	0.002	0.050***
	(0.004)	(0.003)	(0.004)	(0.012)	(0.007)	(0.011)	(0.005)	(0.005)	(0.006)
4-mean Algo $2-C$									
$SP10_{i,t}^{4m(2-C)\sim}$	-0.014***	0.012***	0.010***	-0.022***	0.052***	0.035***	-0.027***	0.018***	0.015***
	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
$SP30_{i,t}^{4m(2-C)\sim}$	-0.032***	0.015***	0.026***	-0.101***	0.093***	0.059***	-0.051***	0.025***	0.035***
	(0.003)	(0.002)	(0.002)	(0.006)	(0.006)	(0.006)	(0.004)	(0.003)	(0.004)
$SP60_{i,t}^{4m(2-C)\sim}$	-0.056***	0	0.045***	-0.202***	0.021***	0.108***	-0.074***	0.003	0.048***
	(0.004)	(0.003)	(0.004)	(0.011)	(0.007)	(0.010)	(0.005)	(0.005)	(0.006)

TABLE 1.33: The effect of *MeanSP* IV on market quality, algo $SPX_{i,t}^{Ym(Z)\sim}$. Table shows the effect of SP on market quality using *MeanSP* as an instrumental variable (IV). The tables below show the coefficient of the IV of SP. $SPX_{i,t}^{Ym(Z)\sim}$ represents a MeanSP ratio for asset i identified by Z algorithms to all placed orders and subsequently cancelled within $X \in \{10, 30, 60\}$ minute time window and with the requirement of minimum volume for the spoofing order to be $Y \in \{2, 4\}$ mean of the average volume for prevailing five consecutive trading days. Below each coefficient, we show the standard errors. Significance levels are denoted by *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

1.8.14 Appendix 14. Hausman test results

TABLE 1.34: This table shows the results of the Hausman test, where the random effects model is preferred under the null hypothesis, while the fixed effects model is preferred under the alternative. If p-value of the test $\leq$ 0.05, then fixed effects model is preferred, otherwise random effects. Panel A shows results for "2-mean" algorithm modification, Panel B: "4-mean" algorithm modification.

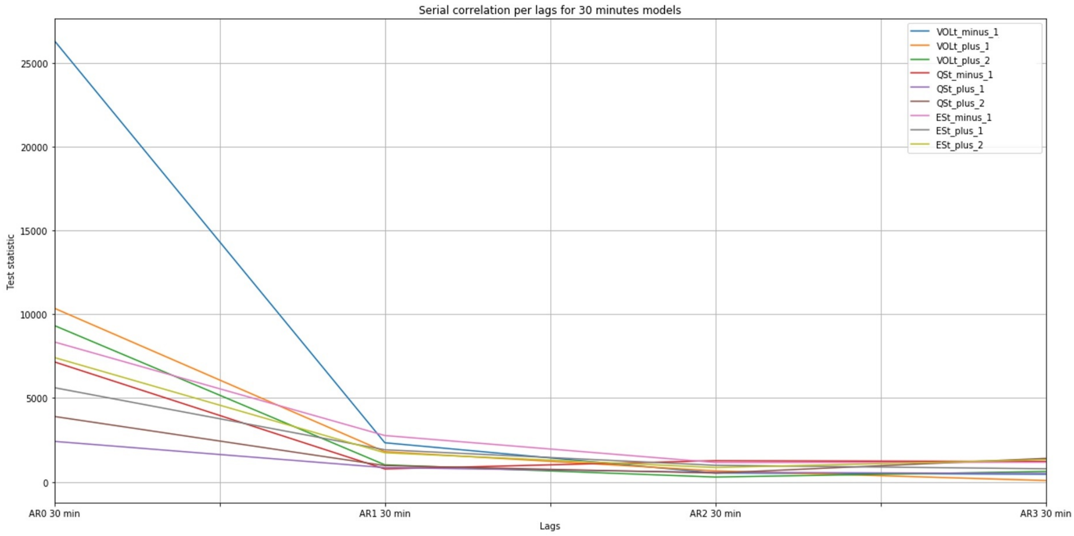
Panel A:	2-mean		$\mathbf{VOLX_{\Delta-1}}$	$\mathbf{VOLX_{\Delta+1}}$	$\mathbf{VOLX_{\Delta+2}}$	$\mathbf{QSX_{\Delta-1}}$	$\mathbf{QSX_{\Delta+1}}$	$\mathbf{QSX_{\Delta+2}}$	$\mathbf{ESX_{\Delta-1}}$	$\mathbf{ESX_{\Delta+1}}$	$\mathbf{ESX_{\Delta+2}}$
0-A	(10 min)	chisq	279.9471	29.4367	365.9061	175.0282	443.2966	478.3429	518.6472	593.1743	1162.6970
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-A	(10 min)	chisq	154.1943	19.2908	2.7442	104.0397	72.5746	320.5488	323.9639	100.2099	105.8782
		p-value	0.0000	0.0017	0.7394	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-A	(60 min)	chisq	60.9112	9.5197	7.3692	59.6234	1483.1649	208.5697	165.4018	59.6269	12.7074
		p-value	0.0000	0.0900	0.1946	0.0000	0.0000	0.0000	0.0000	0.0000	0.0479
0-B	(10 min)	chisq	277.3587	30.8423	360.8053	175.7351	445.0373	485.7075	521.3576	611.2148	1171.8153
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-B	(30 min)	chisq	151.4891	19.7716	2.1570	107.7596	75.5298	64.6286	327.9658	98.1139	109.2490
		p-value	0.0000	0.0014	0.8270	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-B	(60 min)	chisq	58.7527	9.8103	6.6765	62.4804	1474.2159	200.7229	168.9813	62.5106	13.1971
		p-value	0.0000	0.0808	0.2458	0.0000	0.0000	0.0000	0.0000	0.0000	0.0400
0-C	(10 min)	chisq	281.6601	29.1174	362.2826	173.3845	441.4199	477.7559	515.4646	603.9912	1156.6705
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-C	(30 min)	chisq	155.8229	19.2054	2.5128	105.3559	70.9980	78.5747	324.0323	96.1750	105.8194
		p-value	0.0000	0.0018	0.7746	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-C	(60 min)	chisq	61.4837	9.2909	7.1745	60.7245	1499.8623	204.9658	166.5154	57.9939	13.3012
		p-value	0.0000	0.0980	0.2080	0.0000	0.0000	0.0000	0.0000	0.0000	0.0385
1-A	(10 min)	chisq	286.7800	35.0446	486.9196	169.7966	534.2179	434.4795	503.1470	863.6490	1093.6653
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-A	(30 min)	chisq	166.3691	22.4802	9.3580	95.3969	17.3878	59.8553	311.7893	50.1618	93.5009
		p-value	0.0000	0.0004	0.0956	0.0000	0.0080	0.0000	0.0000	0.0000	0.0000
1-A	(60 min)	chisq	77.9608	16.5579	16.7621	47.5753	1699.6839	227.6766	161.6444	21.7483	8.3336
		p-value	0.0000	0.0054	0.0050	0.0000	0.0000	0.0000	0.0000	0.0013	0.2147
1-B	(10 min)	chisq	285.0680	35.8313	469.3982	170.2445	537.2932	437.8978	505.3510	883.6501	1097.6213
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-B	(30 min)	chisq	163.5003	22.6739	8.1498	97.5239	20.0176	50.0207	314.2437	74.7368	95.7675
		p-value	0.0000	0.0004	0.1482	0.0000	0.0028	0.0000	0.0000	0.0000	0.0000
1-B	(60 min)	chisq	74.8083	16.4404	15.1510	48.8010	1681.4740	220.0846	163.4886	23.5540	8.3104
		p-value	0.0000	0.0057	0.0097	0.0000	0.0000	0.0000	0.0000	0.0006	0.2162
1-C	(10 min)	chisq	286.7640	35.1115	486.8582	169.0085	535.0658	434.5071	501.8028	867.7042	1091.6259
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

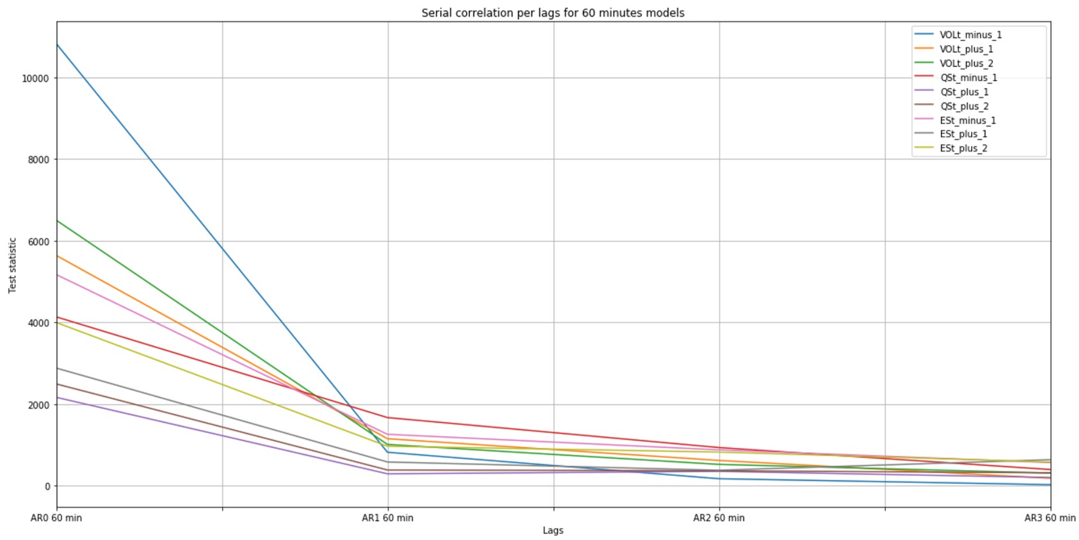
1-C	(30 min)	chisq	166.4219	22.4475	8.6618	95.8836	16.7357	53.9863	311.9122	1280.4925	93.7967
		p-value	0.0000	0.0004	0.1233	0.0000	0.0103	0.0000	0.0000	0.0000	0.0000
1-C	(60 min)	chisq	77.4553	16.3106	16.2656	47.1795	1730.0626	224.0062	161.6536	21.1807	8.1408
		p-value	0.0000	0.0060	0.0061	0.0000	0.0000	0.0000	0.0000	0.0017	0.2280
2-A	(10 min)	chisq	290.5065	30.5247	413.2655	171.5913	467.0777	451.8878	509.9713	672.5111	1127.9772
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-A	(30 min)	chisq	163.0576	20.1277	4.3028	96.8901	45.3405	109.6770	313.1974	84.8634	97.0547
		p-value	0.0000	0.0012	0.5067	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-A	(60 min)	chisq	68.5640	11.3190	8.9054	55.1712	1528.4206	211.0574	160.3800	42.3069	8.0822
		p-value	0.0000	0.0454	0.1129	0.0000	0.0000	0.0000	0.0000	0.0000	0.2321
2-B	(10 min)	chisq	287.4790	31.7557	401.4174	172.3462	470.8362	457.1732	512.7853	691.7832	1134.3459
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-B	(30 min)	chisq	160.0384	20.4821	3.4192	99.7636	48.3147	66.7961	316.5436	82.9307	100.1399
		p-value	0.0000	0.0010	0.6356	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-B	(60 min)	chisq	66.3549	11.4770	8.0225	57.4439	1515.5201	204.3910	163.1824	44.8670	8.7995
		p-value	0.0000	0.0427	0.1550	0.0000	0.0000	0.0000	0.0000	0.0000	0.1852
2-C	(10 min)	chisq	291.2193	30.6023	408.6172	170.6369	468.8206	452.5459	507.9532	683.4407	1124.2887
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-C	(30 min)	chisq	163.2477	20.0439	3.9716	97.7531	44.7006	77.9121	313.4002	80.9344	97.2204
		p-value	0.0000	0.0012	0.5535	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-C	(60 min)	chisq	68.8756	11.0995	8.6233	55.4815	1548.6678	207.8111	160.7671	41.0478	8.5607
		p-value	0.0000	0.0494	0.1251	0.0000	0.0000	0.0000	0.0000	0.0000	0.1998

Panel B:	4-mean		$\mathbf{VOLX}_{\Delta-1}$	$\mathbf{VOLX}_{\Delta+1}$	$\mathbf{VOLX}_{\Delta+2}$	$\mathbf{QSX}_{\Delta-1}$	$\mathbf{QSX}_{\Delta+1}$	$\mathbf{QSX}_{\Delta+2}$	$\mathbf{ESX}_{\Delta-1}$	$\mathbf{ESX}_{\Delta+1}$	$\mathbf{ESX}_{\Delta+2}$
0-A	(10 min)	chisq	269.4449	33.0309	301.1598	179.6339	495.0977	526.1238	531.5282	832.1551	1188.6296
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-A	(10 min)	chisq	136.3795	20.4968	2.0510	116.5595	74.1350	58.9312	348.9947	32.7114	123.2015
		p-value	0.0000	0.0010	0.8420	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-A	(60 min)	chisq	44.2563	9.3770	7.1506	39.4888	1387.2463	183.1556	176.6772	84.9585	24.4052
		p-value	0.0000	0.0949	0.2097	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004
0-B	(10 min)	chisq	268.3946	34.0830	306.3054	180.1194	498.2484	531.9081	532.6046	862.7362	1192.7878
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-B	(30 min)	chisq	134.4507	20.7957	2.2589	120.0345	76.4202	55.7773	351.8787	258.9763	125.0504
		p-value	.0000	0.0009	0.8123	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-B	(60 min)	chisq	43.1627	9.2577	7.0359	43.3704	1393.0735	178.4488	180.3032	86.3887	25.2745
		p-value	0.0000	0.0992	0.2180	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003
0-C	(10 min)	chisq	269.9613	32.7236	290.0485	178.9883	494.7573	532.5912	529.2495	847.9207	1184.6597
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

0-C	(30 min)	chisq	136.0875	20.3680	2.1686	117.8569	74.4003	57.9247	349.0941	1928.2243	122.8346
		p-value	0.0000	0.0011	0.8254	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-C	(60 min)	chisq	43.7804	9.3701	6.9902	40.1402	1394.9386	180.3157	177.2290	81.3038	25.2518
		p-value	0.0000	0.0952	0.2214	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003
1-A	(10 min)	chisq	271.7135	37.6472	390.2049	175.9812	554.2731	498.5177	522.5789	995.4670	1148.8268
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-A	(30 min)	chisq	144.3698	22.0369	3.1310	108.8904	44.9979	64.8608	339.0204	123.4795	111.7341
		p-value	0.0000	0.0005	0.6798	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-A	(60 min)	chisq	50.4197	12.5991	9.1343	16.4468	1435.2913	191.6605	167.4892	47.8278	17.5824
		p-value	0.0000	0.0274	0.1038	0.0115	0.0000	0.0000	0.0000	0.0000	0.0074
1-B	(10 min)	chisq	270.6701	38.1556	382.9665	176.1847	557.2314	500.5363	523.6661	1013.6772	1151.6458
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-B	(30 min)	chisq	142.1206	21.9949	2.9418	110.7082	46.9791	61.9768	340.6485	122.2693	113.6927
		p-value	0.0000	0.0005	0.7090	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-B	(60 min)	chisq	48.9436	12.3142	8.4781	21.5206	1427.5308	186.8163	169.4376	49.1731	18.6493
		p-value	0.0000	0.0307	0.1318	0.0015	0.0000	0.0000	0.0000	0.0000	0.0048
1-C	(10 min)	chisq	271.3842	37.7997	382.3284	175.4827	555.3274	503.6205	521.0191	1001.9988	1148.0919
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-C	(30 min)	chisq	144.6553	21.9528	2.9347	109.0151	43.3324	63.1327	338.5569	115.7853	111.6389
		p-value	0.0000	0.0005	0.7100	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-C	(60 min)	chisq	50.3984	12.2836	8.8630	15.7563	1462.5125	188.6599	167.2148	47.4171	17.5600
		p-value	0.0000	0.0311	0.1147	0.0151	0.0000	0.0000	0.0000	0.0000	0.0074
2-A	(10 min)	chisq	273.0189	35.5909	331.0129	177.1116	526.8452	515.3240	525.9466	935.2927	1161.5959
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-A	(30 min)	chisq	141.9308	21.3324	2.1730	111.1375	57.5126	57.7471	342.1549	151.8453	116.1032
		p-value	0.0000	0.0007	0.8247	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-A	(60 min)	chisq	46.8911	9.7085	7.4816	30.3450	1399.2578	184.1268	170.7925	68.3192	18.9849
		p-value	0.0000	0.0839	0.1872	0.0000	0.0000	0.0000	0.0000	0.0000	0.0042
2-B	(10 min)	chisq	271.5162	36.4356	329.4515	177.5374	531.0697	517.2394	527.2065	960.2032	1164.5472
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-B	(30 min)	chisq	139.6996	21.4338	2.0990	113.8034	59.9115	56.0175	344.4202	146.5669	118.1682
		p-value	0.0000	0.0007	0.8353	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-B	(60 min)	chisq	45.9149	9.4050	7.2478	34.3858	1397.8157	179.9371	173.7344	70.7644	20.3663
		p-value	0.0000	0.0940	0.2029	0.0000	0.0000	0.0000	0.0000	0.0000	0.0024
2-C	(10 min)	chisq	273.2009	35.8024	319.8948	176.6548	529.0453	521.6135	524.4256	950.4563	1160.8101
		p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-C	(30 min)	chisq	141.8398	21.2271	2.1447	111.8669	56.6185	57.5787	342.1013	132.7066	115.9039
		p-value	0.0000	0.0007	0.8288	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-C	(60 min)	chisq	46.9686	9.7329	7.3542	30.4168	1413.5965	181.2267	170.8541	67.0295	19.6566
		p-value	0.0000	0.0832	0.1956	0.0000	0.0000	0.0000	0.0000	0.0000	0.0032

1.8.15 Appendix 15. Breusch-Godfrey test for serial correlation in the models using 30- and 60-minute time windows





1.8.16 Appendix 16. Heteroskedasticity check

TABLE 1.36: Breusch-Godfrey test. This table shows the Breusch-Godfrey test for heteroskedasticity, where the null hypothesis of the test is homoskedasticity. So, if p-value ≤ 0.05, then heteroskedasticity is present in the model (marked in grey). Panel A shows results for "2-mean" algorithm modification, Panel B shows results for "4-mean" algorithm modification.

Panel A:	2-mean		**$VOLX_{\Delta-1}$**	**$VOLX_{\Delta+1}$**	**$VOLX_{\Delta+2}$**	**$QSX_{\Delta-1}$**	**$QSX_{\Delta+1}$**	**$QSX_{\Delta+2}$**	**$ESX_{\Delta-1}$**	**$ESX_{\Delta+1}$**	**$ESX_{\Delta+2}$**
0-A	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6905	0.7296	0.7754	0.7294	0.7449
		p-value	1	1	1	1	1	1	1	1	1
0-A	(30 min)	test-stat	0.8459	0.7781	0.7247	0.7686	0.7150	0.7826	0.7598	0.7417	0.7473
		p-value	1	1	1	1	1	1	1	1	1
0-A	(60 min)	test-stat	1.0275	0.9423	0.8521	0.7475	0.7413	0.7260	0.8556	0.8833	0.8116
		p-value	0.0043	1	1	1	1	1	1	1	1
0-B	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6906	0.7295	0.7754	0.7294	0.7448
		p-value	1	1	1	1	1	1	1	1	1
0-B	(30 min)	test-stat	0.8459	0.7781	0.7247	0.7685	0.7150	0.7825	0.7597	0.7417	0.7473
		p-value	1	1	1	1	1	1	1	1	1
0-B	(60 min)	test-stat	1.0274	0.9423	0.8520	0.7475	0.7413	0.7260	0.8555	0.8833	0.8116
		p-value	0.0043	1	1	1	1	1	1	1	1
0-C	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6905	0.7295	0.7754	0.7294	0.7448
		p-value	1	1	1	1	1	1	1	1	1
0-C	(30 min)	test-stat	0.8459	0.7781	0.7247	0.7686	0.7149	0.7826	0.7598	0.7416	0.7473
		p-value	1	1	1	1	1	1	1	1	1
0-C	(60 min)	test-stat	1.0275	0.9423	0.8521	0.7475	0.7413	0.7260	0.8556	0.8834	0.8116
		p-value	0.0042	1	1	1	1	1	1	1	1
1-A	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7721	0.6906	0.7296	0.7757	0.7294	0.7449
		p-value	1	1	1	1	1	1	1	1	1
1-A	(30 min)	test-stat	0.8460	0.7781	0.7248	0.7688	0.7148	0.7826	0.7603	0.7416	0.7474
		p-value	1	1	1	1	1	1	1	1	1
1-A	(60 min)	test-stat	1.0276	0.9423	0.8520	0.7480	0.7412	0.7261	0.8561	0.8835	0.8118
		p-value	0.0041	1	1	1	1	1	1	1	1
1-B	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7721	0.6906	0.7295	0.7757	0.7294	0.7449
		p-value	1	1	1	1	1	1	1	1	1
1-B	(30 min)	test-stat	0.8460	0.7781	0.7247	0.7688	0.7149	0.7826	0.7602	0.7416	0.7474
		p-value	1	1	1	1	1	1	1	1	1
1-B	(60 min)	test-stat	1.0276	0.9423	0.8520	0.7480	0.7413	0.7261	0.8561	0.8835	0.8118
		p-value	0.0041	1	1	1	1	1	1	1	1
1-C	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7721	0.6906	0.7295	0.7757	0.7294	0.7449
		p-value	1	1	1	1	1	1	1	1	1

1-C	(30 min)	test-stat	0.8460	0.7781	0.7247	0.7689	0.7148	0.7826	0.7603	0.7416	0.7474
		p-value	1	1	1	1	1	1	1	1	1
1-C	(60 min)	test-stat	1.0276	0.9423	0.8520	0.7480	0.7412	0.7261	0.8561	0.8835	0.8117
		p-value	0.0041	1	1	1	1	1	1	1	1
2-A	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6905	0.7296	0.7756	0.7294	0.7449
		p-value	1	1	1	1	1	1	1	1	1
2-A	(30 min)	test-stat	0.8459	0.7781	0.7247	0.7686	0.7149	0.7826	0.7600	0.7416	0.7474
		p-value	1	1	1	1	1	1	1	1	1
2-A	(60 min)	test-stat	1.0274	0.9423	0.8521	0.7476	0.7413	0.7261	0.8557	0.8834	0.8117
		p-value	0.0043	1	1	1	1	1	1	1	1
2-B	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6905	0.7296	0.7755	0.7294	0.7449
		p-value	1	1	1	1	1	1	1	1	1
2-B	(30 min)	test-stat	0.8459	0.7781	0.7247	0.7686	0.7149	0.7825	0.7600	0.7416	0.7474
		p-value	1	1	1	1	1	1	1	1	1
2-B	(60 min)	test-stat	1.0274	0.9423	0.8520	0.7476	0.7414	0.7261	0.8557	0.8834	0.8117
		p-value	0.0044	1	1	1	1	1	1	1	1
2-C	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7721	0.6905	0.7296	0.7756	0.7294	0.7449
		p-value	1	1	1	1	1	1	1	1	1
2-C	(30 min)	test-stat	0.8459	0.7781	0.7247	0.7686	0.7149	0.7826	0.7600	0.7416	0.7474
		p-value	1	1	1	1	1	1	1	1	1
2-C	(60 min)	test-stat	1.0274	0.9423	0.8520	0.7476	0.7413	0.7261	0.8556	0.8834	0.8116
		p-value	0.0043	1	1	1	1	1	1	1	1

Panel B:	4 mean		**VOLX$_{\Delta-1}$**	**VOLX$_{\Delta+1}$**	**VOLX$_{\Delta+2}$**	**QSX$_{\Delta-1}$**	**QSX$_{\Delta+1}$**	**QSX$_{\Delta+2}$**	**ESX$_{\Delta-1}$**	**ESX$_{\Delta+1}$**	**ESX$_{\Delta+2}$**
0-A	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6904	0.7294	0.7755	0.7293	0.7447
		p-value	1	1	1	1	1	1	1	1	1
0-A	(30 min)	test-stat	0.8461	0.7781	0.7247	0.7689	0.7145	0.7825	0.7601	0.7415	0.7472
		p-value	1	1	1	1	1	1	1	1	1
0-A	(60 min)	test-stat	1.0278	0.9423	0.8521	0.7482	0.7407	0.7261	0.8559	0.8830	0.8114
		p-value	0.0039	1	1	1	1	1	1	1	1
0-B	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6905	0.7294	0.7755	0.7293	0.7447
		p-value	1	1	1	1	1	1	1	1	1
0-B	(30 min)	test-stat	0.8461	0.7780	0.7247	0.7689	0.7146	0.7824	0.7601	0.7414	0.7472
		p-value	1	1	1	1	1	1	1	1	1
0-B	(60 min)	test-stat	1.0277	0.9423	0.8520	0.7482	0.7407	0.7261	0.8558	0.8830	0.8115
		p-value	0.0039	1	1	1	1	1	1	1	1
0-C	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6904	0.7294	0.7755	0.7293	0.7447
		p-value	1	1	1	1	1	1	1	1	1

0-C	(30 min)	test-stat	0.8461	0.7780	0.7247	0.7689	0.7145	0.7825	0.7601	0.7414	0.7472
		p-value	1	1	1	1	1	1	1	1	1
0-C	(60 min)	test-stat	1.0278	0.9423	0.8520	0.7481	0.7407	0.7261	0.8558	0.8830	0.8114
		p-value	0.0039	1	1	1	1	1	1	1	1
1-A	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7721	0.6904	0.7294	0.7756	0.7293	0.7447
		p-value	1	1	1	1	1	1	1	1	1
1-A	(30 min)	test-stat	0.8461	0.7781	0.7247	0.7691	0.7144	0.7825	0.7603	0.7414	0.7473
		p-value	1	1	1	1	1	1	1	1	1
1-A	(60 min)	test-stat	1.0279	0.9423	0.8520	0.7486	0.7405	0.7261	0.8562	0.8831	0.8114
		p-value	0.0037	1	1	1	1	1	1	1	1
1-B	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7721	0.6905	0.7294	0.7756	0.7293	0.7447
		p-value	1	1	1	1	1	1	1	1	1
1-B	(30 min)	test-stat	0.8461	0.7781	0.7247	0.7691	0.7144	0.7825	0.7603	0.7414	0.7473
		p-value	1	1	1	1	1	1	1	1	1
1-B	(60 min)	test-stat	1.0279	0.9423	0.8519	0.7486	0.7405	0.7261	0.8561	0.8831	0.8114
		p-value	0.0038	1	1	1	1	1	1	1	1
1-C	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7721	0.6904	0.7294	0.7756	0.7293	0.7447
		p-value	1	1	1	1	1	1	1	1	1
1-C	(30 min)	test-stat	0.8461	0.7781	0.7247	0.7692	0.7144	0.7825	0.7603	0.7414	0.7473
		p-value	1	1	1	1	1	1	1	1	1
1-C	(60 min)	test-stat	1.0279	0.9423	0.8520	0.7486	0.7405	0.7261	0.8561	0.8831	0.8113
		p-value	0.0037	1	1	1	1	1	1	1	1
2-A	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6904	0.7294	0.7755	0.7293	0.7447
		p-value	1	1	1	1	1	1	1	1	1
2-A	(30 min)	test-stat	0.8461	0.7780	0.7247	0.7690	0.7145	0.7825	0.7603	0.7414	0.7473
		p-value	1	1	1	1	1	1	1	1	1
2-A	(60 min)	test-stat	1.0278	0.9423	0.8520	0.7483	0.7407	0.7262	0.8559	0.8831	0.8115
		p-value	0.0039	1	1	1	1	1	1	1	1
2-B	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6904	0.7294	0.7755	0.7293	0.7447
		p-value	1	1	1	1	1	1	1	1	1
2-B	(30 min)	test-stat	0.8461	0.7780	0.7247	0.7690	0.7145	0.7825	0.7603	0.7414	0.7473
		p-value	1	1	1	1	1	1	1	1	1
2-B	(60 min)	test-stat	1.0278	0.9423	0.8520	0.7483	0.7407	0.7262	0.8559	0.8831	0.8115
		p-value	0.0039	1	1	1	1	1	1	1	1
2-C	(10 min)	test-stat	0.7315	0.7236	0.7160	0.7720	0.6904	0.7294	0.7755	0.7293	0.7447
		p-value	1	1	1	1	1	1	1	1	1
2-C	(30 min)	test-stat	0.8461	0.7780	0.7247	0.7690	0.7145	0.7825	0.7603	0.7414	0.7473
		p-value	1	1	1	1	1	1	1	1	1
2-C	(60 min)	test-stat	1.0278	0.9423	0.8520	0.7482	0.7407	0.7262	0.8559	0.8831	0.8115
		p-value	0.0039	1	1	1	1	1	1	1	1

1.8.17 Appendix 17. Sargan-Hansen J-statistic

TABLE 1.38: Sargan-Hansen J-statistic results. This table shows Sargan-Hansen J-statistic to test if SPt is endogenous, where the null hypothesis of the test is exogeneity (marked in grey). Panel A shows results for "2-mean" algorithm modification, Panel B shows results for "4-mean" algorithm modification. If p-value $\leq$ 0.05, SPt is endogenous.

Panel 1:	2-mean		$\mathbf{VOLX_{\Delta-1}}$	$\mathbf{VOLX_{\Delta+1}}$	$\mathbf{VOLX_{\Delta+2}}$	$\mathbf{QSX_{\Delta-1}}$	$\mathbf{QSX_{\Delta+1}}$	$\mathbf{QSX_{\Delta+2}}$	$\mathbf{ESX_{\Delta-1}}$	$\mathbf{ESX_{\Delta+1}}$	$\mathbf{ESX_{\Delta+2}}$
0-A	(10 min)	test-stat	3.183	13.438	3.255	2.898	36.591	13.548	11.667	14.409	7.774
		p-value	0.074	0.000	0.071	0.089	0.000	0.000	0.001	0.000	0.005
0-A	(30 min)	test-stat	0.065	0.220	0.016	0.750	0.059	0.718	1.008	0.268	1.648
		p-value	0.799	0.639	0.899	0.387	0.809	0.397	0.316	0.605	0.199
0-A	(60 min)	test-stat	11.225	0.187	12.191	29.484	1.678	18.304	15.674	0.001	8.561
		p-value	0.001	0.665	0.001	0.000	0.195	0.000	0.000	0.973	0.003
0-B	(10 min)	test-stat	4.994	7.028	3.139	2.248	39.519	10.822	14.863	18.520	5.131
		p-value	0.025	0.008	0.076	0.134	0.000	0.001	0.000	0.000	0.024
0-B	(30 min)	test-stat	0.003	0.223	0.008	0.510	0.027	0.420	0.717	0.397	1.241
		p-value	0.960	0.637	0.929	0.475	0.869	0.517	0.397	0.529	0.265
0-B	(60 min)	test-stat	12.737	0.125	13.346	31.938	1.721	19.181	17.113	0.010	8.938
		p-value	0.000	0.723	0.000	0.000	0.190	0.000	0.000	0.922	0.003
0-C	(10 min)	test-stat	6.250	6.641	1.621	3.686	26.872	8.062	14.327	6.638	3.641
		p-value	0.012	0.010	0.203	0.055	0.000	0.005	0.000	0.010	0.056
0-C	(30 min)	test-stat	0.047	0.169	0.005	0.673	0.123	0.430	0.933	0.496	1.230
		p-value	0.828	0.681	0.942	0.412	0.726	0.512	0.334	0.482	0.267
0-C	(60 min)	test-stat	12.289	0.158	12.987	31.634	1.664	18.969	17.268	0.003	8.869
		p-value	0.001	0.691	0.000	0.000	0.197	0.000	0.000	0.957	0.003
1-A	(10 min)	test-stat	8.379	7.683	15.585	3.342	32.881	19.665	12.625	7.054	14.259
		p-value	0.004	0.006	0.000	0.068	0.000	0.000	0.000	0.008	0.000
1-A	(30 min)	test-stat	0.223	0.535	0.203	2.873	2.831	3.540	2.218	0.903	5.436
		p-value	0.637	0.464	0.652	0.090	0.092	0.060	0.137	0.342	0.020
1-A	(60 min)	test-stat	3.229	0.314	1.743	8.429	1.657	5.820	5.134	0.341	2.834
		p-value	0.072	0.575	0.187	0.004	0.198	0.016	0.024	0.559	0.092
1-B	(10 min)	test-stat	8.094	9.287	5.102	4.192	39.307	13.239	18.305	10.411	10.663
		p-value	0.004	0.002	0.024	0.041	0.000	0.000	0.000	0.001	0.001
1-B	(30 min)	test-stat	0.264	0.661	0.104	3.005	2.729	3.875	2.165	0.845	5.682
		p-value	0.608	0.416	0.748	0.083	0.099	0.049	0.141	0.358	0.017
1-B	(60 min)	test-stat	4.094	0.273	2.361	10.643	1.912	7.360	6.081	0.319	3.534
		p-value	0.043	0.602	0.124	0.001	0.167	0.007	0.014	0.572	0.060
1-C	(10 min)	test-stat	2.263	5.751	0.588	6.472	25.071	4.735	8.674	4.344	2.672
		p-value	0.133	0.017	0.443	0.011	0.000	0.030	0.003	0.037	0.102

1-C	(30 min)	test-stat	0.304	0.788	0.371	3.754	3.729	4.361	2.713	1.078	6.208
		p-value	0.581	0.375	0.543	0.053	0.054	0.037	0.100	0.299	0.013
1-C	(60 min)	test-stat	3.329	0.318	1.776	8.686	1.543	6.029	5.367	0.261	2.832
		p-value	0.068	0.573	0.183	0.003	0.214	0.014	0.021	0.610	0.092
2-A	(10 min)	test-stat	4.906	7.801	7.953	2.301	32.644	13.326	11.721	12.476	8.282
		p-value	0.027	0.005	0.005	0.129	0.000	0.000	0.001	0.000	0.004
2-A	(30 min)	test-stat	1.611	0.399	1.022	6.673	2.446	2.410	5.560	1.324	3.841
		p-value	0.204	0.528	0.312	0.010	0.118	0.121	0.018	0.250	0.050
2-A	(60 min)	test-stat	7.934	0.197	8.509	22.882	1.527	14.002	12.153	0.034	6.080
		p-value	0.005	0.657	0.004	0.000	0.217	0.000	0.001	0.855	0.014
2-B	(10 min)	test-stat	5.513	10.958	4.475	1.901	28.648	8.572	13.158	10.776	5.356
		p-value	0.019	0.001	0.034	0.168	0.000	0.003	0.000	0.001	0.021
2-B	(30 min)	test-stat	1.401	0.509	0.575	6.346	2.366	1.724	4.916	1.569	3.169
		p-value	0.237	0.476	0.448	0.012	0.124	0.189	0.027	0.210	0.075
2-B	(60 min)	test-stat	8.987	0.146	9.147	24.874	1.628	14.930	13.403	0.053	6.400
		p-value	0.003	0.703	0.003	0.000	0.202	0.000	0.000	0.818	0.011
2-C	(10 min)	test-stat	0.610	4.463	1.359	5.138	26.402	9.700	15.259	8.979	3.119
		p-value	0.435	0.035	0.244	0.023	0.000	0.002	0.000	0.003	0.077
2-C	(30 min)	test-stat	1.746	0.376	0.713	6.923	2.878	1.972	5.622	1.556	3.472
		p-value	0.186	0.540	0.399	0.009	0.090	0.160	0.018	0.212	0.062
2-C	(60 min)	test-stat	8.364	0.203	8.637	23.748	1.535	14.097	13.122	0.047	6.036
		p-value	0.004	0.652	0.003	0.000	0.215	0.000	0.000	0.829	0.014

Panel B:	4-mean		**VOLX$_{\Delta-1}$**	**VOLX$_{\Delta+1}$**	**VOLX$_{\Delta+2}$**	**QSX$_{\Delta-1}$**	**QSX$_{\Delta+1}$**	**QSX$_{\Delta+2}$**	**ESX$_{\Delta-1}$**	**ESX$_{\Delta+1}$**	**ESX$_{\Delta+2}$**
0-A	(10 min)	test-stat	5.880	15.816	12.169	13.929	39.916	22.081	22.822	14.929	10.815
		p-value	0.015	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001
0-A	(30 min)	test-stat	14.395	12.889	16.552	30.385	36.369	17.244	28.907	13.368	21.497
		p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0-A	(60 min)	test-stat	0.460	0.839	0.003	0.412	0.265	0.004	0.882	0.120	0.273
		p-value	0.498	0.360	0.960	0.521	0.607	0.952	0.348	0.730	0.601
0-B	(10 min)	test-stat	21.769	18.525	15.016	19.960	76.591	62.488	41.802	31.648	26.764
		p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0-B	(30 min)	test-stat	14.409	13.461	12.344	33.557	36.623	11.710	31.937	14.409	17.392
		p-value	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000
0-B	(60 min)	test-stat	0.085	0.682	0.108	1.051	0.100	0.165	1.609	0.204	0.583
		p-value	0.771	0.409	0.742	0.305	0.751	0.685	0.205	0.652	0.445
0-C	(10 min)	test-stat	14.091	14.541	9.593	16.128	34.024	32.587	25.668	8.633	14.688
		p-value	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.003	0.000

0-C	(30 min)	test-stat	13.836	12.039	14.118	33.247	38.917	14.539	31.212	14.067	19.093
		p-value	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0-C	(60 min)	test-stat	0.141	0.696	0.050	0.936	0.179	0.130	1.502	0.118	0.539
		p-value	0.707	0.404	0.823	0.333	0.672	0.718	0.220	0.732	0.463
1-A	(10 min)	test-stat	12.485	5.319	16.726	8.049	21.697	29.005	15.504	7.000	13.271
		p-value	0.000	0.021	0.000	0.005	0.000	0.000	0.000	0.008	0.000
1-A	(30 min)	test-stat	5.376	2.036	3.544	11.533	27.406	8.768	9.171	10.432	11.327
		p-value	0.020	0.154	0.060	0.001	0.000	0.003	0.003	0.001	0.001
1-A	(60 min)	test-stat	6.383	0.165	3.820	4.516	1.299	4.201	1.143	0.138	1.420
		p-value	0.012	0.685	0.051	0.034	0.254	0.040	0.285	0.711	0.233
1-B	(10 min)	test-stat	12.205	12.330	13.498	14.236	43.160	29.775	22.687	15.414	14.677
		p-value	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1-B	(30 min)	test-stat	6.589	2.388	3.526	14.153	29.281	10.823	10.526	10.398	12.988
		p-value	0.010	0.122	0.060	0.000	0.000	0.001	0.001	0.001	0.000
1-B	(60 min)	test-stat	4.522	0.169	2.554	2.691	0.971	2.504	0.499	0.158	0.766
		p-value	0.034	0.681	0.110	0.101	0.324	0.114	0.480	0.691	0.382
1-C	(10 min)	test-stat	2.108	5.894	5.218	7.338	23.155	8.971	6.600	5.627	3.844
		p-value	0.147	0.015	0.022	0.007	0.000	0.003	0.010	0.018	0.050
1-C	(30 min)	test-stat	6.025	2.624	5.016	15.082	31.576	11.113	11.197	11.562	13.681
		p-value	0.014	0.105	0.025	0.000	0.000	0.001	0.001	0.001	0.000
1-C	(60 min)	test-stat	5.640	0.129	3.404	3.798	1.538	3.151	0.877	0.085	1.084
		p-value	0.018	0.720	0.065	0.051	0.215	0.076	0.349	0.771	0.298
2-A	(10 min)	test-stat	13.165	13.831	19.524	17.089	59.004	36.464	28.037	22.939	20.084
		p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2-A	(30 min)	test-stat	9.373	8.874	12.893	23.436	24.517	11.050	23.102	9.080	14.284
		p-value	0.002	0.003	0.000	0.000	0.000	0.001	0.000	0.003	0.000
2-A	(60 min)	test-stat	0.258	0.820	0.119	1.363	0.046	0.229	2.149	0.349	0.859
		p-value	0.612	0.365	0.730	0.243	0.830	0.632	0.143	0.555	0.354
2-B	(10 min)	test-stat	22.235	31.397	17.794	23.591	81.619	57.421	49.956	29.947	24.533
		p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2-B	(30 min)	test-stat	9.824	10.738	11.029	26.905	27.158	9.512	25.272	10.677	12.990
		p-value	0.002	0.001	0.001	0.000	0.000	0.002	0.000	0.001	0.000
2-B	(60 min)	test-stat	0.006	0.691	0.509	2.651	0.014	0.787	3.474	0.366	1.394
		p-value	0.937	0.406	0.476	0.104	0.905	0.375	0.062	0.545	0.238
2-C	(10 min)	test-stat	10.543	17.696	11.923	14.111	64.703	50.978	28.192	21.898	13.911
		p-value	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
2-C	(30 min)	test-stat	9.408	9.228	11.657	25.442	27.228	10.742	24.835	9.988	13.476
		p-value	0.002	0.002	0.001	0.000	0.000	0.001	0.000	0.002	0.000
2-C	(60 min)	test-stat	0.057	0.728	0.328	2.228	0.029	0.602	3.063	0.310	1.193
		p-value	0.811	0.394	0.567	0.136	0.866	0.438	0.080	0.578	0.275

1.8.18 Appendix 18. Pesaran-Shin (IPS) unit root test

Variable	Time window	Test statistic	P-value
$\mathbf{VOLX_{\Delta}}$	10	-701.52	0.00
	30	-443.96	0.00
	60	-309.73	0.00
$\mathbf{QSX_{\Delta}}$	10	-601.12	0.00
	30	-393.31	0.00
	60	-290.10	0.00
$\mathbf{ESX_{\Delta}}$	10	-618.83	0.00
	30	-399.08	0.00
	60	-296.65	0.00

TABLE 1.40: Pesaran-Shin (IPS) unit root test results for the dependent variables $\mathbf{VOLX_{\Delta}}$, $\mathbf{QSX_{\Delta}}$ and $\mathbf{ESX_{\Delta}}$. If p-value ≤ 0.05 the variable is stationary.

1.8.19 Appendix 19. The effect of SP on total market quality changes

Panel A: 2-mean

	VOL_T	QS_T	ES_T	VOL_T	QS_T	ES_T	VOL_T	QS_T	ES_T
Algo	**10-min**			**30-min**			**60-min**		
0-A	0.0001***	0.001***	0.001***	0.0001*	0.002***	0.001***	0.0002**	0.002***	0.001***
	(0.00003)	(0.0001)	(0.00004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)
0-B	0.0001***	0.001***	0.001***	0.00005	0.001***	0.001***	0.0002*	0.002***	0.001***
	(0.00003)	(0.0001)	(0.00004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)
0-C	0.0001***	0.001***	0.001***	0.0001	0.002***	0.001***	0.0003***	0.002***	0.001***
	(0.00003)	(0.0001)	(0.00004)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)
1-A	0.0002***	0.001***	0.001***	0.0004***	0.002***	0.001***	0.001***	0.002***	0.001***
	(0.00003)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0003)	(0.0002)
1 -B	0.0002***	0.001***	0.001***	0.0003***	0.002***	0.001***	0.0005***	0.002***	0.001***
	(0.00003)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0003)	(0.0002)
1-C	0.0002***	0.001***	0.001***	0.0004***	0.002***	0.001***	0.001***	0.002***	0.001***
	(0.00003)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0003)	(0.0002)
2-A	0.0002***	0.001***	0.001***	0.0002***	0.002***	0.001***	0.0003***	0.002***	0.001***
	(0.00003)	(0.0001)	(0.00004)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0003)	(0.0002)
2 -B	0.0001***	0.001***	0.001***	0.0002***	0.002***	0.001***	0.0003***	0.002***	0.001***
	(0.00003)	(0.0001)	(0.00005)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0003)	(0.0002)
2 -C	0.0002***	0.001***	0.001***	0.0002***	0.002***	0.001***	0.0003***	0.002***	0.001***
	(0.00003)	(0.0001)	(0.00005)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0003)	(0.0002)

Panel B: 4-mean

	VOL_T	QS_T	ES_T	VOL_T	QS_T	ES_T	VOL_T	QS_T	ES_T
Algo	**10-min**			**30-min**			**60-min**		
0-A	0.0001*	0.001***	0.001***	0.00005	0.002***	0.001***	0.0002	0.002***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.0003)
0-B	0.00001	0.001***	0.001***	-0.0001	0.002***	0.001***	0.00003	0.002***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.0004)
0-C	0.00003	0.001***	0.001***	-0.00003	0.002***	0.001***	0.0002	0.002***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.0004)
1-A	0.0002***	0.001***	0.001***	0.001***	0.003***	0.002***	0.001**	0.003***	0.002***
	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.0004)
1-B	0.0001*	0.001***	0.001***	0.0004**	0.003***	0.002***	0.0005*	0.003***	0.002***
	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.0004)
1-C	0.0002**	0.001***	0.001***	0.0004***	0.003***	0.002***	0.001**	0.003***	0.002***
	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0004)	(0.0003)	(0.0003)	(0.001)	(0.0004)
2-A	0.0001**	0.001***	0.001***	0.0002*	0.002***	0.001***	0.0001	0.002***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.0004)
2-B	0.0001	0.001***	0.001***	0.0002	0.002***	0.001***	-0.00000	0.002***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.0004)
2-C	0.0001	0.001***	0.001***	0.0002	0.002***	0.001***	0.0001	0.002***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.001)	(0.0004)

TABLE 1.42: The table shows the baseline results of the effect of SP on total market quality changes. We present regression coefficients for the panel regressions (Equation 1.6, 1.7, 1.8) on the change in short-term volatility (VOL_T), quoted spread (QS_T), and effective spread (ES_T) across the periods from (t) to ($t+2$). Panel A shows results for "2-mean" algorithm modification, Panel B shows results for "4-mean" algorithm modification. Columns "10-min", "30-min", and "60-min" show the result for the Sp modification measured in 10, 30- and 60-minutes time windows respectively. Below each coefficient, we show the standard errors in parenthesis. Significance levels are denoted by $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

1.8.20 Appendix 20. Instrumental variable correlation test

Panel A: 2-mean										
Algo	**Time window**	**$VOLX_{\Delta-1}$**	**$VOLX_{\Delta+1}$**	**$VOLX_{\Delta+2}$**	**$QSX_{\Delta-1}$**	**$QSX_{\Delta+1}$**	**$QSX_{\Delta+2}$**	**$ESX_{\Delta-1}$**	**$ESX_{\Delta+1}$**	**$ESX_{\Delta+2}$**
0-A	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-A	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-A	60 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-B	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-B	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-B	60 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-C	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-C	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-C	60 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-A	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-A	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-A	60 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-B	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-B	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-B	60 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-C	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-C	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-C	60 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-A	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-A	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-A	60 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-B	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-B	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-B	60 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-C	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-C	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-C	60 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Continued on next page

Panel B: 4-mean										
Algo	**Time window**	**$VOLX_{\Delta-1}$**	**$VOLX_{\Delta+1}$**	**$VOLX_{\Delta+2}$**	**$QSX_{\Delta-1}$**	**$QSX_{\Delta+1}$**	**$QSX_{\Delta+2}$**	**$ESX_{\Delta-1}$**	**$ESX_{\Delta+1}$**	**$ESX_{\Delta+2}$**
0-A	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-A	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-A	60 min	0.0000	0.0004	0.0000	0.0000	0.0008	0.0000	0.0000	0.0006	0.0000
0-B	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-B	30 min	0.0000	0.0002	0.0001	0.0001	0.0003	0.0002	0.0001	0.0003	0.0001
0-B	60 min	0.0001	0.0016	0.0021	0.0001	0.0026	0.0033	0.0001	0.0020	0.0025
0-C	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0-C	30 min	0.0000	0.0001	0.0000	0.0000	0.0001	0.0001	0.0000	0.0001	0.0000
0-C	60 min	0.0000	0.0016	0.0007	0.0001	0.0027	0.0013	0.0001	0.0021	0.0009
1-A	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-A	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-A	60 min	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000
1-B	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-B	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-B	60 min	0.0000	0.0001	0.0000	0.0000	0.0002	0.0000	0.0000	0.0001	0.0000
1-C	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-C	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1-C	60 min	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000
2-A	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-A	30 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-A	60 min	0.0000	0.0003	0.0000	0.0000	0.0005	0.0000	0.0000	0.0004	0.0000
2-B	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-B	30 min	0.0000	0.0001	0.0000	0.0000	0.0002	0.0001	0.0000	0.0002	0.0001
2-B	60 min	0.0000	0.0009	0.0003	0.0000	0.0013	0.0005	0.0000	0.0011	0.0004
2-C	10 min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2-C	30 min	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000
2-C	60 min	0.0000	0.0008	0.0001	0.0000	0.0011	0.0002	0.0000	0.0009	0.0001

TABLE 1.43: Instrumental variable correlation test results. The table shows the p-values of regression coefficients for the variable $MeanSP$, which indicated the correlation between the instrumental variable ($MeanSP$) and the variable being instrumented (SP). Panel A shows results for "2-mean" algorithm modification, Panel B shows results for "4-mean" algorithm modification. If the p-value is less than 0.05, the coefficient is significant.

Chapter 2

Forecasting Financial Market Manipulation using Machine Learning Methods

Parts of this chapter have been presented by Tatiana Franus on:

- June 2023, Finance and Business Analytics Conference, Lefkada, Greece.
- June 2023, International Symposium of Forecasting, Charlottesville, Virginia, USA.
- August 2023, Wold Finance Conference, Arger, Norway.

2.1 Introduction

Electronic markets, the automation of trading, and high-speed trading have transformed the financial market landscape and introduced new possibilities for disruptive practices, when traders make tremendous profits by artificially affecting market beliefs and negatively affecting other market participants. Market manipulation strategies represent a source of price distortion and the creation of artificial market conditions. They significantly threaten the trust and integrity of capital markets through mispricing and market imperfections. They harm investors' confidence, resulting in less participation and hence may adversely affect efficiency, liquidity, integrity, and development of the stock market (Guiso et al. 2008, Imisiker & Tas 2013, Punniyamoorthy & Thoppan 2013). In 2017, Citigroup was fined $25 million in a spoofing case for manipulating the U. S. Treasury futures market. In 2020, U. S. regulators levied a $920 million fine on JP Morgan Chase for eight years of spoofing manipulation in markets for precious metals and treasury bills.

The form of manipulation Citigroup and JP Morgan Chase committed is called "spoofing," which involves placing and cancelling orders to create an artificial supply or demand. Our research focuses on spoofing as a widespread type of trade-based manipulation. Significant resources have been invested in automated surveillance systems to detect price manipulating behaviours, especially since the Dodd-Frank Act made spoofing illegal (Dodd-Frank (2010)). In 2018, the CFTC created a Task Force on spoofing to target this sort of misconduct, indicating that this type of market manipulation is of high practical importance. Interestingly, the responsibility for detecting manipulation lies with those who trade. All firms that engage in trading must have monitoring in place to check their activity for signs of market manipulation. This gave rise to a sizable and fast-growing "trade surveillance" industry that aims to monitor the clients' trades and detect illegal trading activity to improve the quality of financial markets (Cumming et al. 2011, Aitken, de B Harris & Ji 2015).

However, while financial exchange markets surveillance has attracted much attention across different markets, the lack of an effective detection and prediction algorithm of spoofing manipulation causes challenges for regulators in their ability to monitor huge amounts of trading activities in real time.

This paper provides a threefold contribution to the literature on spoofing manipulation in financial markets: it develops a novel a data-driven approach to forecast the market state when spoofing manipulation is likely to appear, it identifies its link to market state features and order book variables, and it proposes an effective spoofing prediction measure, the Real-Time Spoofing Probability (RTSP), based on several machine learning algorithms. The validity of our approach is then tested on a unique data set of suspected spoofing cases detected by the Moscow Exchange (MOEX).

Our research is related to a small but growing body of literature addressing trade-based manipulation issues. Aggarwal & Wu (2006) empirically show that manipulation is associated with higher stock volatility, greater liquidity, and high returns during the manipulation period. Lee et al. (2013) conduct a comprehensive empirical study of spoofing using data from the Korean Exchange. Using a particular type of spoofing strategy, they show that spoofing leads to substantial extra profits. Their results suggest that stocks with higher volatility of returns, lower market capitalization, lower price levels, and lower marginal transparency are more prone to spoofing. Recent years have witnessed a growing interest in the literature on the application of machine learning algorithms to identify market manipulation. Öğüt et al. (2009) use artificial neural networks and support vector machines to detect stocks' manipulation in the Istanbul Stock Exchange. Diaz et al. (2011) introduced the data set that was further used by Golmohammadi & Zaiane (2015) in their extensive experiments, where they adopted different supervised learning algorithms. In our research, we make a step further and seek a realistic approach to detect market state with spoofing manipulation event by identifying all possible features that could describe the market state at a specific time, including self-designed features like order book filling ratio and order book spread (Section 2.2.2). We sort the features into three groups: market quality variables, variables related to trades and their frequencies, and features that describe the order book state. We propose a complexity reduction approach to identify, out of hundred variables, those important to detect and forecast the preferable market state for the manipulator to enter the order book in the next tick. This data-driven approach allows us to capture the main economic forces present in limit order markets using a large number of variables, while keeping the analysis tractable by state-of-art machine learning algorithms.

Cao et al. (2014) also use machine learning algorithms to detect price manipulation. Their results show that k-Nearest Neighbour and One-Class Support Vector Machine effectively detect spoofing behaviour. Cao et al. (2015) use Adaptive Hidden Markov Model compared to standard machine learning algorithms such as Gaussian Mixture Models, kNN to identify manipulation activities from NASDAQ and the London Stock Exchange. Their results show that Adaptive Hidden Markov Model is preferable for detection. Both studies test models on synthetically generated data. Martínez-Miranda et al. (2016) use a reinforcement learning framework within the full and partial observability of Markov decision processes and analyse the underlying behaviour of the manipulator.

Another problem is the dependence of the results on training data, as it could be challenging to attain enough trading data for supervised learning. Leangarun et al. (2016) focus on detecting fraudulent activities as spoofing and pump-and-dumping feed forward neural network on labelled level 1 data from NASDAQ. While their neural network model achieves 88.28% in pump-and-dump detection, it fails to model spoof trading effectively. The reason is that level 1 data does not contain enough information. We address this issue by using a dataset with complete historical information of orders placed on the market and show that a limit order book is a valuable data source for detecting spoofing.

Tuccella et al. (2021) use Gated Recurrent Unit (GRU)-based architecture for spoofing detection on several cryptocurrency exchanges, while Tao et al. (2022) proposes a quantification instrument to monitor real-time spoofing behaviour on the TMX market using a conditional Wasserstein distance. However, both studies identify spoofing orders using their algorithm, which questions the robustness of the detection method. In contrast, in this paper, we train ML models on a dataset of suspicious spoofing orders identified by the Moscow Exchange. We show that combining machine learning techniques and publicly available data of orders and trades can effectively substitute the privileged data of traders' IDs, trading styles, etc.

In practice, spoofing can take many forms, but all involve placing and cancelling limit orders. Moreover, spoofing orders occur not far from the bid and ask prices inside the order book as manipulators want to be seen by others as genuine informed traders. However, the current literature on manipulation detection uses level 2 with the depth up to 5th level inside the order book. We address this limitation and seek a more realistic approach using our dataset with up to 50 levels inside the order book to identify those market states when spoofing is more likely. We specify all possible features that could describe the market state at a specific time, including self-designed features like order book filling ratio and order book spread (Section 2.2.2). We sort the features into three groups: market quality variables, variables related to trades and their frequencies, and features that describe the order book state. We propose a complexity reduction approach to identify, out of a hundred variables, those essential to detect and forecast the preferable market state for the potential manipulator to enter the order book in the next tick. This data-driven approach allows us to capture the main economic forces present in limit order markets using a large number of variables while keeping the analysis tractable by state-of-art machine learning algorithms.

We expand our methodology to introduce a forecasting measure of predicting market states with spoofing manipulation events. Our RTSP measure indicates a risk when intraday manipulative activity is highly probable in real-time as the trading process continues. We demonstrate how regulators, exchanges and other data vendors may denote a market state with potential spoofing manipulation practices. So, exchanges could avoid the harmful influence of disruptive activities and thus improve the market quality by implementing our RTSP measure in their surveillance systems.

The proposed methodology is one of the examples of how regulators could detect the LOB with spoofing order and simultaneously increase market quality. Regulation designed to catch and penalise spoofers, rather than just educating market participants about the possibility of spoofing, can be beneficial. Knowing that market surveillance systems include a spoofing detection mechanism signals manipulators to avoid spoofing trading on that markets. Moreover, market participants will trade cautiously using an RTSP spoofing risk indicator, which may change the market state and make the conditions unpreferable for spoofers to earn a low-risk profit, thus increasing market quality per se.

With respect to the current literature, we focus on real-time forecasting methodology trained on suspected spoofing events detected by the exchange, which makes our research unique. Another contribution to the literature is that we use several machine learning algorithms for detection with sophisticated robustness checks, which allow us to introduce a real-time spoofing probability measure. So, we are replicating MOEX's spoofing identification strategy using much more limited data with state-of-art technology and determine periods when we might see suspicious order activity.

To test our measure on out-of-sample data, we first train chosen ML algorithms on five previous trading days and then forecast for the next 10-, 30-, and 60-minute period in the current day. We use each ML algorithm's probabilities to predict the spoofing order occurrence in the next tick. Secondly, we construct the measure of a real-time spoofing probability as a simple average of ML outcomes. Consequently, the RTSP measure gives us a probability in real time that the market state is preferable for the potential spoofer to place his order. Moreover, we compare the forecasting performance of our measure to the performance of other ML algorithms and show that the designed methodology works better in the given environment.

In the empirical application, we create a unique dataset by combining data provided by MOEX with suspected spoofing orders identified by the exchange's internal algorithm on ten liquid stocks and data of all orders and trades over the same period. Those ten stocks are not the most traded stock on the Moscow exchange. So the choice of stocks is in line with the results of Williams & Skrzypacz (2020), which suggest that spoofing should be most prevalent in markets which are sufficiently liquid but not too liquid. Using files from MOEX with all orders and trades, we may track all orders placed on the market, identify whether they are filled or cancelled and reconstruct an entire limit order book on each tick. To the best of our knowledge, we are the first to use this level of granularity to address the spoofing forecasting problem.

Overall, in the paper, we show how effectively forecast the market state with spoofing manipulation activity and present a real-time forecasting measure as a trading risk indicator. We contribute to the empirical literature on detecting and forecasting spoofing manipulation in financial markets and make a step further to real-time signalling of manipulative environment, opening a new niche for further research.

The same algorithm could identify similar scenarios in different markets. If one wants to export this idea to a different stock exchange, she would need to train the same method on a different set of suspicious orders. Moreover, our approach can be applied to other disruptive financial market practices, especially high-frequency trading manipulation. The model is self-training, so the user does not need to make manual adjustments depending on the manipulative practices. Users need to consider only macro-level effects on the trading process, such as news announcements, dividend payments or similar.

The main contribution of our work is that we not only show the methodology to detect potential spoofing orders on historical information, but we also offer the real-time measure of detecting the market state when the manipulation is highly probable. The RTSP measure could be used as a warning sign for market participants, exchanges and regulators. We look into the future and indicate the real-time trading risk in the manipulative environment. This makes our study unique and opens a new area for further research.

The remainder of the paper is organized as follows. Section 2.2 describes the data and the variables. The following parts of the paper are dedicated to the methodology we use in our study (Section 2.3) and the results that we obtain to forecast spoofing (Section 2.4). In Section 2.5, we introduce the novel forecasting measure RTSP and show its predictive power (Section 2.5.1). Finally, some conclusive remarks are offered in Section 2.6.

2.2 Data and variables description

2.2.1 Data

In our empirical investigation, we are the first to use unique market data for six consecutive months from January until June 2019 from the Moscow Exchange (MOEX) consisting of two datasets.

The first dataset essential for the study is the information on 51,706 orders that MOEX identified as possible spoofing orders using their internal algorithm. This dataset of cancelled orders of 10 liquid stocks is in the following format: security id, time, buy/sell, volume, and price. Data providers do not release precise algorithms the exchange surveillance department used for spoofing identification; however, they could track potential spoofing orders by the trader's id numbers, how the order influences the market, whether the order tightens the spread, what lifetime of cancelled order is and many other internal and external information. Moreover, the exchange tracks the trading style of each trader and could mark potential spoofers. We know that suspicious spoofing orders in provided dataset were identified based on the information regarding proven detected spoofing cases from 2010 to 2019 in the USA, Europe, and the UK.

The second dataset consists of all orders placed on the stock market and includes the following information for each order: record number; security code; type of order (sell or buy); time (in the format of microseconds); order number; type of action (placed, cancelled, executed); price; volume; trade number (if executed); trade price (if executed). The data files include all orders on MOEX, all trades, and best bid and ask prices. This information allows us to reconstruct the full limit order book (LOB) up to maximum available level at any time and in any stock. We construct the book 50 levels deep.

We match suspicious spoofing orders identified by MOEX with the second dataset of orders and trades, which allows us to track the manipulative order's lifetime and identify its order book price level and the time of cancellation. So, we build a unique dataset that allows us to methodologically show how to forecast spoofing orders' appearance in real-time and empirically prove it.

Table 2.1 includes description and general statistics of the dataset with the spoofing orders.

format:	security id, time, buy/sell, volume, price
time:	6 months from January till June 2019
10 stocks:	FIVE', 'GMKN', 'MGNT', 'NVTK', 'PLZL', 'POLY', 'ROSN', 'SNGS','TATN', 'YNDX'
number of orders:	51 706 spoofing orders
lifetime:	on average the lifetime of spoofing order is 0.0045 min with maximum lifetime 0.33 min
price level:	on average spoofing order appears on the 5th price level below or above best bid or ask in the order book
max price level:	the maximum price level is 48
trades information:	on average 84% of spoofing orders do not have any trade during their lifetime
orders frequency:	spoofing orders appear every day on each stock
buy or sell:	55% of spoofing orders are sell orders
time of the day:	orders appear equally during the day, except YNDX, which spoofing orders appear only after 4pm

TABLE 2.1: Description and general statistics of the spoofing orders

2.2.2 Variables

The data allow us to reconstruct the limit order book at any time or tick. For our analysis, we rebuild the order book every 10 seconds for the ten stocks (Table 2.1). Each reconstructed order book consists of 50 ask prices and 50 bid prices, so the depth of the reconstructed order book is 50 levels. We use price levels depending on the stock's minimum price increment, including zero volume price levels. We only consider normal trading hours[1] and omit the first and last half hours of each trading day, which are often periods of high volatility and increased spreads (Cartea et al. 2019). Every price level in the reconstructed order books represents the volume of orders placed on the market from the beginning of the trading day minus cancelled and executed orders on the same price level.

Having the data, we can find the exact time when the suspicious spoofing order was placed. And we can observe the market state just before that order, consisting of the reconstructed limit order book, the trading price and volume. Having this information, we calculate different market state variables based on the literature. We divide variables into three large groups: market quality variables, variables of trades and their frequency, and order book variables.

[1]Normal trading hours on MOEX are from 10:00 till 18:45.

Market quality variables

1. Spread measures:

   - QS. Quoted spread is $(a - b)/m$, where a is the best ask, b is the best bid, and m is the midquote;
   - QS_delta. The difference in quoted spreads measures ten and twenty seconds before the spoofing order.
   - QS_delta_t2. The difference in quoted spreads measures ten and thirty seconds before the spoofing order.
   - ES. Effective (half) spread is $(p - m)/m$, where p is the trade price and m is the prevailing midquote (prior to execution) (Lee 1993, Blume & Goldstein 1992).

2. Short-term volatility measures:

   - VOL_1_Nmin. N-minute price volatility is a standard deviation of the midquotes within the interval divided by the last midquote (Aitken & Frino 1996). $N \in \{1, 2, 5, 10\}$ minutes.
   - VOL_2_Nmin. N-minute price volatility is measured as a standard deviation of price returns, where $N \in \{1, 2, 5, 10\}$ minutes.
   - VOL_3_Nmin. Average realized volatility in the time interval is measured as
     $1/n\Sigma_{s=1}^{N} \left| ln(m_{t-s} - m_{t-s-1}) \right|$, where m_s is the midquote at end of minute s, and $N \in \{1, 2, 5, 10\}$ minutes.

Variables of trades and its frequency

- $Hour$ is a trading hour from 10 am till 6 pm in which the tick prevailing the spoofing order occurred.

- UF. Ultra-fast activity measure is a fraction of limit orders submitted and cancelled very quickly out of all cancelled orders (Cartea et al. 2019). UF_X_N is an ultra-fast measure cancelled within X ms, where $X \in \{1, 10, 50, 100, 600\}$ ms in one minute, and N indicates how many minutes before spoofing order ultra-fast activity happened, where $N \in [1; 10]$ minutes. For example, spoofing order was placed at 10:30, so UF_10_5 is a fraction of the limit orders cancelled within 10 ms measured in the minute from 10:25 to 10:26.

The logic of the interval choice is taken out of the data and should be economically reasonable for answering the main research question. As we have data with orders that live on average less than a second, we consider short intervals of less than 10 minutes. One-minute intervals coincide with research on quote stuffing events (Egginton, Van Ness & Van Ness 2016), while 10-minute intervals are used by Cartea et al. (2019) and Hasbrouck & Saar (2013). For a better picture, we include 1-, 2-, 5-, 10-minute intervals.

Order book variables

Order book state is a "snapshot" of the LOB at every point in time when at least one action (trades, placement or cancellation of limit orders) was present. Two neighbouring order book states could have the time index at least a one-time increment possible on the exchange. [2] However, neighbouring order book states may be several seconds apart during slow trading times. We consider the order book state just before the spoofing order was placed [3]

We compute separate order book variables for the LOB's ask and bid sides. If we aim to determine the order book state before the manipulative order and then find variables with explanatory and predictive power, we have to consider that spoofing orders appear on one side of the LOB. Does it mean that the market state on this side is the one that attracts manipulators? While separating the dataset into two subsets of buy and sell orders, we address this question. Also, we calculate the order book depth variables separately for the ask and bid sides of the LOB.

1. Order book imbalance

   Volume imbalance of the limit order book is the ratio of volume posted at the best bid price minus the volume posted at the best ask price to the sum of the volume at the best bid and best ask prices (Cartea et al. 2020). We compute those imbalances for several visible price levels and the total order book imbalance.

   - Total order book imbalance:
     $IMB = (SB - SA)/(SA + SB)$,
     where SB is a sum of the orders' volume on all bid price levels, SA is a sum of the orders' volume on all ask price levels.
   - Order book imbalance until i price level:
     $IMB\_i = (SB_i - SA_i)/(SA_i + SB_i)$,
     where $i \in \{0, 1, 2, 3, 4, 5\}$ is a price level in the order book, where $i = 0$ is the best ask and best bid price levels.
   - Order book imbalance until the price level of the spoofing order:
     $IMB\_order = (SB^* - SA^*)/(SA^* + SB^*)$,
     where SB^* and SA^* are the sums of the orders' volume from the bid and ask sides of the order book, respectively, that lie within the price band defined by the spoofing order. For example, a sell spoofing order was placed inside the order book on the price five basis points higher than the best ask. So, SA^* is an accumulated volume from the best ask until the fourth price level above the best ask, while SB^* is an accumulated volume from the best bid until the fourth price level below the best offer.
   - Difference in the order book imbalance:
     $IMB\_delta = IBM\_10 - IBM\_20$,
     where IBM_10 is an order book imbalance measured ten seconds before the spoofing order. IBM_20 is the order book imbalance variables for the order book twenty seconds before the spoofing order.
     $IMB\_delta\_t2 = IBM\_10 - IBM\_30$,

[2] MOEX provides data in the format HHMMSSZZZXXX, which gives us the minimum time increment of 0.000001 seconds.

[3] The mean lifetime of spoofing orders is 0.005 seconds.

where IBM_30 is the variable for the order book thirty seconds before the spoofing order.

2. Order book spread

   - Order book spread
     $DistNormN = (AvAskN - AvBidN)/m$,
     where $AvAskN$ and $AvBidN$ are a mean ask and prices respectively that lies within $N \in \{10, 20, 50\}$ ask and bid price levels including those with zero volume; m is a midquote;
   - Clean order book spread
     $DistNormCleanN$ is the same measure as $DistNormN$, but we do not consider price levels with zero volume. $N \in \{10, 20, 50\}$ not including zero volume levels. For example, $N = 10$, but we observe two price levels from the ask side with zero volume. So, we must observe the next two price levels from the ask side to measure a mean ask price.

3. Order book depth

   Biais et al. (1995) found that thin books elicit orders and thick books result in trades. Ranaldo (2004) uses similar measures like pending volume in the number of shares divided by 10,000 at the best quote on the same market side as the incoming trader and on the opposite side of the market for the incoming trader. We follow similar logic by introducing the following variables.

   - Average order book depth
     $DistVolN = (VolAsk + VolBidN)/Vol\_m$ is the average cumulative volume resting in the LOB within $N \in \{10, 20, 50\}$ basis points of the best bid and ask (Cartea et al. 2019). $VolAskN$ and $VolAskN$ are cumulative volumes on the order book's ask and bid sides, resting within the N price level.
     $Vol\_m = (VolAsk + VolBid)/2$ is an average volume on the best bid and ask price levels.
     This measure accounts for the tick size by measuring the depth at given distances relative to the current best bid and ask. We normalize this measure by average volume on the best bid and ask price levels.
   - Average clean order book depth
     $DistVolCleanN$ is the same measure as $DistVolN$, but we do not consider price levels with zero volume. $N \in \{10, 20, 50\}$ not including zero volume levels. For example, $N = 10$, but we observe two price levels from the ask side with zero volume. So, we must count cumulative volume from the ask side until the twelfth price increment from the best ask.

4. Order book filling ratio:

   - $FRA = FA/TotalA$ and $FRB = FB/TotalB$,
     where FA and FB are the counts of price levels in the order book with at least one sell and buy limit order, respectively. $TotalA$ and $TotalB$ are the maximum price levels from the order book's ask and bid sides.

- $FRA\_delta = FRA\_10 - FRA\_20$ and
  $FRB\_delta = FRB\_10 - FRB\_20$,
  where FRA_10 and FRB_10 are order book filling ratios measured ten seconds before the spoofing order. FRA_20 and FRB_20 are order book filling ratios for the order book twenty seconds before the spoofing order.
  $FRA\_delta\_t2 = FRA\_10 - FRA\_30$ and
  $FRB\_delta\_t2 = FRB\_10 - FRB\_30$,
  where FRA_30 and FRB_30 are order book filling ratios for the order book thirty seconds before the spoofing order.
  So, the order book filling ratio is the number of filled price levels with the limit orders divided by the number of all price levels.
  The logic underneath this metric is the following. The order book has several visible price levels; however, the difference between neighbouring levels could be more significant than a minimum price increment. In other words, the order book could be filled with orders on each price level or many empty price levels. We don't find similar measures in the literature.

2.3 Methodology

While information content has seriously expanded over time, and its delivery speed increased dramatically, the methodological approaches have also improved. In recent years, we have witnessed several applications of ML techniques in empirical research on market microstructure. We utilize the forecasting abilities of ML methods in our paper in two steps. First, after identifying all possible variables or features that could describe the market state at a specific time we choose only those important for the current analysis. Secondly, with ML algorithms, we forecast if the market state is preferable for the manipulator to enter the order book in the next tick.

2.3.1 Step 1: Importance of variables

We identified many different variables to forecast spoofing events. All variables are standardized, which makes them comparable across stocks and easier to interpret. We remove outliers with more than three standard deviations away from the mean for each feature. However, one may argue whether all variables are helpful for the analysis.

First, we need to choose essential features for the prediction. We use the lasso as a regularization technique for variable selection. The lasso continuously shrinks the coefficients toward zero (Tibshirani 1996, Hastie et al. 2009) (Appendix 2.7.1). Allowing for prespecified nonlinearities and interactions, the lasso, however, can behave poorly if predictors are highly correlated (Zou & Hastie 2005). Correlation matrix is presented in Appendix 2.7.2. Due to the high correlation with reference to Zou & Hastie (2005), we also use elastic net (Enet) as a variable selection method. After the first step, we keep 67 features as predictors of the spoofing manipulation trading activity (Appendix 2.7.3).

2.3.2 Step 2: Machine learning for forecasting

We work with a unique extensive dataset that includes all orders placed tick-by-tick during six consecutive months. Furthermore, spoofing events are not equally spaced on the timeline. So, for our research, we choose machine learning as a methodology because of unevenly spaced data within a big dataset. ML methodology can also capture nonlinearities and variable interactions, as we identify as many as 67 predictors of spoofing manipulation events. For example, spoofing manipulation could be more likely if trading volume is below, for example, the 95th percentile and stock volatility is below the median. Another advantage of chosen approach is that ML methods are not making any a-priory assumptions on the form of the relationship. Machine learning also focuses on prediction, which is our research goal. Standard measures of market quality struggle to tell whether and when spoofing trading occurred on a given day and minute. ML methods allow us to identify the best prediction model driven by the data.

We have to make several design choices. We pick a dataset where spoofing activity is directly observed (so-called training sample in ML terms) so that a ML method can learn from contrasting model predictions with observed outcomes (so-called supervised learning). We pick orders disclosed by MOEX, which reports suspicious spoofing orders that were cancelled. We call a training sample of such spoofing orders "True" orders. Also, we randomly pick non-spoofing orders during the same period and call the sample "False" orders.

Section 2.2.2 presents a list of features that describes a market state before the manipulative order appears in the order book. In this section, we show how we use the machine learning (ML) approach to predict whether the current market state is preferable for the manipulator to place his spoofing order in the next tick. Later in this section, we report the results currently obtained for ML models, determining whether the order is "True" or "False".

We test commonly used ML models to predict the spoofing order event. Training on 75% data and testing on the rest of the models show the average prediction accuracy on the data from 50% to 69%. (Table 2.2). Further, we run the Diebold-Mariano test to compare the predictive qualities of the models in pairs. The test shows that Random Forest is better than all other models, while Decision Tree and XGBoost are better or not worse than other models (Appendix 2.7.4). We also implement the Model Confidence Set (MCS) procedure for model comparison. The MCS test shows the same results as the Diebold-Mariano test, confirming that Random Forest is the best model. To find the second-best model, we exclude Random Forest from the comparison test and get four other models (KNN, Decision Trees, Gradient Boosting, and XGBoost) that perform similarly well (Appendix 2.7.5).

We select ML methods that rely on decision trees (Chapters 9 and 10 in Hastie et al. (2009)). For example, if trading volume is above the 95th percentile, split on whether the quoted spread is above or below the median; otherwise, split on order book imbalance's 25th percentile. Each of the four leaves is then assigned an expected frequency of spoofing activity (historical average). Decision trees are invariant to variable scaling and robust to outliers. Random forest, introduced by Ho (1995), is a supervised ensemble learning method based on the decision tree. Random Forest takes an average over many random decision trees (Breiman 2001). Each of these many trees is constructed on a sample bootstrapped from a random subset of all predictors from the original dataset (e.g., 10 our of 100, Chapter 15 in Hastie et al. (2017)). Finally, our preferred method is eXtreme Gradient Boosting or XGBoost (Chen & Guestrin 2016), which efficiently implements Gradient Tree Boosting

Logistic regression:	52%
Support Vector Machines (SVM):	57%
K-nearest neighbors (KNN):	68%
Naive Bayes (NB):	50%
Stochastic Gradient Descent (SGD):	52%
Decision Tree (DT):	68%
Random Forest (RF):	76%
Gradient Boosting (GB) :	67%
XGBoost:	68%
Adaptive Markov Chains	50%
LSTM Neural Network (10 hidden layers)	55%
LSTM Neural Network (5 hidden layers)	55%
GRU neural network	51%
Percepton neural network	53%

TABLE 2.2: Prediction accuracy of ML models. The table shows results of commonly used ML models with True and False orders proportion of 1:1, splitting the data into 75% training and 25% validating sets.

(Chapter 10 in Hastie et al. (2017)). While Random Forest averages over random trees (so-called bagging), in Gradient Tree Boosting, each new tree focuses on examples that previous trees find problematic (so-called Boosting). In general, Boosting produces forecasts better than bagging but is much slower to estimate. XGBoost makes Gradient Tree Boosting almost as fast as Random Forest. It also recognizes that trees are prone to overfitting and penalizes trees with many leaves in favour of simpler, shorter trees (called regularization). The XGBoost package includes an efficient linear model solver and tree learning algorithm. It supports various objective functions, including regression, classification and ranking. Regularization makes the models perform worse in-sample but improves out-of-sample performance, which is our goal. We use the scikit-learn package in Python that implements the lasso and Random Forest and provides an XGBoost interface.

While we focus on Random Forest because it yields the best performance, XGBoost and Decision Tree results are the second best. As shown by Chen et al. (2015), in our research, the XGBoost algorithm similarly provides computational efficiency for handling big datasets. Although the Random Forest algorithms could perform better, we take XGBoost as another method as it utilises decision tree logic but works much faster and is thus preferable for practitioners. The XGBoost, similar to the Random Forest, is tuned using hyperparameters which is highly preferable on our data, where we want to use different order book variables for spoofing prediction and need to use clusters, rates, and other parameters inside one algorithm.

In recent years, ensemble algorithms including tree-based algorithms have become important for solving prediction and classification problems in many different fields with certain achievements (Zhou 2012). Instead of fitting a single model, tree-based methods combine multiple single tree models to obtain optimal prediction performance. This approach produces better predictions, require less data preprocessing and provide better fits to nonlinear relationships. These advantages make the tree-based approach a good choice when addressing financial market microstructure analysis. As we work with big data, and our dataset is unequally spaced with many predictive variables, ML algorithms based on trees perform better than others in our settings, as shown above in Table 2.2 and later in Table 2.15.

Finally, we make further improvements to the selected models. Random Forest, Decision Tree, and XGBoost are calibrated by looking for additional columns to serve as predictors and optimize hyperparameters available to improve outcomes (Appendix 2.7.6).

2.4 Empirical findings

2.4.1 Model evaluation method

We judge the model according to $accuracy$, $precision$, $recall$ and $f1 - score$ measures. We use the $accuracy$ measure as a prediction indicator. To check for $precision$, $recall$ and $f1 - score$, we split the data into 75% training and 25% validating sets.

$Accuracy$ is the proportion of correctly forecasted market states to the total number of all forecasted market states. In other words, accuracy measures the portion of all testing samples classified correctly. The trader can decide whether to enter the market using the $accuracy$ result produced by our model. If the $accuracy$ is almost one, the spoofer will likely enter the market using manipulative order, and the trader should be cautious. Whereas, if the prediction is close to zero, the market state is not preferable for the spoofer to enter and get profit so that the trader may enter the market safely. Any wrong prediction can cause the trader a money loss.

Hence, the model should be evaluated for its robustness. The other parameters used to evaluate a binary classifier's robustness are $precision$, $recall$ (also known as sensitivity) and $f1 - score$. The formulas to calculate parameters are given below:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}, \; Precision = \frac{tp}{tp + fp},$$

$$Recall = \frac{tn}{tn + fp}, \; F1 - score = \frac{2 * Recall * Precision}{Recall * Precision},$$

where, tp is a number of true positive values, tn is a number of true negative values, fp is a number of false positive values, and fn is a number of false negative values.

In other words, $precision$ measures the proportion of all correctly identified samples in a population of samples which are classified as positive labels. $Precision$ answers the question: Of all "True" predicted market states with spoofing orders, how many actually had spoofing orders in the LOB? For example, if $precision$ of "True" spoofing orders equals 0.82, it means that 82% of all predicted market states with spoofing orders in the LOB were correctly identified by the ML algorithm. Similarly, if $precision$ of "False" spoofing orders equals 0.74, it means that 74% of predicted market states without spoofing orders in the LOB were correctly identified.

$Recall$ or sensitivity measures the ability of a classifier to identify positive labels correctly. $Recall$ answers the question: Of all actual "True" market states with spoofing orders, how many were predicted correctly by the ML algorithm? For example, if $recall$ of "True" spoofing orders equals 0.71, it means that 71% of market states with spoofing orders in the LOB were identified correctly by the ML algorithm. Similarly, if $recall$ of "False" spoofing orders equals 0.84, it means that 84% of market states without spoofing orders in the LOB were identified correctly.

$F1 - score$ combines $precision$ and $recall$ into a single metric, which makes it more convenient for users to have only one performance metric rather than multiple. $F1 - score$ is the harmonic mean of $precision$ and $recall$, and it is never higher than the geometrical mean. It also tends towards the least number, minimizing the impact

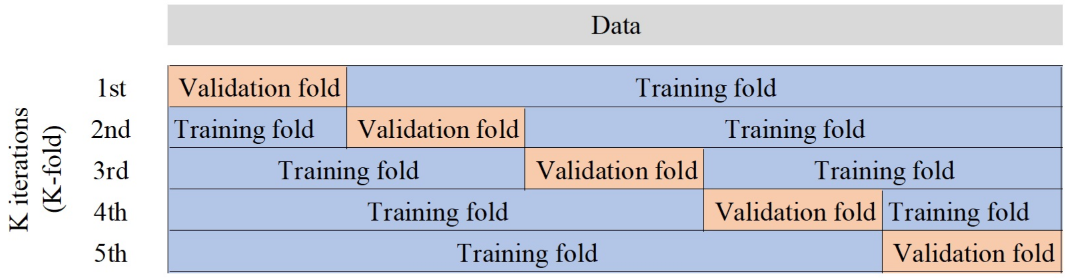


FIGURE 2.1: 5-fold cross-validation

of the large outliers and maximizing the impact of small ones. $F1 - score$, therefore, tends to privilege balanced systems.

2.4.2 Robustness of ML models

To show the robustness of our results, we use cross-validation that helps us avoid overfitting and keeps the analysis out-of-sample. Cross-validation is a resampling procedure used to evaluate the consistency of prediction quality of machine learning models on a limited data sample. The process has a single parameter k that refers to the number of groups a given data sample will be split into. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample to evaluate how the model is expected to perform in general when used to make predictions on data not used during the model's training. Bergmeir et al. (2018) theoretically show that cross-validation can control for overfitting the data when machine learning methods are used for prediction.

We start with the 5-fold cross-validation for each ML model and obtain average prediction accuracy. It generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train-test split. The general procedure is as follows: firstly, the dataset is shuffled randomly; secondly, the dataset is split into k groups to take each group as a hold-out or test data set and then fit a model on the training set and evaluate it on the test set; and finally, retain the evaluation score and discard the model. Then we summarize the model's skill using the sample of model evaluation scores. Then we assess the model forecasting performance based on the mean accuracy of the prediction of each fold. The following Figure 2.1 shows the cross-validation logic.

We use K-fold, Stratified K-fold, and Shuffle cross-validations, splitting the spoofing events into K non-overlapping parts in calendar time. We run cross-validation with different Ks (3, 4, 5, 6, 8, 10) to test for the stability of the outcomes.

<u>K-fold cross-validation</u>

The technique is repeated K times to use every fold as a validation set and other left-outs as a training set. This validation technique is not considered suitable for imbalanced datasets as the model will not get adequately trained owing to the proper ratio of each class data. We train data using the shuffling and block approach.

The typical approach when using k-fold cross-validation is to shuffle the data randomly and split it in K equally-sized folds. By shuffling the dataset, we ensure that the model is exposed to a different sequence of samples in each part, which can help to prevent it from memorizing the order of the training data and overfitting to specific patterns. The blocked k-folds cross-validation procedure is similar to the standard form described above. The difference is that there is no initial random shuffling of observations.

Stratified K-fold cross-validation
K-fold validation splits into k-folds with a uniform probability distribution. However, stratified k-fold, an enhanced version of the k-fold cross-validation technique, splits the dataset into K equal folds. Each fold has the same ratio of instances of target variables. This method enables us to work with imbalanced datasets.

Shuffle cross-validation
And finally, to check the robustness of our results we use the shuffle cross-validation technique that involves splitting the whole data in the percentage of your choice and keeping the train-test split percentage different. This method allow us to get the test errors and access the outcomes.

Different True-False proportion
Moreover, we check how the models perform with the change in the proportion of True and False spoofing orders in our sample. So, we add more than 51,706 randomly selected False orders, which makes the True-False proportion 1:2. We run the models with cross-validations with the new imbalanced dataset. We also use different cross-validation checks for imbalanced data.

2.4.3 Results of ML models

Results for balanced data (1:1 True-False proportion)
Overall three ML models chosen on the first step of the analysis predict correctly from 66% to 78% of spoofing cases for balanced data. Training Ml algorithms on 75% of our dataset and then testing on the rest 25% and repeating several times with keeping the split ratio tell us that the chosen methodology has a valuable forecasting accuracy. Further, we use different validation methods to show that our models are not overfitting, so the models will perform well on the new data. Tables 2.3, 2.4, 2.5, 2.6 below show the results.

	Random Forest			XGBoost			Decision Tree		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
False	0.74	0.84	0.79	0.78	0.63	0.69	0.64	0.72	0.68
True	0.82	0.71	0.76	0.70	0.82	0.75	0.69	0.60	0.64
Accuracy		0.78			0.73			0.66	

TABLE 2.3: Accuracy of ML models. The table shows results of Random Forest, XGBoost and Decision Tree ML models with True and False orders proportion are equal to 1:1, splitting the data into 75% training and 25% validating sets. *Accuracy* measures the portion of all testing samples classified correctly. *Recall* measures the ability of a classifier to correctly identify positive labels. And *Precision* measures the proportion of all correctly identified samples in a population of samples which are classified as positive labels. $F1-score$ combines *precision*, *recall* into a single metric, $F1-score = (2 * Recall * Precision)/(Recall * Precision)$.

Different cross-validation checks illustrate that the models show good performance no matter how we change the testing sample. We have the same forecasting ability regardless of how we choose the data sample for validation. A small variability in the results is acceptable, as some subsamples are noisier than others.

	1. Shuffle			
K	Accuracy (%)	Random Forest	XGBoost	Decision Tree
3	Min accuracy	76.5	72.5	66.0
	Max accuracy	76.9	72.7	66.6
	Average accuracy	76.7	72.6	66.3
4	Min accuracy	77.1	72.4	65.8
	Max accuracy	77.4	72.9	66.4
	Average accuracy	77.2	72.7	66.1
5	Min accuracy	77.2	72.4	65.6
	Max accuracy	77.7	73.3	66.4
	Average accuracy	77.5	72.9	66.1
6	Min accuracy	77.3	72.0	65.7
	Max accuracy	78.2	73.4	66.4
	Average accuracy	77.7	72.9	66.1
8	Min accuracy	77.4	72.0	65.9
	Max accuracy	78.3	73.2	66.8
	Average accuracy	77.9	72.7	66.1
10	Min accuracy	77.4	72.4	65.5
	Max accuracy	78.4	73.3	66.9
	Average accuracy	78.0	73.0	66.3
	2. Blocks			
K	Accuracy (%)	Random Forest	XGBoost	Decision Tree
3	Min accuracy	56.6	62.0	62.8
	Max accuracy	61.0	66.4	66.7
	Average accuracy	58.4	64.3	64.7
4	Min accuracy	54.9	61.0	59.0
	Max accuracy	61.9	67.9	65.8
	Average accuracy	57.9	64.7	62.2
5	Min accuracy	55.9	63.0	60.1
	Max accuracy	64.2	70.2	69.2
	Average accuracy	59.3	65.9	64.4
6	Min accuracy	55.3	60.0	59.3
	Max accuracy	63.4	68.5	68.4
	Average accuracy	58.6	64.8	64.2
8	Min accuracy	55.3	56.5	56.1
	Max accuracy	68.7	68.6	69.1
	Average accuracy	58.9	64.5	62.8
10	Min accuracy	54.7	59.6	57.2
	Max accuracy	70.3	71.9	71.3
	Average accuracy	59.8	65.9	64.1

TABLE 2.4: K-fold cross-validation check for of ML models using balanced data with 1:1 True-False proportion. The table shows the results of two cross-validation approaches: *1. Shuffle*: the complete dataset is shuffled in random k-parts; *2. Blocks*: the dataset is sorted by time into blocks without shuffling.

K-fold cross-validation with blocks shows that XGBoost outperforms the random forest algorithm, while the shuffle k-fold cross-validation shows the opposite. Therefore, developing a combination rather than a unique algorithm is essential, as no uniquely dominant methodology exists. We confirm the fact in the literature that forecast combination is always better than individual forecast (Claeskens et al. 2016).

K	Accuracy (%)	Random Forest	XGBoost	Decision Tree
3	Min accuracy	76.4	72.0	65.7
	Max accuracy	76.9	73.0	66.3
	Average accuracy	76.6	72.5	66.1
4	Min accuracy	77.0	72.2	65.3
	Max accuracy	77.3	72.9	66.8
	Average accuracy	77.1	72.7	66.0
5	Min accuracy	77.2	71.9	65.8
	Max accuracy	78.0	73.4	66.8
	Average accuracy	77.5	72.8	66.2
6	Min accuracy	77.4	72.5	65.8
	Max accuracy	78.0	73.5	66.2
	Average accuracy	77.6	72.8	66.0
8	Min accuracy	77.5	72.4	65.5
	Max accuracy	78.4	73.5	67.2
	Average accuracy	77.9	73.0	66.2
10	Min accuracy	77.5	71.7	65.1
	Max accuracy	79.2	73.9	67.1
	Average accuracy	78.0	72.9	66.2

TABLE 2.5: Stratified K-fold cross-validation check for of ML models using balanced data with 1:1 True-False proportion.

K	Accuracy (%)	Random Forest	XGBoost	Decision Tree
3	Min accuracy	76.0	71.9	65.8
	Max accuracy	76.3	72.7	66.1
	Average accuracy	76.2	72.4	65.9
4	Min accuracy	76.1	72.2	65.5
	Max accuracy	76.6	72.6	66.8
	Average accuracy	76.3	72.4	65.9
5	Min accuracy	75.8	72.1	65.8
	Max accuracy	76.7	73.0	66.4
	Average accuracy	76.2	72.6	66.2
6	Min accuracy	75.6	71.7	65.8
	Max accuracy	76.6	72.7	66.4
	Average accuracy	76.2	72.3	66.1
8	Min accuracy	75.7	71.8	65.7
	Max accuracy	76.7	72.8	66.4
	Average accuracy	76.3	72.2	66.1
10	Min accuracy	75.9	71.6	65.6
	Max accuracy	76.6	72.9	66.6
	Average accuracy	76.4	72.3	66.1

TABLE 2.6: Shuffle K-fold cross-validation check for of ML models using balanced data with 1:1 True-False proportion, training on 60% of the data and validating on 20% of the data

Results for imbalanced data (1:2 True-False proportion)

Random Forest model gives 82.3%, which is a good result, but since the dataset is imbalanced, only 57% of True orders were predicted correctly, while for False orders, prediction accuracy is 95%. XGBoost model gives overall 78.2% prediction accuracy with 59% accuracy in predicting True orders and 88% False orders. The decision tree also shows similar results with a little less forecasting accuracy (Tables 2.7, 2.8, 2.9, 2.10).

	Random Forest			XGBoost			Decision Tree		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
False	0.81	0.95	0.88	0.81	0.88	0.84	0.76	0.85	0.80
True	0.85	0.57	0.69	0.71	0.59	0.65	0.61	0.46	0.53
Accuracy		0.82			0.78			0.72	

TABLE 2.7: This table shows results of Random Forest, XGBoost and Decision Tree ML models with True and False orders proportion are equal to 1:2, splitting the data into 75% training and 25% validating sets. *Accuracy* measures the portion of all testing samples classified correctly. *Recall* measures the ability of a classifier to correctly identify positive labels. And *precision* measures the proportion of all correctly identified samples in a population of samples which are classified as positive labels. $F1 - score$ combines $precision$, $recall$ into a single metric, $F1 - score = (2 * Recall * Precision)/(Recall * Precision)$.

K	Accuracy (%)	Random Forest	XGBoost	Decision Tree
3	Min accuracy	81.6	77.9	71.3
	Max accuracy	81.9	78.0	72.0
	Average accuracy	81.7	77.9	71.6
4	Min accuracy	81.8	77.9	71.7
	Max accuracy	82.2	78.4	72.1
	Average accuracy	82.1	78.1	71.9
5	Min accuracy	81.9	77.8	71.5
	Max accuracy	82.5	78.4	72.3
	Average accuracy	82.3	78.2	71.9
6	Min accuracy	82.0	77.6	71.2
	Max accuracy	82.7	78.4	72.2
	Average accuracy	82.3	78.2	71.8
8	Min accuracy	82.1	77.8	71.3
	Max accuracy	83.0	78.5	72.2
	Average accuracy	82.6	78.3	71.8
10	Min accuracy	82.1	77.8	71.2
	Max accuracy	83.1	78.6	72.3
	Average accuracy	82.7	78.3	71.8

TABLE 2.8: Stratified K-fold cross-validation check for of ML models using imbalanced data with 1:2 True-False proportion.

	1. Shuffle			
K	Accuracy (%)	Random Forest	XGBoost	Decision Tree
3	Min accuracy	81.7	77.8	71.8
	Max accuracy	81.8	77.9	72.0
	Average accuracy	81.7	77.9	71.9
4	Min accuracy	81.9	77.9	71.7
	Max accuracy	82.3	78.2	72.2
	Average accuracy	82.0	78.0	71.9
5	Min accuracy	82.0	77.8	71.6
	Max accuracy	82.5	78.3	72.1
	Average accuracy	82.2	78.1	71.8
6	Min accuracy	82.2	77.9	71.7
	Max accuracy	82.7	78.4	72.1
	Average accuracy	82.5	78.2	71.9
8	Min accuracy	82.2	78.1	71.3
	Max accuracy	83.0	78.7	72.3
	Average accuracy	82.6	78.3	71.9
10	Min accuracy	82.3	78.0	71.4
	Max accuracy	83.2	78.6	72.1
	Average accuracy	82.7	78.3	71.8
	2. Blocks			
K	Accuracy (%)	Random Forest	XGBoost	Decision Tree
3	Min accuracy	66.4	68.9	69.3
	Max accuracy	66.5	70.3	70.7
	Average accuracy	66.4	69.7	69.8
4	Min accuracy	66.1	67.6	67.5
	Max accuracy	66.7	72.1	71.6
	Average accuracy	66.5	69.8	69.6
5	Min accuracy	64.7	69.2	68.0
	Max accuracy	70.6	72.7	71.3
	Average accuracy	67.3	70.8	69.7
6	Min accuracy	65.8	68.1	66.2
	Max accuracy	66.9	71.6	71.7
	Average accuracy	66.5	69.7	69.6
8	Min accuracy	64.0	66.0	66.2
	Max accuracy	68.3	73.4	73.7
	Average accuracy	66.6	69.7	69.6
10	Min accuracy	65.0	66.9	66.2
	Max accuracy	74.8	77.3	74.5
	Average accuracy	67.4	71.0	69.8

TABLE 2.9: K-fold cross-validation check for of ML models using imbalanced data with 1:2 True-False proportion. We show the results of two cross-validation approaches: *1. Shuffle*: the complete dataset is shuffled in random k-parts; *2. Blocks*: the dataset is sorted by time into blocks without shuffling.

K	Accuracy (%)	Random Forest	XGBoost	Decision Tree
3	Min accuracy	81.0	77.8	71.6
	Max accuracy	81.3	78.0	71.9
	Average accuracy	81.2	77.9	71.8
4	Min accuracy	81.1	77.7	71.3
	Max accuracy	81.6	78.0	72.1
	Average accuracy	81.4	77.9	71.6
5	Min accuracy	81.2	77.7	71.7
	Max accuracy	81.5	78.1	72.2
	Average accuracy	81.4	78.0	71.9
6	Min accuracy	81.2	76.9	71.3
	Max accuracy	81.6	78.2	72.2
	Average accuracy	81.4	77.6	71.8
8	Min accuracy	81.0	77.3	71.4
	Max accuracy	81.7	78.0	72.1
	Average accuracy	81.3	77.7	71.7
10	Min accuracy	80.9	77.4	71.3
	Max accuracy	81.9	78.4	72.3
	Average accuracy	81.5	77.8	71.8

TABLE 2.10: Shuffle K-fold cross-validation check for of ML models using imbalanced data with 1:2 True-False proportion, training on 60% of the data and validating on 20% of the data.

Although one could argue that 1:1 and even 1:2 proportion of True/False spoofing orders are unrealistic, ML literature addresses this question methodologically. When researchers lack data for training purposes, the literature suggests balancing data using several techniques, such as adding synthetically generated data (SMOTE, etc) (More & Rana 2017, Gong & Kim 2017). Balancing the data improve forecasting accuracy. We show empirically how the algorithm behaves if we change the proportion of True and False orders without generating additional orders as our data sample is relatively big with 51,706 suspicious spoofing orders.

To illustrate the performance of the models using the imbalanced dataset, we run the chosen ML algorithms increasing the proportion of False orders up to 15 by cutting the number of True spoofing orders. The results are presented in Table 2.11 and show that validity of the method strongly relies on the proportion of the data trained.

Various state-of-the-art learning techniques have been suggested in the past few years to address classification problem in imbalance dataset, for example, algorithm level methods and data level methods. Algorithm driven approach or classifier level approach keeps the training dataset invariable and adjusts the inference algorithm to facilitate learning specifically related to the minority class. In other words, while adds more priority to predicting True spoofing orders, some techniques can make models work better by changing internal parameters in the code. To illustrate this approach, we run Random Forest for imbalanced datasets called "BalanceRandomForestClassifier" (Appendix 2.7.6), which randomly under-samples each bootstrap sample to balance it. We get a higher predictive accuracy of 80% with a True order prediction of 71%. For XGBoost to balance order weights, we significantly change the parameter 'scale_pos_weight' from 1 to 500 (Appendix 2.7.6), which puts 500 times more priority on True orders prediction. We obtained 81% overall accuracy. We show the results in Table 2.12.

True:False	Random Forest			XGBoost			Decision Tree		
	Accuracy	Recall False	Recall True	Accuracy	Recall False	Recall True	Accuracy	Recall False	Recall True
1:1	0.77	0.84	0.70	0.72	0.63	0.82	0.66	0.72	0.59
1:2	0.80	0.95	0.50	0.77	0.88	0.57	0.72	0.90	0.38
1:3	0.83	0.97	0.41	0.81	0.95	0.42	0.77	0.98	0.14
1:4	0.86	0.98	0.36	0.84	0.97	0.35	0.81	0.99	0.13
1:5	0.88	0.99	0.33	0.87	0.99	0.29	0.84	0.99	0.09
1:6	0.89	0.99	0.33	0.89	0.99	0.27	0.86	0.99	0.11
1:7	0.90	0.99	0.28	0.90	0.99	0.23	0.88	0.99	0.09
1:8	0.91	0.99	0.28	0.91	1.00	0.23	0.89	1.00	0.05
1:9	0.92	0.99	0.27	0.92	1.00	0.22	0.90	1.00	0.05
1:10	0.92	0.99	0.24	0.92	1.00	0.20	0.91	1.00	0.06
1:11	0.93	0.99	0.25	0.93	1.00	0.21	0.92	1.00	0.04
1:12	0.93	0.99	0.22	0.93	1.00	0.19	0.92	1.00	0.6
1:13	0.94	0.99	0.24	0.94	1.00	0.21	0.93	1.00	0.3
1:14	0.94	0.99	0.23	0.94	1.00	0.18	0.93	1.00	0.3
1:15	0.95	1.00	0.21	0.94	1.00	0.17	0.94	1.00	0.6

TABLE 2.11: This table shows the results of training ML algorithms on imbalanced data, achieved by cutting the initial dataset of True spoofing orders to achieved the proportion presented in the first column. We split the data into 75% training and 25% validating sets. We report *Accuracy* and *Recall* measures. *Accuracy* measures the portion of all testing samples classified correctly. *Recall* measures the ability of a classifier to correctly identify positive labels. In other words, *Recall* could be seen as an accuracy for True and False orders separately.

	Balanced RF (1)			XGBoost (500) (2)			XGBoost (10) (3)		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
False	0.85	0.86	0.86	0.83	0.91	0.87	0.79	0.86	0.82
True	0.72	0.70	0.71	0.77	0.63	0.70	0.66	0.55	0.60
Accuracy		0.81			0.81			0.75	

TABLE 2.12: This table shows results of modified parameters for Random Forest model using 'BalanceRandomForestClassifier' (1); and XGBoost ML model with 500 and 10 modification in 'scale_pos_weight' parameter, (2) and (3) respectively. We use imbalanced True and False orders proportion equal to 1:2, splitting the data into 75% training and 25% validating sets. *Accuracy* measures the portion of all testing samples classified correctly. *Recall* measures the ability of a classifier to correctly identify positive labels. And *precision* measures the proportion of all correctly identified samples in a population of samples which are classified as positive labels. $F1 - score$ combines *precision*, *recall* into a single metric, $F1 - score = (2 * Recall * Precision)/(Recall * Precision)$.

Data driven approach includes random undersampling and oversampling. Under sampling eliminates instance of majority class randomly to balance the dataset (Kotsiantis et al. 2006). Oversampling techniques includes focused oversampling, synthetic sampling, random oversampling and advanced heuristic techniques like synthetic minority oversampling (SMOTE) (Chawla et al. 2002).

Our dataset of suspicious spoofing orders includes 51,706 of True and the same number of False orders allowing us to avoid using the abovementioned sample adjustment techniques and get adequate robust results.

2.5 Real-time spoofing probability (RTSP) measure

In this section, we introduce a new measure, Real-Time Spoofing Probability (RTSP), as an average of three machine learning algorithms (Random Forest, XGBoost and Decision Tree), that we have chosen and tested in Step 2 of our analysis (Section 2.3.2). RTSP is a real-time indicator of the risk of interacting with a spoofer.

Unlike liquidity measures such as effective or quoted bid-ask spreads, a spoofing manipulation measure is not easy to identify. Manipulators usually hide behind retail trades' order flow to avoid being detected. Empirical measures are typically motivated by theoretical studies of spoofing trading to overcome these limitations. For example, Cartea et al. (2020) model the trading strategy of an investor who spoofs the limit order book (LOB) and computes the LOB volume imbalance as a measure that is important for spoofing strategy to be profitable. Our research proposes a data-driven approach to measuring trading risk when spoofing manipulation is highly probable. State-of-the-art ML techniques help us learn from the data and find how the LOB when suspected spoofers trade differ from the LOB when they do not trade. This flexible approach accounts for non-linearities and interactions between variables, while cross-validation and regularization, standard in ML, prevent it from overfitting the data.

After the model is estimated on a training sample, we extrapolate it to the next period and compute RTSP by applying the model parameters to the set of input variables observed for a given LOB state. In the literature, empirical evidence and extensive simulations show that the estimated optimal forecast combination typically does not perform well, and that the arithmetic mean often performs better (Claeskens et al. 2016). Our cross-validation checks also show that algorithms' results differ depending on the data split and there is no uniquely dominant methodology. So, we construct the RTSP measures as a simple average of ML outcomes from Section 2.3.2, where we show that chosen models (Random Forest, XGBoost, and Decision Tree) correctly predict over 70% of spoofing events for balanced data.

To test our measure on out-of-sample data, we train models on five previous consecutive trading days and the current trading day before 14:00. As the lifetime of spoofing order is less then a minute according to Table 2.1, we aim to short-term forecasting time window. Moreover, taking into account that features in our model are based on limit order book information, that changes quickly, we need to add new information to the training set often. So, we forecast spoofing orders for the next 10, 30, and 60 minutes from 14:00 to 18:00. We forecast each market state or tick. We keep the variety of 10, 30, and 60 minutes forecast for the robustness of our findings. For example, in a 10-minute setting, we train data on the previous five trading days and today's morning, then forecast spoofing risk for each tick, and every 10 minutes, we retrain the model with new data.

ML models give us the result from zero to one, which is the probability of the spoofing order occurring in the next tick. So, RTSP is higher when spoofing trading is more likely. We do not use rounding to compute RTSP measure a simple average of three numbers. We train models on the balanced dataset with a 1:1 True-False orders proportion. Then we predict the imbalanced dataset with a 1:2 True-False orders proportion as it is closer to the actual trading environment. We use two validation approaches where we add more data to train the dataset. Fig. 2.4 shows the logic of the expanding validation approach when we add more data for the training period while keeping the testing period 10, 30 or 60 minutes. Moreover, we also run a rolling validation approach when we keep the training period the same for all the splits (Fig.2.5). We show the results of both approaches in Table 2.13.

2.5.1 Forecasting power of RTSP

We forecast 70,347 orders including of 19,971 True and 50,376 False suspect spoofing orders. Table 2.13 shows the forecasting results of all three models and RTSP measures with validation approaches illustrated in Fig. 2.4 and Fig. 2.5.

10-minute re-estimation frequency			
Model	All correct (%)	True correct (%)	False correct (%)
XGBoost	67.2 / 66.4	74.4 / 46.9	75.0 / 75.3
Random Forest	68.7 / 68.8	22.2 / 22.1	87.1 / 87.3
Decision Tree	65.0 / 64.5	49.5 / 47.4	71.1 / 71.3
RTSP	68.2 / 68.3	42.5 / 41.9	78.4 / 78.8
30-minute re-estimation frequency			
Model	All correct (%)	True correct (%)	False correct (%)
XGBoost	66.8 / 66.7	47.1 / 46.1	74.6 / 74.8
Random Forest	68.2 / 68.5	21.4 / 21.9	86.8 / 86.9
Decision Tree	65.3 / 64.6	50.2 / 48.7	71.3 / 70.9
RTSP	68.1 / 68.3	42.7 / 42.4	78.1 / 78.6
60-minute re-estimation frequency			
Model	All correct (%)	True correct (%)	False correct (%)
XGBoost	66.3 / 66.4	47.0 / 46.2	74.0 / 74.4
Random Forest	67.9 / 68.2	21.3 / 21.9	86.4 / 86.5
Decision Tree	65.0 / 64.3	50.1 / 48.2	70.9 / 70.7
RTSP	67.9 / 67.8	42.7 / 42.1	77.9 / 78.0

TABLE 2.13: Forecasting performance of RTSP. The table shows the results of ML models and the RTSP measure with 10-, 30- and 60-minute forecasting re-estimation settings, where column *All correct* shows the percentage of correctly predicted all market states, while columns *True correct* and *False correct* show the percentage of correctly predicted market states with True and False spoofing orders respectively. The first number shows the result using the expanding validation approach (Fig. 2.4) when we add new data to the training dataset while keeping the testing time frame the same (10, 30, or 60 minutes). The second number after the slash shows the results using the rolling validation approach illustrated in Fig. 2.5.

Overall, we do not observe significant differences in predictive quality across 10-, 30- and 60-minute rolling forward predictions. So, whether we re-estimate our model every 10 minutes or every hour does not significantly improve the results. XGBoost and Decision Tree models predict around 50% of True and 70-75% of False spoofing cases correctly with an overall predictive quality of approximately 67%. Random Forest predicts only 20-25% of True cases; however, 87% of False cases, with overall approximately 68% events predicted correctly. At the same time, the RTSP measure captures all three models. Moreover, different forecasting validation approaches show similar results, and we continue our analysis with a rolling cross-validation method.

Additionally, we test our method on buy and sell suspect spoofing orders separately. Table 2.14 shows that forecasting accuracy for market states with buy spoofing orders is slightly better. This example indicates that one could train models differently depending on the required task; for example, the exchange wants to know the risk of spoofing occurrence only from the bidding side of the limit order book.

10-minute re-estimation frequency						
Model	All Buy (%)	True Buy (%)	False Buy (%)	All Sell (%)	True Sell (%)	False Sell (%)
XGBoost	70.4	40.9	80.2	67.9	40.3	77.0
Random Forest	72.2	24.1	88.2	70.5	22.7	86.3
Decision Tree	65.9	43.5	73.2	63.6	45.2	69.7
RTSP	70.5	37.6	81.4	68.4	37.9	78.5
30-minute re-estimation frequency						
Model	All Buy (%)	True Buy (%)	False Buy (%)	All Sell (%)	True Sell (%)	False Sell (%)
XGBoost	70.6	41.6	79.8	67.7	41.1	76.6
Random Forest	72.1	24.1	87.9	70.4	22.8	86.2
Decision Tree	65.6	42.4	73.3	63.6	45.5	69.6
RTSP	70.2	37.0	81.2	68.3	38.1	78.2
60-minute re-estimation frequency						
Model	All Buy (%)	True Buy (%)	False Buy (%)	All Sell (%)	True Sell (%)	False Sell (%)
XGBoost	70.3	41.9	79.6	67.6	41.3	76.4
Random Forest	72.0	24.0	87.8	65.1	22.9	85.9
Decision Tree	65.0	42.9	72.3	63.2	45.7	69.0
RTSP	70.0	37.6	80.7	68.2	38.7	77.9

TABLE 2.14: The table shows the performance of ML models and the RTSP measure for buy and sell spoofing risk with 10-, 30- and 60-minute forecasting re-estimation settings. The column *All Buy* shows the percentage of correctly predicted market states with buy spoofing orders, while columns *True Buy* and *False Buy* show the percentage of correctly predicted market states with True and False buy orders, respectively. The same logic is applied for forecasting market state with sell spoofing orders.

Consequently, RTSP shows a probability in real time that the market state is preferable for the spoofers to place their order. Out-sample forecasting reveals that a forecast combination is better than an individual forecast. Table 2.13 shows that Random Forest slightly overperforms RTSP measure while assessing all spoofing and non-spoofing market states forecast. However, for example, in 10-minute re-estimation frequency, we observe RTSP measure significantly overperforming in forecasting "True" spoofing orders, which is a high priority for the regulator and traders as final users and beneficiaries of the developed algorithm. The aim of the machine learning application is equally important to forecast the market state when spoofing is likely to occur and the market state which is safe to trade without manipulative orders. RTSP smoothes the prediction of the individual machine learning algorithms, making the forecasting outcomes more reliable.

As the next logic step, we compare the forecasting performance of RTSP to the performance of other ML algorithms. We show that our designed methodology performs better in the given environment. We run a rolling cross-validation forecast using other ML models to assess their predictive power. Table 2.15 shows the forecasting power of alternative models, and we observe worse results than those from the RTSP measure.

Model	All correct (%)	True correct (%)	False correct (%)
Logistic Regression	57.7	37.4	65.7
Logistic Regression (Lasso)	57.9	37.6	65.9
Logistic Regression (Ridge)	57.7	37.4	65.7
K-nearest neighbors	58.5	38.2	66.5
Stochastic Gradient Descent	53.9	43.8	57.9
Support Vector Machines	29.9	17.5	34.8
LSTM with 10 layers	54.6	44.0	58.9
GRU	61.4	25.9	75.5
Percepton	46.6	55.4	43.1

TABLE 2.15: Forecasting results of alternative models. The table shows 60-minute forecasting results using alternative models, where column *All correct* shows the percentage of correctly predicted all orders, while columns *True correct* and *False correct* show the percentage of correctly predicted True and False orders, respectively.

2.5.2 Practical application of RTSP

Finally, we want to illustrate one of many possible applications that could be developed from our model and RTSP measure, as our method could be used as a real-time risk indicator for authorities or exchanges showing the market conditions when spoofing orders may appear in the order book leading to a worse market state for the order execution. One could run an automatic signalling application of the spoofing risk. Fig. 2.2 shows an example of the graphical spoofing risk indication. We set a critical value for the RTSP measure as 0.4 and marked it as a horizontal black dotted line on the bottom chart. We observe that RTSP rises above the critical value at the end of the trading day around 18:30. On the top chart, we see that many True suspicious spoofing orders appear at that time, which we mark as green dotted lines. Another example on the graph is the time between 17:30 and 17:45 when RTSP significantly rises above 0.4 and correctly forecasts the spoofing order placement. The user may change the critical value parameter and adjust the signal appearance as a number, warning, or trading feature on the chosen platform. So, the RTSP measure could be incorporated into trading platforms or surveillance systems. The model is self-training; no adjustments are needed to detect other manipulative practices.

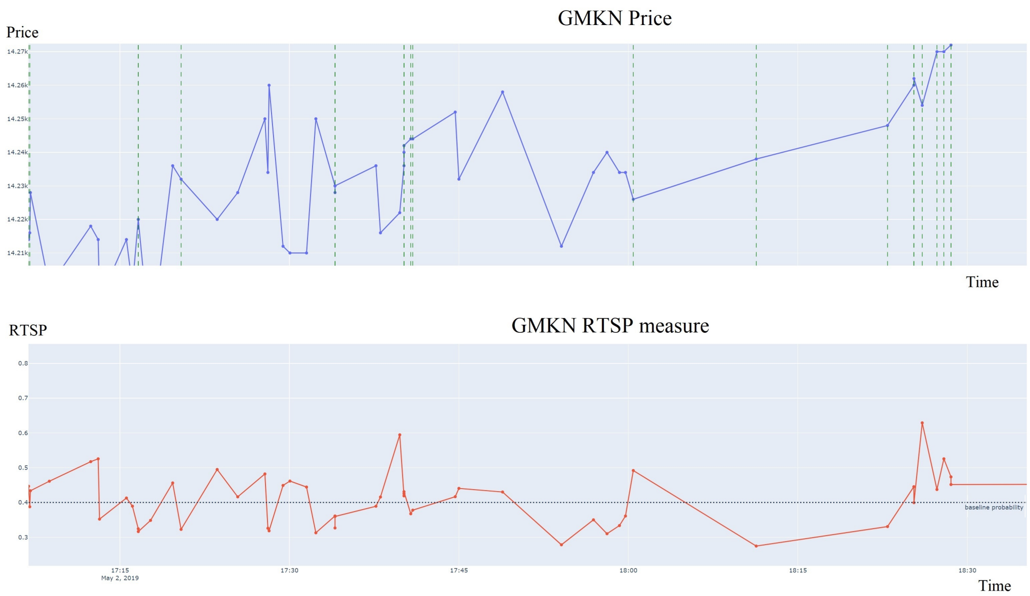


FIGURE 2.2: RTSP implementation. The figure illistates an example of RTSP measure signals on GMKN stock. The vertical green dotted lines on the top chart indicate when True spoofing orders happened. The horizontal black dotted line on the bottom chart indicates a critical value for the RTSP measure of 0.4.

2.6 Conclusion and further research

The major contribution of our research is that we introduce a novel data-driven approach to the real-time prediction of market state when a spoofing event is highly probable. We introduce a Real-Time Spoofing probability (RTSP) measure that indicates the risk of intraday manipulative activity. We show methodologically how to identify periods when we might see suspicious activity in the order book.

The more broad contributions of our empirical work and its stark differences from other studies in the related literature can be summarised in several layers. First, we characterise the limit order book state putting together the majority of measures available in the financial microstructure. We explain how we remove unimportant variables from the analysis. Second, the state of art approach shows the importance of the limit order book information for detecting a fraudulent trading activity using machine learning techniques. Third, we show the model selection process among different machine learning techniques and further implement the robustness validation approach to test the model's efficiency, including various cross-validation methods. On top of that, as our main contribution, based on the previous layers and machine learning methodology, we introduce the RTSP measure, which could be implemented in real-life trading. Exchanges and regulators could benefit from the research as we show how to implement a real-time measure for the market surveillance system. Using our approach, we show how exchanges could identify periods with suspicious fraudulent activity in the order book. Having information on previous suspicious orders, which exchanges typically identify post factum after analysing combined information of traders id's and their historical trading strategies, exchanges could use our method to indicate the risk of similar fraudulent activity in the next tick and act accordingly, depending on their goals. Implemented by the exchange as a part of the surveillance system, our method could lead to significant market quality improvement and a reduction of manipulating cases.

The designed measure has an essential future as an adjustment possibility depending on the asset, market microstructure, and the type of manipulative activity. So, as we construct the RTSP measure using ML algorithms trained on the given dataset, the measure may forecast spoofing events and other fraudulent activities on different financial markets. The most valuable part of the research is showing how to approach such market requests step by step.

Our work does not consider the model to account for market shocks, news, dividend activity or similar macro events. We focus on intraday high-frequency data, where spoofing manipulations tend to occur. However, on top of our model, the model users may add one more layer of accumulation of spoofing activities with the overall market performance, intraday shocks and other events. Nevertheless, this analysis provides valuable information for regulators, market data vendors, and exchanges who aim to prevent their clients from artificial, manipulative orders to improve the health of market quality.

Researchers may further use the designed methodology for other disruptive activities, playing with market variables and model parameters. Moreover, the same approach could be implemented on any financial market and asset; hence, the models capture essential features such as volatility, tick size and trading activity and automatically adjust depending on the training dataset. That flexibility and the robustness of the developed model make our research an essential brick in market microstructure literature.

2.7 Appendices

2.7.1 Appendix 1. Lasso regularization for variables choice

Feature	Coef	Std.Err.	z	P> $\|z\|$	[0.025	0.975]
IMB	-0.3212	0.10522	-3.05267	0.002268	-0.52743	-0.11497
IMB_order	-0.00843	0.027257	-0.30911	0.75724	-0.06185	0.044997
FRA	-1.60787	0.144537	-11.1243	9.56E-29	-1.89116	-1.32458
FRB	0.880062	0.14531	6.056448	1.39E-09	0.595259	1.164864
QS	383.8753	34.00574	11.28854	1.49E-29	317.2252	450.5253
ES	122.714	49.63106	2.472524	0.013416	25.43888	219.9891
IMB_0	0.006483	0.014956	0.433459	0.664682	-0.02283	0.035795
IMB_1	0.097356	0.014714	6.616393	3.68E-11	0.068516	0.126196
IMB_2	0.077485	0.015252	5.080211	3.77E-07	0.047591	0.107378
IMB_3	0.078047	0.015426	5.059438	4.20E-07	0.047813	0.108282
IMB_4	-0.07434	0.015573	-4.77395	1.81E-06	-0.10487	-0.04382
IMB_5	-0.03868	0.01498	-2.58226	0.009816	-0.06804	-0.00932
UF_1ms_1	0.000422	0.000366	1.155149	0.24803	-0.00029	0.001139
UF_1ms_2	-0.00073	0.000496	-1.46944	0.141714	-0.0017	0.000243
UF_1ms_3	0.000657	0.000566	1.159823	0.246121	-0.00045	0.001766
UF_1ms_4	0.001032	0.000529	1.950864	0.051073	-4.80E-06	0.002068
UF_1ms_5	-0.00035	0.000502	-0.69271	0.48849	-0.00133	0.000636
UF_1ms_6	0.000565	0.000578	0.976952	0.328593	-0.00057	0.001698
UF_1ms_7	3.50E-05	0.000778	0.044997	0.96411	-0.00149	0.001559
UF_1ms_8	-0.00119	0.000701	-1.70177	0.088799	-0.00257	0.000181
UF_1ms_9	0.001496	0.000532	2.814629	0.004883	0.000454	0.002538
UF_1ms_10	-0.00107	0.000321	-3.32851	0.000873	-0.0017	-0.00044
UF_10ms_1	-0.00077	0.000741	-1.03415	0.301067	-0.00222	0.000686
UF_10ms_2	0.001162	0.000977	1.189622	0.234195	-0.00075	0.003076
UF_10ms_3	0.001405	0.001069	1.314504	0.188677	-0.00069	0.0035
UF_10ms_4	-0.00322	0.001053	-3.05752	0.002232	-0.00528	-0.00116
UF_10ms_5	0.003332	0.001106	3.012135	0.002594	0.001164	0.0055
UF_10ms_6	-0.00368	0.001174	-3.13422	0.001723	-0.00598	-0.00138
UF_10ms_7	0.002428	0.001261	1.924943	0.054237	-4.40E-05	0.0049
UF_10ms_8	0.001425	0.001161	1.227215	0.219742	-0.00085	0.0037
UF_10ms_9	-0.00373	0.000965	-3.86287	0.000112	-0.00562	-0.00184
UF_10ms_10	0.002529	0.000562	4.500931	6.77E-06	0.001428	0.00363
UF_50ms_1	0.007226	0.0018	4.015637	5.93E-05	0.003699	0.010754
UF_50ms_2	-0.00568	0.002153	-2.63767	0.008348	-0.0099	-0.00146
UF_50ms_3	0.000266	0.002007	0.132356	0.894703	-0.00367	0.004199
UF_50ms_4	0.000947	0.002137	0.443233	0.657597	-0.00324	0.005137
UF_50ms_5	-0.00443	0.00232	-1.90867	0.056305	-0.00897	0.000119
UF_50ms_6	0.01228	0.00243	5.053454	4.34E-07	0.007518	0.017043
UF_50ms_7	-0.0053	0.002431	-2.1814	0.029154	-0.01007	-0.00054
UF_50ms_8	-0.00452	0.001985	-2.27897	0.022669	-0.00841	-0.00063
UF_50ms_9	0.00126	0.001604	0.785475	0.432175	-0.00188	0.004405
UF_50ms_10	-0.001	0.001095	-0.91578	0.359781	-0.00315	0.001143
UF_100ms_1	-0.00551	0.001667	-3.30508	0.000949	-0.00878	-0.00224
UF_100ms_2	0.004442	0.002013	2.20594	0.027388	0.000495	0.008388
UF_100ms_3	-0.00302	0.001863	-1.61985	0.105264	-0.00667	0.000634
UF_100ms_4	0.002616	0.001967	1.330229	0.183443	-0.00124	0.006471
UF_100ms_5	0.000583	0.002108	0.276438	0.782211	-0.00355	0.004714
UF_100ms_6	-0.00897	0.002183	-4.11074	3.94E-05	-0.01325	-0.00469
UF_100ms_7	0.004389	0.002155	2.036936	0.041656	0.000166	0.008612
UF_100ms_8	0.001557	0.001819	0.856017	0.391988	-0.00201	0.005123
UF_100ms_9	0.004167	0.00153	2.722724	0.006475	0.001167	0.007166
UF_100ms_10	-0.0023	0.001034	-2.22182	0.026296	-0.00432	-0.00027
UF_600ms_1	-0.00046	0.000486	-0.93692	0.348798	-0.00141	0.000497
UF_600ms_2	0.000218	0.000583	0.37335	0.708888	-0.00092	0.00136
UF_600ms_3	0.001369	0.000599	2.283564	0.022397	0.000194	0.002544
UF_600ms_4	-0.00127	0.000594	-2.13249	0.032967	-0.00243	-0.0001
UF_600ms_5	0.000823	0.000608	1.35409	0.175708	-0.00037	0.002015
UF_600ms_6	0.000205	0.00062	0.330751	0.740833	-0.00101	0.001421
UF_600ms_7	-0.00108	0.000626	-1.7199	0.08545	-0.0023	0.00015
UF_600ms_8	0.001556	0.000608	2.559037	0.010496	0.000364	0.002748
UF_600ms_9	-0.00257	0.000598	-4.30355	1.68E-05	-0.00374	-0.0014
UF_600ms_10	0.00148	0.000354	4.175298	2.98E-05	0.000785	0.002175
VOL_1_1min	-332.465	112.7972	-2.94746	0.003204	-553.544	-111.387
VOL_1_2min	67.88061	179.9601	0.377198	0.706026	-284.835	420.596

VOL_1_5min	-55.8454	271.0023	-0.20607	0.836736	-587	475.3093
VOL_1_10min	36.52509	286.0571	0.127685	0.898399	-524.137	597.1868
VOL_2_1min	6.095202	1.104971	5.516166	3.46E-08	3.929499	8.260905
VOL_2_2min	-5.96658	1.590235	-3.75201	0.000175	-9.08338	-2.84978
VOL_2_5min	-4.02738	2.149298	-1.87381	0.060956	-8.23993	0.185166
VOL_2_10min	19.84205	2.331745	8.509528	1.75E-17	15.27192	24.41219
VOL_4_1min	2.754861	6194.337	0.000445	0.999645	-12137.9	12143.43
VOL_4_2min	2.870529	10254.95	0.00028	0.999777	-20096.5	20102.19
VOL_4_5min	2.577828	13808.9	0.000187	0.999851	-27062.4	27067.52
VOL_4_10min	3.352821	11427.39	0.000293	0.999766	-22393.9	22400.62
Hour	0.019725	0.002944	6.700296	2.08E-11	0.013955	0.025495
DistNorm50	20.25476	4455.851	0.004546	0.996373	-8713.05	8753.562
DistNorm20	-199.294	20273.16	-0.00983	0.992157	-39934	39535.38
DistNorm10	-194.514	18044.84	-0.01078	0.991399	-35561.7	35172.72
DistNormClean50	149.5656	27.31266	5.476056	4.35E-08	96.03381	203.0975
DistNormClean20	-176.587	31.28896	-5.64374	1.66E-08	-237.912	-115.262
DistNormClean10	93.17991	48.21986	1.932397	0.053311	-1.32927	187.6891
DistNormVol50	0.553286	0.078023	7.091356	1.33E-12	0.400364	0.706207
DistNormVol20	0.089309	0.030394	2.938428	0.003299	0.029739	0.14888
DistNormVol10	-0.01351	0.018983	-0.71187	0.476548	-0.05072	0.023692
DistNormVolClean50	-0.61615	0.09366	-6.57852	4.75E-11	-0.79972	-0.43258
DistNormVolClean20	-0.05694	0.036533	-1.55872	0.119063	-0.12855	0.014659
DistNormVolClean10	-0.08247	0.025564	-3.22598	0.001255	-0.13257	-0.03236
IMB_order_delta	-0.25393	0.188881	-1.3444	0.178818	-0.62413	0.116268
FRA_delta	-0.04651	0.375481	-0.12387	0.901414	-0.78244	0.689416
FRB_delta	-0.97608	0.377774	-2.58377	0.009773	-1.71651	-0.23566
QS_delta	70.65095	65.19404	1.083702	0.278497	-57.127	198.4289
IMB_order_delta_t2	0.366445	0.155094	2.362738	0.01814	0.062468	0.670423
FRA_delta_t2	-0.08441	0.315573	-0.26749	0.789091	-0.70293	0.534099
FRB_delta_t2	-0.36953	0.31831	-1.16091	0.245679	-0.99341	0.254347
QS_delta_t2	45.82324	57.36325	0.798826	0.424391	-66.6067	158.2531

2.7.2 Appendix 2. Correlation matrix between predictors

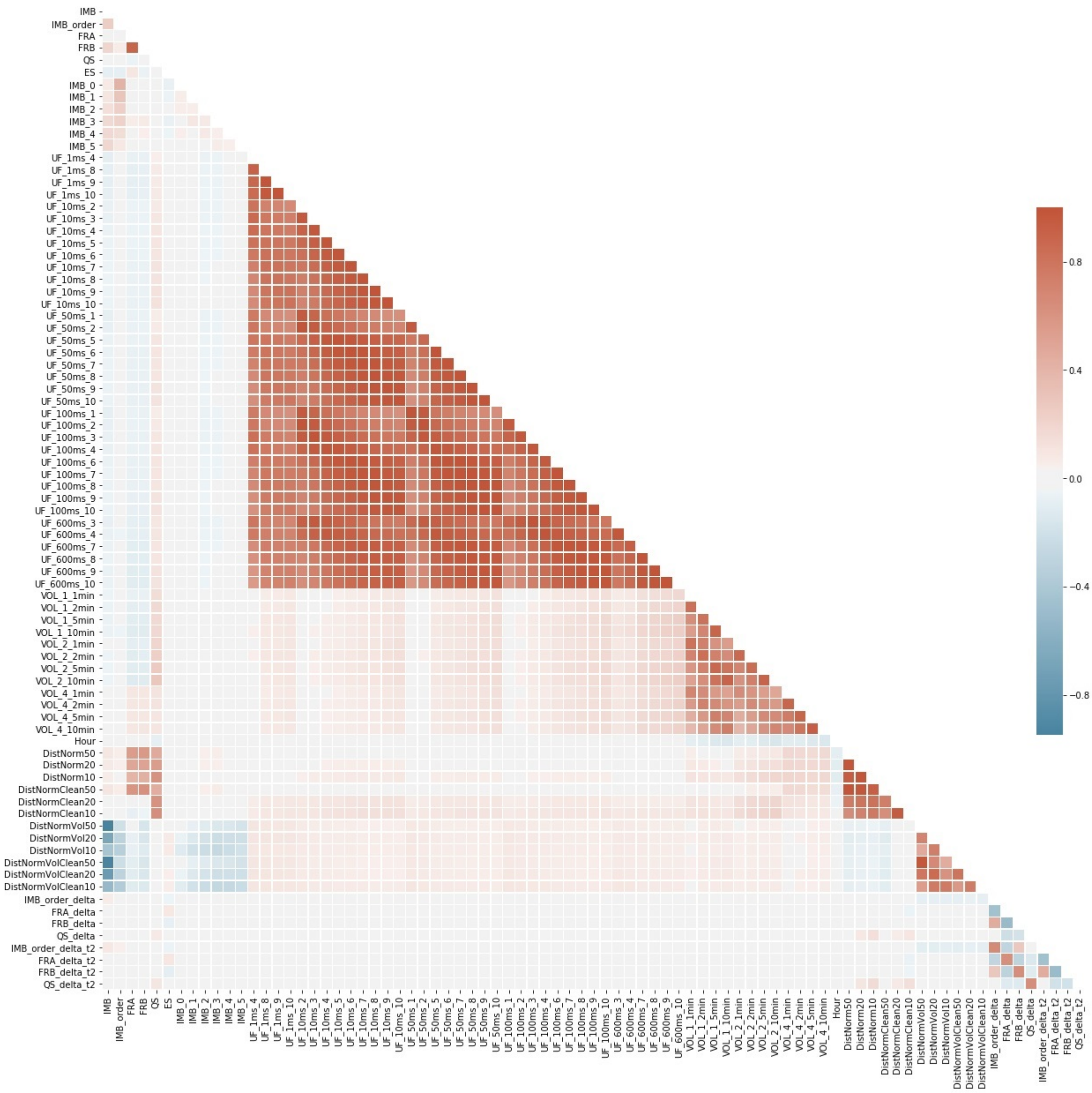


FIGURE 2.3: Correlation matrix between predictors

2.7.3 Appendix 3. List of features

We remove from the further analysis unnecessary features using the elastic net (Enet) variable selection method with coefficients equal to 0. Unnecessary features are the following: QS, ES, VOL_1_1min, VOL_1_2min, VOL_1_5min, VOL_1_10min, VOL_3_1min, VOL_3_2min, VOL_3_5min, VOL_3_10min, QS_delta, QS_delta_t2. Variable 'Hour' leads to over-fitting of the model due to the absence of False orders in the first and last trading hours. 'VOL_2_1min' variable has many unidentified values. Therefore, we remove both features 'Hour' and 'VOL_2_1min'.

The final selection of predictors for further analysis is the following: IMB, IMB_order, FRA, FRB, IMB_0, IMB_1, IMB_2, IMB_3, IMB_4, IMB_5, UF_1ms_4, UF_1ms_8, UF_1ms_9, UF_1ms_10, UF_10ms_2, UF_10ms_3, UF_10ms_4, UF_10ms_5, UF_10ms_6, UF_10ms_7, UF_10ms_8, UF_10ms_9, UF_10ms_10, UF_50ms_1, UF_50ms_2, UF_50ms_5, UF_50ms_6, UF_50ms_7, UF_50ms_8, UF_50ms_9, UF_50ms_10, UF_100ms_1, UF_100ms_2, UF_100ms_3, UF_100ms_4, UF_100ms_6, UF_100ms_7, UF_100ms_8, UF_100ms_9, UF_100ms_10, UF_600ms_3, UF_600ms_4, UF_600ms_7, UF_600ms_8, UF_600ms_9, UF_600ms_10, VOL_2_2min, VOL_2_5min, VOL_2_10min, DistNorm50, DistNorm20, DistNorm10, DistNormClean50, DistNormClean20, DistNormClean10, DistVol50, DistVol20, DistVol10, DistVolClean50, DistVolClean20, DistVolClean10, IMB_delta, FRA_delta, FRB_delta, IMB_delta_t2, FRA_delta_t2, FRB_delta_t2.

2.7.4 Appendix 4. Diebold-Mariano test

First model	Second model	Diebold-Mariano statistic	P-value
Random Forest	Logistic regression	-54.09	0.00
Random Forest	SVM	-48.59	0.00
Random Forest	KNN	-24.51	0.00
Random Forest	Naive Bayes	-62.61	0.00
Random Forest	SGD	-59.31	0.00
Random Forest	Decision Tree	-22.97	0.00
Random Forest	Gradient Boosting	-26.32	0.00
Random Forest	XGBoost	-26.81	0.00
Random Forest	LSTM with 10 layers	-50.05	0.00
Random Forest	GRU	-52.20	0.00
Random Forest	Percepton	-54.35	0.00
Decision Tree	Logistic regression	-37.09	0.00
Decision Tree	SVM	-26.91	0.00
Decision Tree	KNN	-1.83	0.07
Decision Tree	Naive Bayes	-37.16	0.00
Decision Tree	SGD	-36.24	0.00
Decision Tree	Gradient Boosting	-1.66	0.10
Decision Tree	XGBoost	-2.12	0.03
Decision Tree	LSTM with 10 layers	-33.51	0.00
Decision Tree	GRU	-37.10	0.00
Decision Tree	Percepton	-34.42	0.00
XGBoost	Logistic regression	-33.59	0.00
XGBoost	SVM	-25.31	0.00
XGBoost	KNN	0.25	0.80
XGBoost	Naive Bayes	-39.49	0.00
XGBoost	SGD	-36.87	0.00
XGBoost	Gradient Boosting	1.63	0.10
XGBoost	LSTM with 10 layers	-29.46	0.00
XGBoost	GRU	-32.16	0.00
XGBoost	Percepton	-33.03	0.00

TABLE 2.17: Diebold-Mariano test. The table shows the test result that compares the models' predictive qualities in pairs. We compare Random Forest, Decision Tree and XGBoost with other ML models. The null hypothesis is that models are equal. If the p-value is less than 0.05, then the models are not equal. We mark the results with the '-' sign if the first model is better than the second model; the '+' sign shows the opposite.

2.7.5 Appendix 5. Model Confidence Set test

Panel A							
Model	Elimination result						
Logistic regression	eliminated						
SGD	eliminated						
GRU	eliminated						
Naive Bayes	eliminated						
LSTM with 10 layers	eliminated						
SVM	eliminated						
XGBoost	eliminated						
Gradient Boosting	eliminated						
KNN	eliminated						
Random Forest	eliminated						
Superior Set Model created	Rank_M	v_M	MCS_M	Rank_R	v_R	MCS_R	Loss
Random Forest	1	-28.06988	1	1	-28.06988	1	0.2332108
p-value : [1] 0							

Panel B							
Model	Elimination result						
Logistic regression	eliminated						
Naive Bayes	eliminated						
GRU	eliminated						
SGD	eliminated						
LSTM with 10 layers	eliminated						
SVM	eliminated						
Superior Set Model created	Rank_M	v_M	MCS_M	Rank_R	v_R	MCS_R	Loss
KNN	3	0.6730877	0.769	3	1.886552	0.212	0.3200074
Decision Tree	1	-2.3702309	1.000	1	-1.876958	1.000	0.3126542
Gradient Boosting	2	0.6042045	0.813	2	1.876958	0.223	0.3192163
XGBoost	4	1.7018618	0.159	4	2.368780	0.081	0.3210779
p-value : [1] 0.159							

TABLE 2.18: Model Confidence Set test. The table shows the implementation of the Model Confidence Set (MCS) procedure for model comparison at the 95% confidence level (alfa is 0.05) and MAE (mean absolute error) loss function. Panel A shows that comparing all the models we use in the research, Random Forest is the best model. Panel B shows the results of the model's comparison test, excluding Random Forest, to find the second-best model.

2.7.6 Appendix 6. Optimization of the parameters for ML models

To optimize the chosen models, we tune the parameters while training models on 75% of the data and testing on 25% of the data. For Random Forest, increasing 'max_depth' to 41 and 'n_estimators' to 600 improves the accuracy. Further increasing 'max_depth' does not improve the results, and an increase of 'n_estimators' to 1000 improves the accuracy by 0.5%; however, it dramatically raises the complexity of the models, causing a decrease in model performance. For XGBoost, we choose 'n_estimators' of 500, as the increase to 1000 adds only 1.5% to the model's accuracy, raising the complexity and decreasing the performance speed and decreasing 'n_estimators' to 100 results in a significant drop of the accuracy by 5%. Decreasing 'max_depth' below 10 results in a drop of the accuracy. For Decision Tree, 'max_depth' over 5 does not yield the results; also, increasing 'min_samples_leaf' over 3 does not improve results either.

So, we choose the following parameters for selected machine learning models.

Random Forest parameters:

$clf = RandomForestClassifier\ (random\_state = 0$
$n\_estimators = 600,\ min\_samples\_split = 5,$
$min\_samples\_leaf = 1,\ max\_features =' sqrt',$
$max\_depth = 41,\ bootstrap = False)$

XGBoost parameters:

$clf = xgb.XGBClassifier\ (learning\_rate = 0.01,$
$n\_estimators = 500,$
$max\_depth = 10,$
$min\_child\_weight = 1,$
$gamma = 0.1,$
$subsample = 0.8,$
$colsample\_bytree = 0.8,$
$objective =' binary : logistic',$
$nthread = 4,$
$scale\_pos\_weight = 1.5$,random_state = 0)

Decision Tree parameters:

$clf = DecisionTreeClassifier$
$(random\_state = 0, max\_depth = 5, min\_samples\_leaf = 3)$

2.7.7 Appendix 7. Expanding and rolling validation

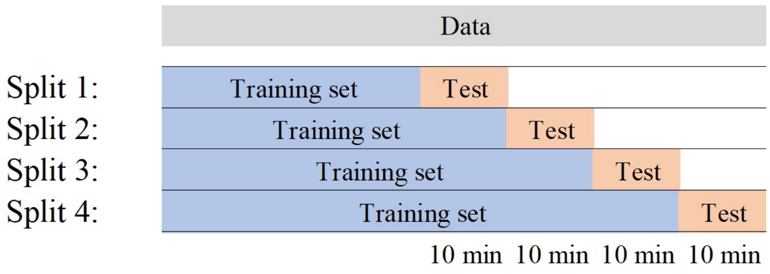


FIGURE 2.4: Illustration of the expanding validation approach

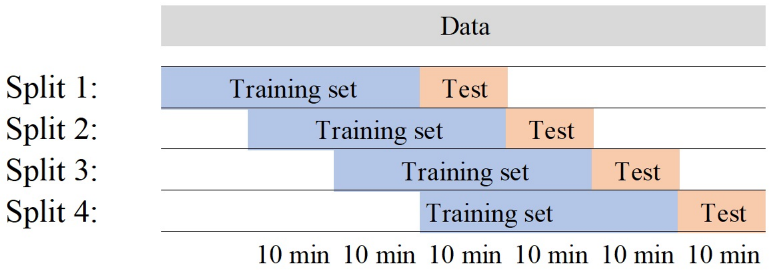


FIGURE 2.5: Illustration of the rolling validation approach

Chapter 3

Informed Trading in Futures Market

3.1 Introduction

Researchers in finance question how the behaviour of different types of investors and their interaction in the market affect returns. As price responses signal informed trades, consistent profits gained from positions or trading activity indicate who is informed. Researchers detect the presence of informed traders from price responses to order flow, which is known to be one of the key micro-level price determinants (Love & Payne 2008). Most previous studies have analysed informed trading with data on different investor types in the foreign exchange, equity and bond markets. Only a little research is done on the futures market due to data limitations on the investor type indication.

This paper contributes to the debate on informed trading by exploiting comprehensive data of the future market from the Moscow Exchange (MOEX) with the identification of the investor type who initiated the trade. The paper addresses several market microstructure and market design questions. First, does customer or investor order flow in the futures market contain predictive information? How do different trading practices by investor types affect market outcomes differently? Second, how does risk sharing take place in the developing futures market? Do investors systematically trade in opposite directions, or is their trading activity positively correlated? Third, what characterises different investors' trading behaviour? By answering these questions, we can improve our understanding of what ultimately drives end-users' demand for futures contracts. We aim to answer these questions both from an intraday point of view and a longer investment perspective.

We tackle these questions empirically using a dataset covering 20 months in 2022 and 2021 of every trade on the most liquid 27 futures contracts on MOEX. The data includes information on the client or investor type who initiated the trade. Clients are grouped into six categories: corporate traders, dealers, retail traders, non-residents, institutional traders and unidentified traders. Thus, we cover the trading behaviour of various investors with specific unique features and are quite heterogeneous in their motives for market participation and use of information.

We analyse buying and selling trading volume by each investor group and order flows as their net trading volume. Cespa et al. (2022) find volume helps predict next-day currency returns and argue that the predictive information in volume is different from that contained in order flow. In our analysis, buyer- and seller-initiated trading volume and order flow lead to similar outcomes.

Our results show that the retail traders' order flow has predictive power for the returns in the futures market for the next trading period in short-term time dimensions of 30 and 60 seconds. In longer time frequencies, the trading behaviour of retail traders loses its predictive power. Corporate clients tend to trade in the opposite direction to retail traders, and their behaviour does not show informativeness. When the corporate client buys futures, the market goes down.

Analysing longer intraday forecast performance, we find that 60-second order flow signals from corporate investors, retail traders, and institutional investors are informative for future returns for the following ten periods or ten minutes; however, the first minute has the highest return. 10-minute order flows also show longer horizon forecast performance, up to eight minutes after the signal formation. However, this is only true for corporate clients, retail traders and institutional investors.

Contemporaneous analysis reveals that non-residents, retail traders, and institutional investors pushing the prices up in the current period in a high-frequency time dimensions. Corporate clients and dealers trade in the opposite direction to the market.

Relying on the portfolio approach, we find the impact of institutional and retail traders on futures intraday return, knowing that institutional traders have relatively significant mean volume per trade[1], while the most significant market share belongs to retail traders. Neither dealers nor corporate clients process the fundamental information and have no future price predictive power.

The order flow information becomes insignificant after the first day for every client type. Hence, the information in daily flows is short-lived and incorporated into returns relatively quickly. The findings align with the literature on FX order flow, even though we run the analysis on another market and asset class. Menkhoff et al. (2016) find that the order flow is informative for the first two-three days, while our results show only one-day order flow forecasting power. Our findings' decrease in the time horizon might be associated with the rising trading speed over the last decade.

Another contribution of our paper is that the flow signal differs for daily and intraday trading, and the analysis needs to be built separately for intraday frequencies such as high-frequency trading and algorithmic trading, and daily trading.

We also contribute to the discussion about retail investor's behaviour and future returns. Firstly, we find a positive relationship between retail flows and future prices, which could be seen from the perspective that retail investors drive future returns. As we run an analysis on the futures market that is by definition driven by the price of the underlying asset, we conclude that retail investors are informed about future price movements of the underlying asset (Kaniel et al. 2012, Kelley & Tetlock 2013, Boehmer et al. 2021). Kaniel et al. (2012), Barrot et al. (2016) claim that retail investors are rewarded for providing liquidity to institutional investors. Barber et al. (2008) show that retail investors' order flows are positively autocorrelated and thus forecast short-term price pressure. The willingness of retail investors to provide liquidity and autocorrelated order flows may contribute to the short-term predictiveness of the retail order flow.

Secondly, we find that the retail selling volume anticipates negative returns in the next period. Nevertheless, we can not observe whether the seller-initiated trade was a close of the long position or a short sell; our results are most consistent with the information hypothesis that retail short sellers possess and act on unique information beyond that held by other investors. Under this theory, retail short selling predicts negative returns as prices of the underlying assets converge to their fundamental values, just as informed order flow predicts returns in models such as Kyle (1985). Moreover, our research using contemporaneous regressions shows institutional investors' order flows' impact on the current price movements, which is another evidence to support the argument that retail traders seem to follow order flow signals from large institutional trades and trade in the same direction in the current period.

With institutional investors' order flow, our research shows its impact on the current price movements. However, the market overreacts to large institutional trades, and we observe prices decrease in the next period.

Thirdly, we find that in a very short-term (30- and 60-second period analysis), retail traders provide liquidity to institutional investors that need to execute their trades immediately, as suggested by Kaniel et al. (2008). Also, our results indicate the liquidity provision role of retail trades to corporate clients.

[1]On average, institutional investors' mean trading volume is twice as high as retail traders' on MOEX data. Given the current automated and algorithm-driven market structure, the trade size of institutional investors could be split into several smaller-volume trades.

Even if each retail trader has imprecise information, the resulting signal is relatively precise when the information is aggregated through the trades of many individuals. We show that 46% of futures trading volume on MOEX is associated with retail traders. Also, retail traders have fewer constraints than institutional investors, at least with respect to diversification requirements or short-selling. Thus retail traders are better positioned to trade aggressively when they are informed.

While our main contribution highlights that retail traders and institutional investors correctly anticipate intraday and daily returns, we also identify ways in which these types of traders differ. Retail traders', institutional investors' and non-residents' order flows are good predictors for intraday returns in commodity futures, while only institutional investors' order flow positively predicts returns in stock futures. Intraday order flows of corporate clients, retail traders and non-residents positively predict returns on currency futures, while the order flow of institutional investors does not.

Our paper relates to a vast empirical literature studying information content of the flow of different investor types and orders executed by informed traders. Scholars investigate traders' informativeness on foreign exchange, stock, bond and derivative markets.

Cerrato et al. (2011) examines how customer order flows affect exchange rates using weekly net order flow. Evans & Lyons (2002) suggest that order flow is crucial for understanding how information is incorporated into exchange rates. Cespa et al. (2022) find volume helps predict next-day currency returns and argue that the predictive information in volume differs from that contained in the order flow. Menkhoff et al. (2016) show that order flows are highly informative about future exchange rates and provide significant economic value. They also find that different customer groups can share risk effectively and differ markedly in their predictive ability, trading styles, and risk exposure (Menkhoff et al. 2016).

Many scholars investigate the relationship between different investors' trading and stock returns. Several studies that analyse the trading behaviour of groups of traders focus only on subgroups. Kaniel et al. (2008), Kelley & Tetlock (2017) analyse retail trader's informativeness. Using proprietary data on millions of trades by retail investors, Kelley & Tetlock (2017) provides large-scale evidence that retail short selling predicts negative stock returns. Van Kervel & Menkveld (2019) show that institutional orders are mostly information-motivated.

It is known from the literature that order flow is positively associated with contemporaneous returns in all asset classes. Hasbrouck (1991*a*,*b*) investigates the effect on stock market, while Brandt & Kavajecz (2004) and Czech et al. (2021) - on the US and the UK bond markets respectively.

Fishe & Smith (2012) use daily positions of futures market participants to identify informed traders and show that the intraday informed group is dominated by managed money traders or hedge funds and swap dealers, with commercial hedgers under-represented.

Most recently, Menkveld & Saru (2022) analyse the trading behaviour of different types of traders using Euro STOXX 50 index futures data. Their results indicate that the classification of who is informed is endogenous to market conditions and that aggressive orders contain more information scaled by volume. Relative order informativeness differs between the agent's and principal's orders, with the agent's orders being more informative (Menkveld & Saru 2022)

Overall, we contribute to the literature by analysing intraday and daily customer flows and buying and selling volumes in the futures market. Our evidence shows that retail investors have an advantage over other market participants in a high-frequency environment and gain quick short-term returns from both predicting other investors' future demand and quick responses to the arrival of news. Through their active trading, retail investors help impound private information into future prices and expedite the price discovery process.

Additionally, our finding show that non-residents obtain information about price movements of currency and commodity futures and correctly predict intraday returns, which makes economic sense based on the specificity of the Russian market. In contrast to our other findings, corporate clients' flow positively predicts future returns in currency exchange rate futures. Corporate clients on the Russian market trade currency futures to hedge their exchange rate risk exposure.

The remainder of the paper is structured as follows. In Section 3.2, we provide institutional details on MOEX futures, describe the data, and give a statistical overview of the Russian futures market. In Section 3.3, we give an order flow definition and run correlation analysis. We present our results on relative differences in order flow informativeness by trading type in Section 3.4 and results of portfolio analysis in Section 3.5. We analyse sources of return predictability in Section 3.6 and conclude in Section 3.7.

3.2 Data

Our analysis focuses on futures from the Moscow Exchange, the largest exchange in Russia, operating trading markets in equities, bonds, derivatives, foreign exchange, money markets, and precious metals. Russian derivative market trading volume was RUB 4.3 trillion in February 2023, which equals USD 56 bn. MOEX futures trading volume was 6.4 mln contracts in March 2023, while trading volume in CME was around 30 mln contracts.

The Russian derivatives market is relatively new, with contracts of different liquidity. Despite its size and short history, the future market is sophisticated and transparent. The total Russian market size has risen dramatically over the last ten years, from 450 trillion rubles in 2012 to 1100 trillion rubles in 2021. The Moscow Exchange is the third bond trading market in the world in by dollar trading volume after CME Group (USA) and BME (Spain), the 11th future trading market and the 26th stock market. Moreover, MOEX is the 15th market in world by capitalization (5.3 bn USD), straight after TMX Group (Canada) with 5.7 bn USD [2].

We employ the dataset based on every trade for 27 most liquid futures contracts traded on MOEX [3] from January 6, 2020, to September 30, 2021, excluding April 2021[4], a total of 407 trading days. Hence, in contrast to much earlier literature, we analyse daily and intraday data. Table 3.14 contains information on the average number of trades per minute and the average trading ruble volume per minute for each contract.

[2]www.moex.com

[3]The data set includes future on exchange rates: USDRUB (Si), EURRUB (Eu), EURUSD (Ed); futures on the index of Russian Trading System (RTS), the index on RTS mini (RTSM), the index on Moscow Exchange (MIX), index on Moscow Exchange mini (MXI), SP 500 ETF trust (SPYF); futures on commodities: Brent oil (Br), gold (GOLD), silver (SILV) platinum (PLT), natural gas (NG); futures on stocks: Sberbank (SBRF), Gazprom (GAZR), VTB bank (VTBR), Norilsk Nickel (GMKN), Lukoil (LKOH), Rosneft (ROSN), Aeroflot (AFLT), Magnit (MGNT), Alrosa (ALRS), Sberbank preferred shares (SBPR), Yandex (YNDF), Tatneft (TATN), Surgutneftegas (SNGR), NLMK group (NLMK).

[4]The provided data does not contain one month, April 2021, for unknown reasons. No noteworthy events happened during the missing month; market volatility, spreads, and price return did not deviate significantly from its annual means. Therefore, the absence of one month in the data does not affect the research outcomes in a meaningful way.

The data set includes information on the price, the ruble volume, the direction of the trade (buy or sell), and an indication of the client who initiated the trade. There are six client groups: corporate traders ($CORP$); dealers (DLR), retail traders (RET), institutional traders ($INST$), non-residents (NER), and unidentified traders ($UNDEF$)[5]. Client groups show substantial heterogeneity in the motives of market participation[6], and these groups are likely to differ in their degree of information awareness and sophistication. Our data only contains information on customer-initiated trades, which means that this paper has nothing to say about the trading strategies of the market-makers.

We omit the morning (7:00-10:00) and evening (19:05 - 23:50) trading sessions with low trading volumes, leaving the regular trading hours between 10:00 and 18:45 when the stock market is open. The level of activity on the futures market is structurally different in those two periods than in the main body of the equity trading day.[7] Informed traders prefer to trade during high trading volume periods to mask their informativeness and limit their price impact (Jin et al. 2020). Our research applies the typical approach in the literature focusing on futures contracts with the closest maturity, where most of the trading volume is concentrated (Schlag & Stoll 2005).

In Figure 3.1, we show that retail traders account for 49% of the total trades in the Russian future market. The second biggest market share of 43% belongs to non-residents; however, data do not allow us to identify types of non-resident investors more specifically. Dealers' market share is 4.44%, corporate traders' market share is around 1%, and institutional traders' market share less then 1%.

In Figure 3.2, we show the dynamic of the market share measured in volume by investor type per month. We do not observe any significant difference between the total market share disaggregated by the investor type for all futures contracts and most five liquid contracts. However, non-residents have a higher share of up to 70% in commodities futures, such as Brent oil (BR) and silver (SILV). At the same time, the non-residents' trading volume decreases in less liquid stock futures, such as ALRS, MAGN, YNDX, and SGNG. For example, for MAGN, the market share of non-residents drops to 33%, while for YNDF to 9%. We also observe a rise in dealers' market share in stock futures. Mentioned differences show the interest in liquid futures for non-residents and retail traders and reasonable interest for non-residents in hedging their portfolios by trading commodity futures.

In Figure 3.3, we show a monthly dynamic of the mean ruble volume per trade disaggregated by the investor type. Corporate clients, dealers, and retail traders have higher trading volume in less liquid futures on stocks (e.g. ALSR, MGNT, NLMK, TATN), while institutional investors have higher trading volume in most liquid futures on the market, including commodity futures (GOLD, BR), index futures (MIX, RTS) and currency futures (EU, ED). Retail traders' average trade size is relatively small, while the market share is the highest after non-resident clients (Figure 3.2).

[5]It is essential to note that we do not have data on individual customers and hence cannot use any information on customers' identities. We have only data on client types.

[6]We ignore unidentified customers, however, we include them in the results for statistical and representation purposes.

[7]Morning and evening sessions have ten and three times less average trading volume per hour than regular trading sessions. On average, 83% of futures trading volume is executed during the regular trading hours, being 94% for the stock futures.

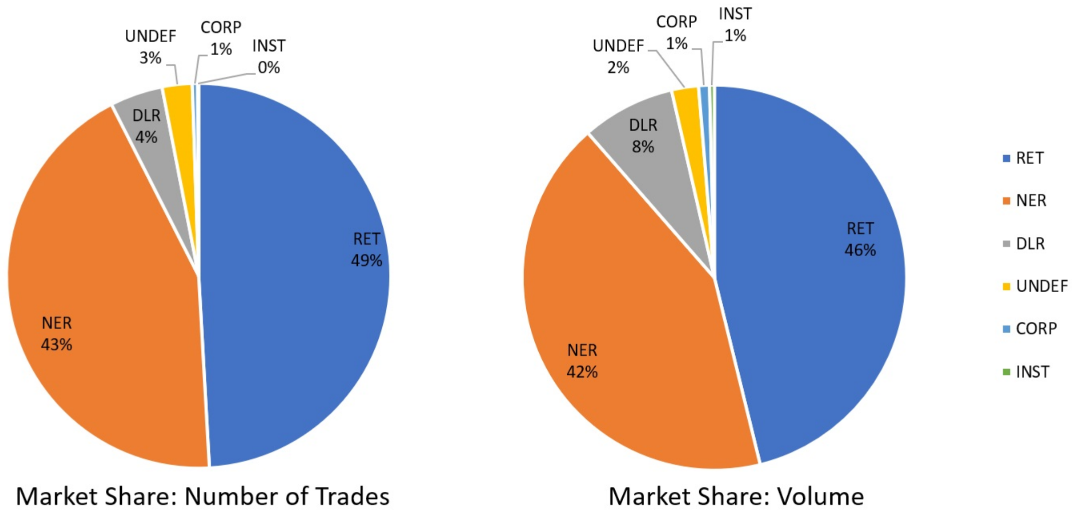


FIGURE 3.1: MOEX future market shares by investor type. This figure shows the breakdown of the total trading ruble volume and number of trades in the Russian futures market. The sample period is January 2020 to September 2021, excluding April 2021. Six investors types are defined as follows: corporate traders ($CORP$); dealers (DLR), retail traders (RET), institutional traders ($INST$), non-residents (NER), and unidentified traders ($UNDEF$).

3.3 Order flows definition and correlation analysis

3.3.1 Order flow definition

Order flow is defined as the difference between buyer-initiated and seller-initiated trading activity in a given market and, thus, corresponds broadly to what practitioners might describe as aggressive buying or selling pressure (Love & Payne 2008). Order flow explains contemporaneous price movements because it contains information about fundamentals or long-run risk premia that were previously dispersed among market participants (Lyons (1995), Perraudin & Vitale (1996), and Evans & Lyons (2002)). Menkhoff et al. (2016) study the information in order flows in the foreign exchange market, and measure order flows as net buying pressure against the USD, that is, the USD volume of buyer-initiated minus seller-initiated trades of a currency against the USD. Czech et al. (2021) examine trading by different investor types in government bond markets and calculate the order flow of each investor type as a difference in buy and sell volume by investor group scaled by the total trading volume of the investor group.

Cespa et al. (2022) find volume helps predict next-day currency returns and argue that the predictive information in volume differs from that contained in the order flow. Thus, in our research, we separately analyse buying and selling trading volume by each investor group. Furthermore, we also measure an order flow as a net trading volume which is the buyer-initiated volume minus the seller-initiated volume. We use order flow as the key micro-level price determinant, which is known to partially impound information that is publicly released to all market participants (Love & Payne 2008).

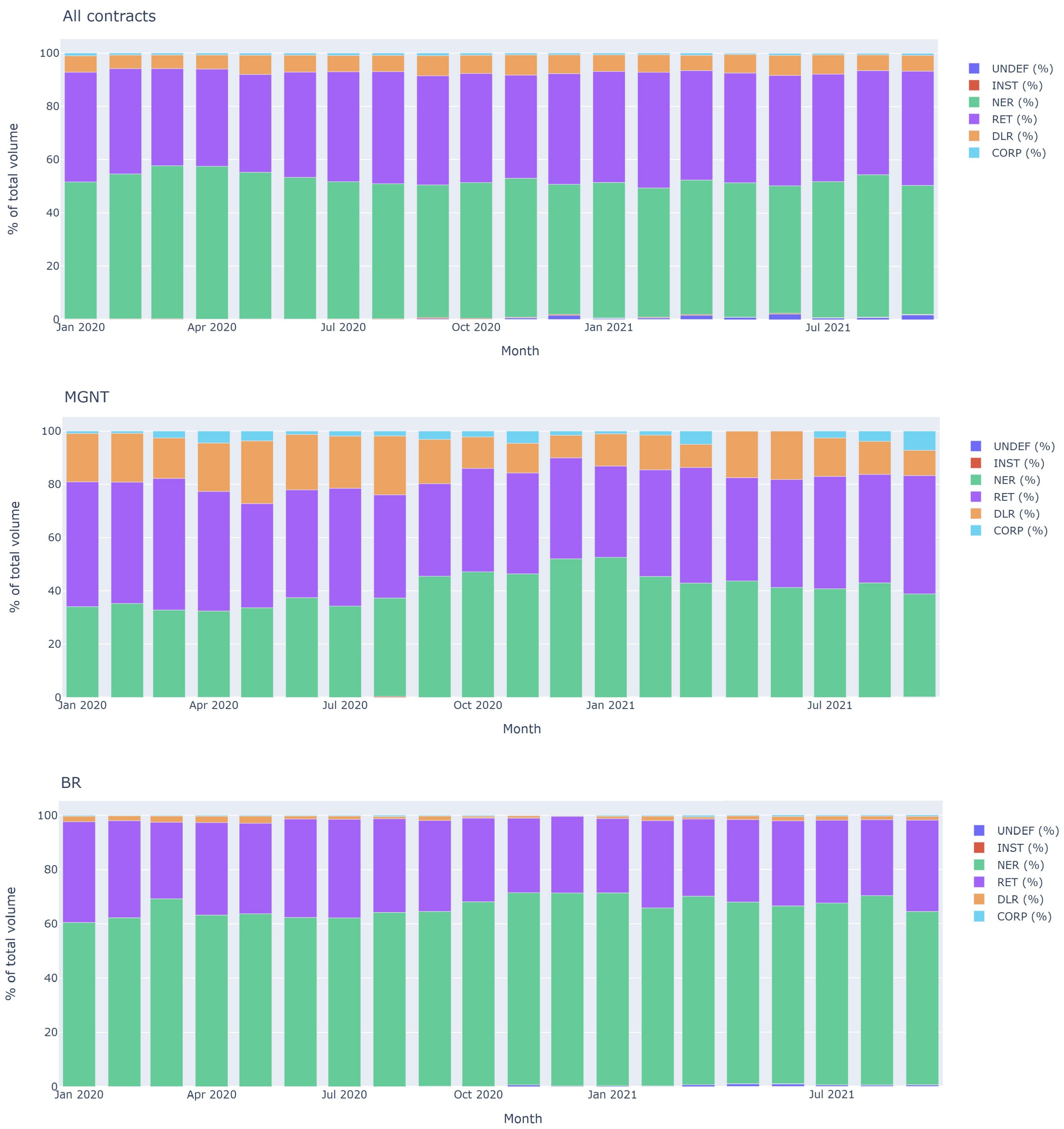


FIGURE 3.2: The monthly dynamic of the market share in ruble volume by the investor type. The top figure shows the breakdown of the total trading volume in the Russian futures market from January 2020 to September 2021, excluding April 2021. We compare the market share by investor types for the stock future contract (MGNT) in the middle and for the Brent oil future contract (BR) in the bottom. Six investors types are defined as follows: corporate traders ($CORP$); dealers (DLR), retail traders (RET), institutional traders ($INST$), non-residents (NER), and unidentified traders ($UNDEF$).

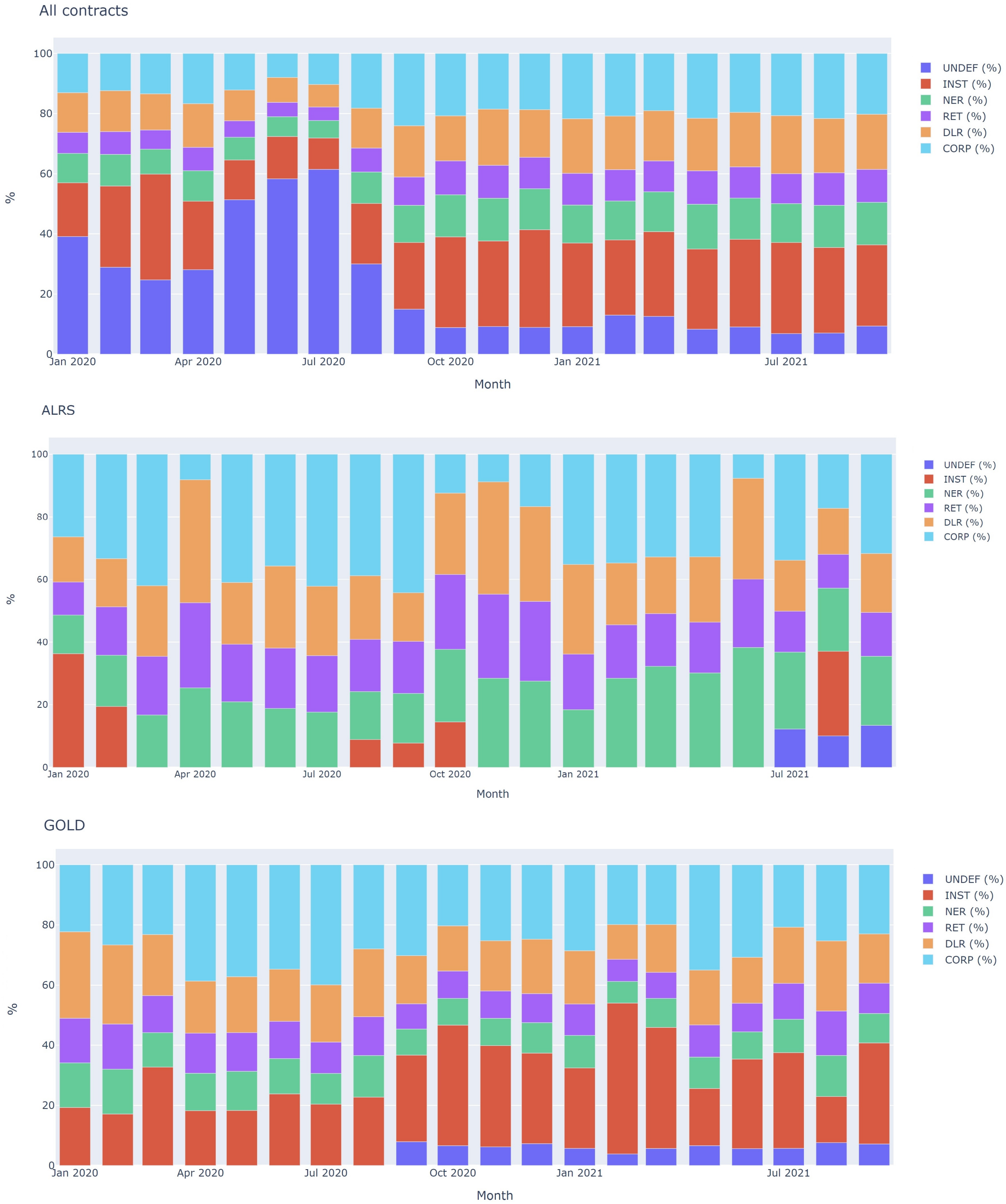


FIGURE 3.3: The monthly dynamic of the mean ruble volume per trade by the investor type. The top figure shows the average ruble volume per trade per client type for all future contracts from January 2020 to September 2021, excluding April 2021. We show the average trade size by the investor for the stock futures contract (ALRS) in the middle and for the gold futures contract (GOLD) in the bottom. Six investors types are defined as follows: corporate traders (*CORP*); dealers (*DLR*), retail traders (*RET*), institutional traders (*INST*), non-residents (*NER*), and unidentified traders (*UNDEF*).

Standardizing order flows

Futures on MOEX have significant variations in trading volume and hence a large absolute size in order flows compared to other future contracts, especially less liquid futures. The first five most traded futures have 95% of all trading volume in the market, so one cannot easily compare order flows across different futures. Some form of standardization is needed to make meaningful comparisons. We take this into account in our empirical analysis below.

So, for order flow standardization, we use the same logic as Menkhoff et al. (2016) and divide flows by their standard deviation to remove the difference in absolute order flow size across futures. As we make intraday analysis, we compute the standard deviation of order flows via a rolling scheme over a 5-day trading window.

Time frequencies

In our research, we are interested in analysing daily and intraday returns. We contribute to the literature by discussing the information content of the order flow signals measured in different time frequencies. Having the data of all trades for 20 months, we construct several time series depending on the time frequencies. We measure total buying and selling volumes and the order flow in time windows of 30 and 60 seconds, 10, 30, and 60 minutes. Moreover, separately, we measure buying and selling volumes and order flows for every trading day. Thus we conduct six separate analyses depending on the time frequency or time windows in which we measure the flows.

When we further run regressions, all the variables we use are calculated based on the time frequencies. So, if we run a regression with a 30-second frequency, we measure the 30-second returns. Similarly, further in portfolios analysis (Section 3.5), when we talk about periods, we use the same logic of 60-second, 10-minute and daily time windows, where one window is one period. For example, in the daily analysis, we show a post-formation return for 30 periods, which is 30 trading days.

3.3.2 Returns correlation

As we are interested in analysing daily and intraday returns, we start by checking the correlation between futures returns in different time frequencies.

Correlation matrices in Tables 3.4, 3.5 and Appendix 3.8.2 show that the shorter the time window inside the day, the less correlated futures returns are. So, 30-second futures returns have a very poor correlation. However, we observe a positive correlation between 60-minute returns of futures on the stock in the same sector and negative correlations between stocks and commodities futures returns. Daily returns correlate differently than intraday returns. The results show a positive correlation of daily returns in nearly most futures, and we do not observe a negative correlation between commodity and stock futures returns. That comes from the homogeneity of the market we analyse and shows the differences between daily and intraday investments.

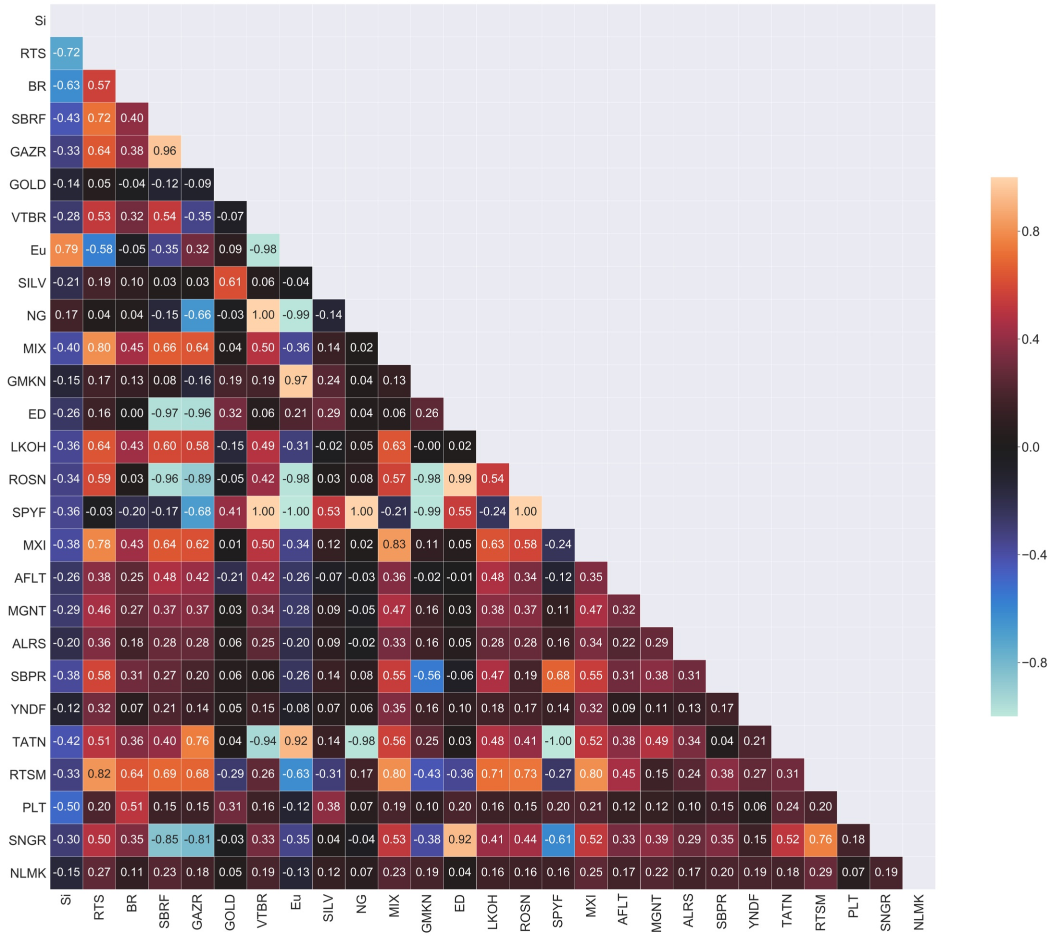


FIGURE 3.4: Correlation matrix between 30-second futures returns

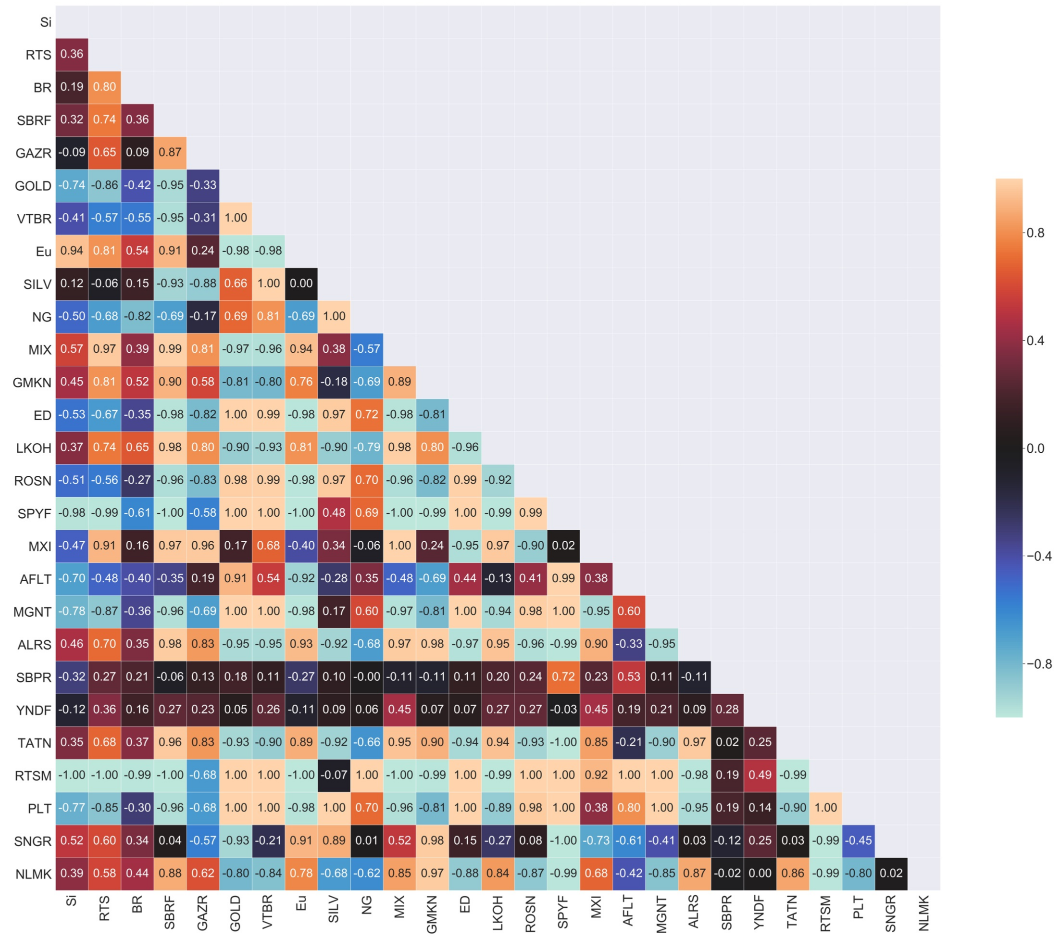


FIGURE 3.5: Correlation matrix between 60-minute futures returns

3.3.3 Order flows autocorrelation

Order flows disaggregated by client groups tend to be positively autocorrelated with the most significant correlation within the first lag (around 15-25%), followed by a drop in the correlation below 10% up to the fifth lag. The correlation is close to zero after the sixth lag. We found the most apparent autocorrelation in the 10-minute order flows while the least obvious in the daily order flows (Figure 3.6).

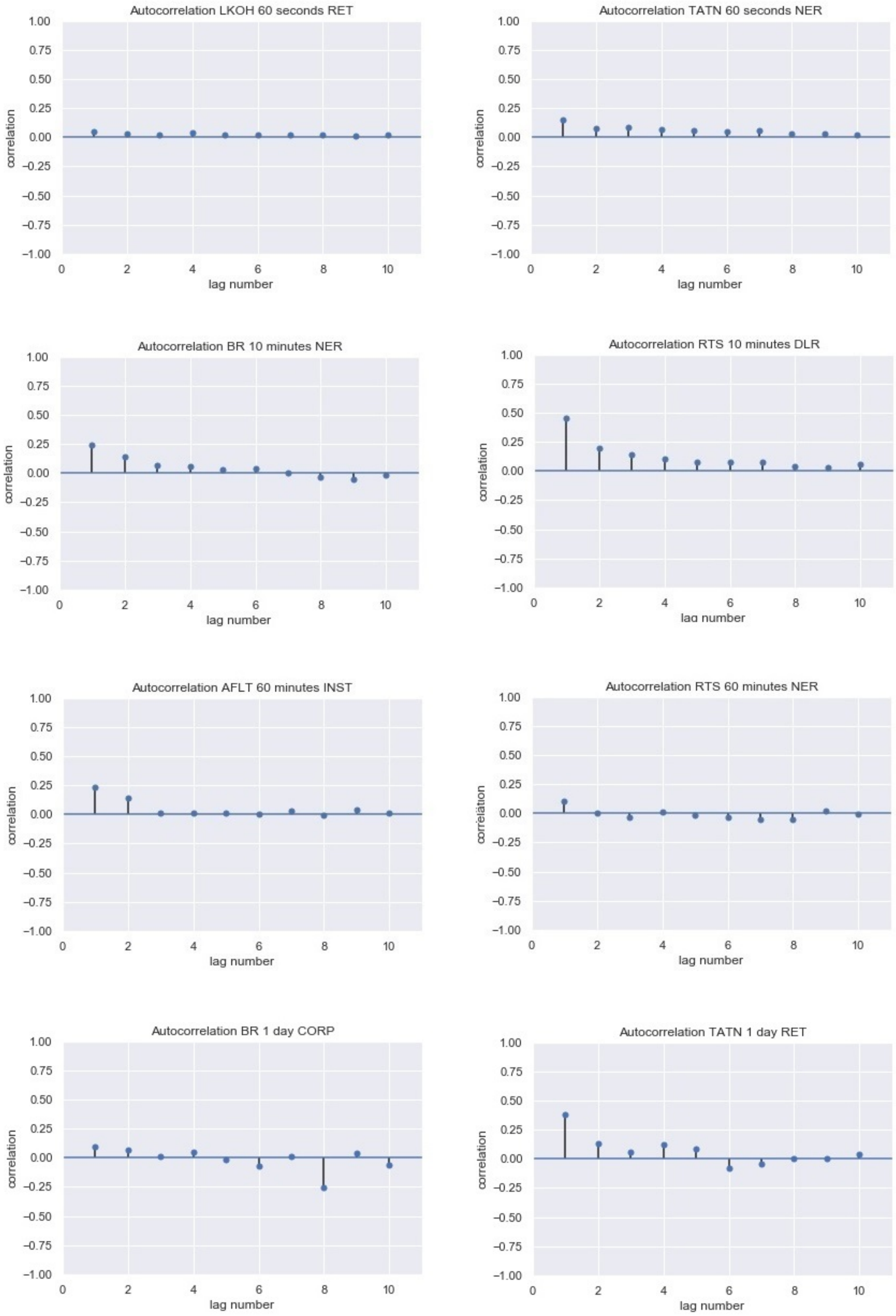


FIGURE 3.6: Order flows autocorrelation. The figure presents autocorrelation of 60-second, 10-minute, 60-minute, and daily disaggregated order flows. This figure shows several examples of different futures.

3.3.4 Order flows correlation

	60-second order flow (1)					10-minute order flow (2)				
	CORP	DLR	RET	INST	NER	CORP	DLR	RET	INST	NER
DLR	-0.17					-0.20				
RET	-0.85	0.02				-0.83	0.01			
INST	-0.51	0.03	0.12			-0.62	0.03	0.22		
NER	-0.01	-0.02	-0.26	-0.06		0.01	0.07	-0.27	-0.07	
UNDEF	-0.01	-0.02	-0.14	-0.08	0.04	0.00	-0.03	-0.16	-0.13	0.07
	60-minute order flow (3)					**daily order flow (4)**				
	CORP	DLR	RET	INST	NER	CORP	DLR	RET	INST	NER
DLR	-0.20					-0.21				
RET	-0.83	0.01				-0.85	0.03			
INST	-0.63	0.04	0.23			-0.56	0.03	0.15		
NER	0.02	0.10	-0.26	-0.08		-0.01	0.10	-0.19	-0.03	
UNDEF	0.02	-0.05	-0.17	-0.19	0.13	0.00	-0.03	-0.12	-0.17	0.13

TABLE 3.1: Orders flows correlation. The table shows (1) the 60-second time-frequency non-overlapping correlation of the investors' order flows. Similarly, (2) is a 10-minute, (3) 60-minute and (4) daily order flows correlations.

Table 3.1 shows the 60-second, 10-minute, 60-minute and daily time-frequency non-overlapping correlations of the investors' order flows. We do not observe significant differences between intraday and daily order flows correlations. Our analysis shows a highly negative (0.83-0.85) correlation between corporate and retail traders' order flows. Also, we observe a negative correlation (0.51-0.63) between corporate and institutional traders' order flows. Moreover, corporate clients order flows have a negative correlation with dealers' order flows. A similar negative correlation of 0.19-0.26 is found between retail traders' and non-residents' order flows. A positive correlation is observed between institutional and retail investors' order flows. So, corporate clients trade in the opposite direction to retail traders and institutional investors; however, we need to take into account that institutions' market share is much less significant in the Russian market than retail traders (Figure 3.2).

3.3.5 Order flows correlation over longer horizon

Given these findings, we look at the correlations among client groups' order flows over long horizons to determine if different types of investors tend to trade in the same or opposite directions. Figures 3.7, 3.8, 3.9 plot contemporaneous Pearson correlations between standardised order flows of different investor groups for horizons up to 60 periods. Average correlations between flows are based on the average correlation across all 27 futures contracts.

We keep the variety of time frequencies to check the difference in daily and intraday order flows correlations. A horizon of one period corresponds to 60-second, 10-minute or daily observations, whereas correlations for longer horizons are based on overlapping sums. The overlapping approach smooths the results because we have more data points. In Appendix 3.8.7, we report non-overlapping correlations over a long horizon, which show the same outcomes for intraday time-frequencies, while figure (3) with daily order flows correlations is noisy, which makes an overlapping approach preferable.

In an overlapping approach for the 10-minute time-frequency represented on Figure 3.8, we are calculating two periods, which is 20 minutes correlation from the 10-minutely data, then 30 minutes correlation from the 10-minutely data and further up to 60 periods or 600 minutes. That implies that sums are overlapping, which smooths the lines on the graph.

We do not observe a significant difference between 60-second and 10-minute order flows correlations (Figure 3.7, 3.8). We track positive correlations between retail and institutional traders in both time dimensions. Dealers and non-residents tend to trade in opposite directions for the first 60 seconds, and then we observe positive correlations in their order flows. Similarly, corporate clients and non-residents trade in opposite directions for the first 8 minutes, then trade in the same direction intraday. We observe the mirror situation with dealers and institutional traders. They trade in the same direction short-term for up to 45 minutes and then trade in the opposite direction. Other client pairs have negative order flows correlations inside the day.

Figure 3.9 shows daily order flows correlations over a longer horizon of up to 60 days. We observe opposite results for some pairs compared to intraday order flow correlations. Dealers and non-residents, corporate clients and non-resident have positive intraday order flows correlations while moving into the negative side over the longer horizon.

The pair of dealers and institutional investors tend to have positive order flow correlations up to 45 minutes, then negative intraday, and then going to the positive side over the next trading week. However, we observe a negative correlations over longer investment horizons after the second week.

Interestingly enough, the situation is with the pairs of corporate traders and dealers, institutional and non-resident traders, and dealers and retail traders. Their order flows correlations move from negative to positive after 12-14 trading days. Our results indicate that risk-sharing can take place in the short run due to a negative correlation between the order flows of different market segments. However, there is no risk sharing when clients hold the position longer than three weeks.

The only pair of retail and institutional traders hold a positive correlation of order flows during all 60 trading days, with a minor exception between the 18^{th} and 26^{th} trading day.

Two pairs of corporate and retail traders, corporate and institutional traders, have stable negative correlations of order flows over the longer horizon both inside the day and for the next 60 days. Retail and institutional investors have a positive correlation inside the day; however the correlation drops and becomes negative after twenty trading days, and then again positive after 25^{th} trading day. Also, correlations between order flows of corporate clients and dealers differ inside the day (negative) and daily (become positive after 15^{th} trading day).

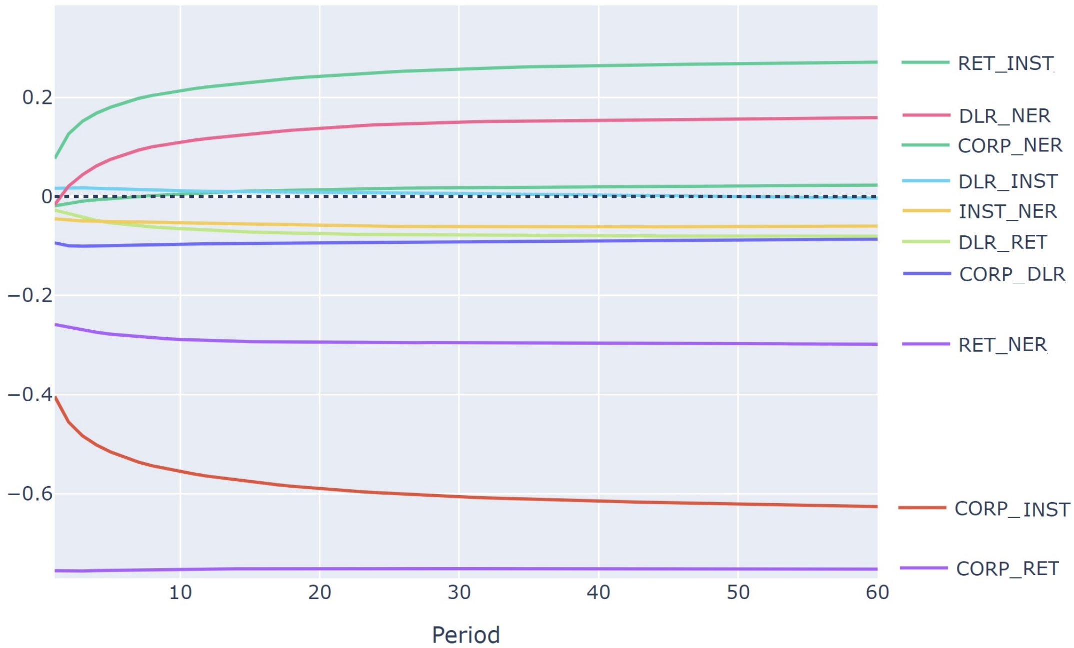


FIGURE 3.7: Correlations of 60-second customers' order flows over a long horizon. The figure plots contemporaneous Pearson correlations between standardised order flows of different investor groups for horizons up to 60 periods, where one period is a 60-second time window; periods overlap.

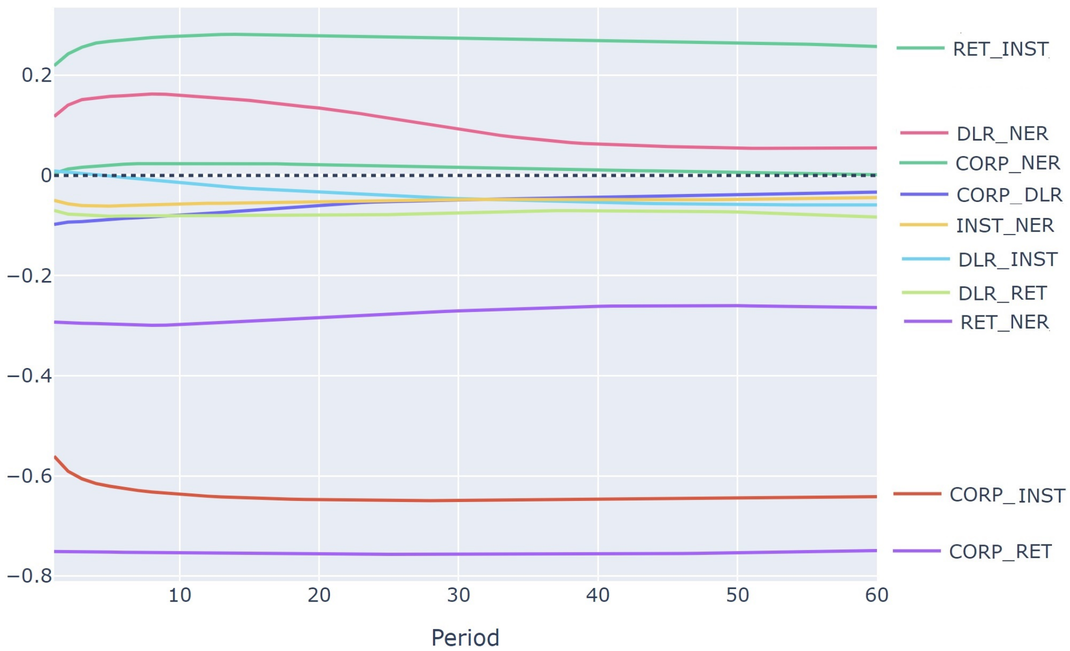


FIGURE 3.8: Correlations of 10-minute customers' order flows over a long horizon. The figure plots contemporaneous Pearson correlations between standardised order flows of different investor groups for horizons up to 60 periods, where one period is a 10-minute time window; periods overlap.

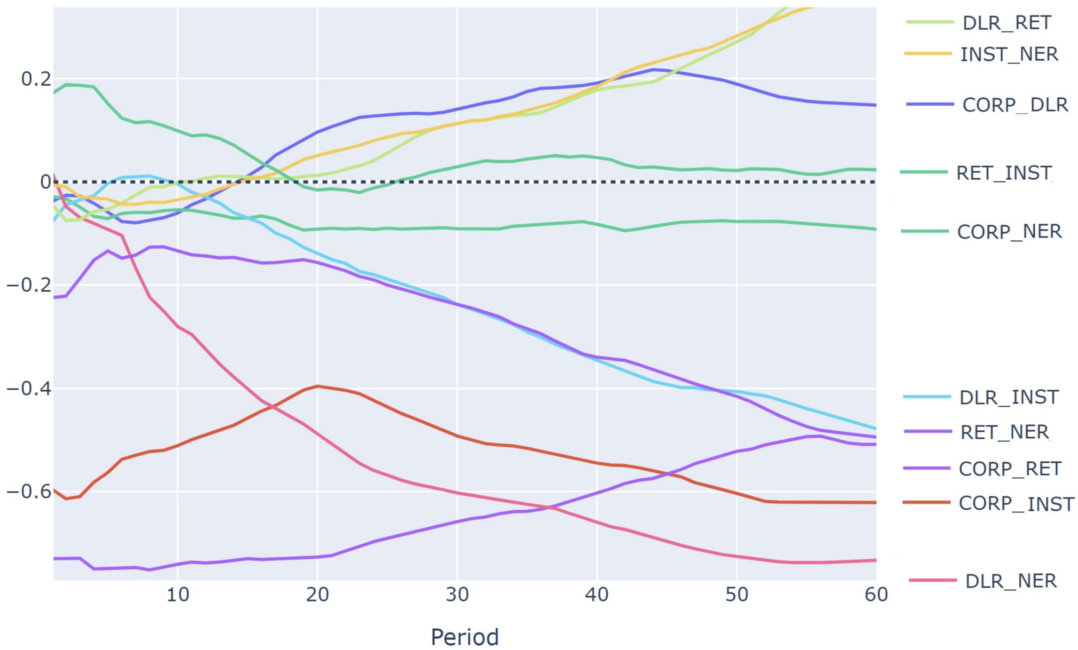


FIGURE 3.9: Correlations of daily customers' order flows over a long horizon. The figure plots contemporaneous Pearson correlations between standardised order flows of different investor groups for horizons up to 60 periods, where one period is a daily time window; periods overlap.

3.4 Predictive power of flows

To explore the price impact of different investor types on the futures market, we estimate models with buying and selling volumes and order flows of different investor or client types as dependent variables. For all models, we analyse intraday and daily returns. We are interested in high-frequency periods such as 30 and 60 seconds for intraday returns and estimate them separately for each month. Also, we check longer intraday periods used for algorithmic trading, such as 5, 10, 30, and 60 minutes. The data are stationary for all time dimensions.

For robustness, we run models with the five most actively traded futures[8] and the rest 22 futures. The results stay the same as when we run regressions across all 27 futures. Further, we continue our analysis without splitting the dataset.

We run a panel regression with fixed effects[9], as shown below, where data are pooled across 27 future contracts traded on MOEX using data for twenty months from January 2020 to September 2021.

We conduct our empirical investigation by examining how investors' trading activity is related to future returns. In the first model, we measure investors' activity on the asset i as a volume of the buyer- and seller-initiated trades by the client group in the current time window. We construct the measure separately for 30- and 60-second, 5-, 10-, 30-, 60-minute, and daily time windows.

$$R_{i,t+1} = \beta_0 + \sum_{m=1}^{5} \beta_m Volume^{Z}_{m,i,t} + \epsilon_{i,t}, \tag{3.1}$$

where $Z \in \{CORP, DLR, RET, NER, INST\}$ is a client type, so $Volume^{CORP}_{i,t}$ is a traded volume for the future contract i by corporate trader in current time window; similarly, $Volume^{DLR}_{i,t}$, $Volume^{RET}_{i,t}$, $Volume^{NER}_{i,t}$, and $Volume^{INST}_{i,t}$ are the buyer or seller trading volumes for the future contract i by dealers, retail traders, non-residents, and institutional investors respectively in current time window.

In the second model, we use the order flow to measure investors' activity on the asset i as a buyer and seller trading volume difference by the client group in the current time window. We also construct the measure separately for 30- and 60-second, 5-, 10-, 30-, 60-minute, and daily time windows.

$$R_{i,t+1} = \beta_0 + \sum_{m=1}^{5} \beta_m OrderFlow^{Z}_{m,i,t} + \epsilon_{i,t}, \tag{3.2}$$

where $Z \in \{CORP, DLR, RET, NER, INST\}$ is a client type, so $OrderFlow^{CORP}_{i,t}$ is a difference in an order flow for the future contract i by corporate traders in current time window; similarly $OrderFlow^{DLR}_{i,t}$, $OrderFlow^{RET}_{i,t}$, $OrderFlow^{NER}_{i,t}$, and $OrderFlow^{INST}_{i,t}$ are the order flows for the future contract i by dealers, retail traders, non-residents, and institutional investors respectively in current time window.[10]

[8] 78% of all trading volume is concentrated in the top five futures contracts: Si, BR, GAZR, RTS, SBRF.

[9] We have experimented with fixed and random effects, and none have any significant differences. We report fixed effects results further in the chapter.

[10] We don't include undefined ($UNDEF$) client group in our models.

We estimate panel regressions for each month separately and for the entire period, which gives us 1,323 panel regression results[11]. We divide the trading volume by 10,000 for better results representation.

Table 3.2 shows the typical estimation results of the panel regression for one month, October 2020, using a 30-second time-frequency window for the order flow calculation.

	Dependent variable: Returns		
	Buy	**Sell**	**Order Flow**
CORP	-0.013***	0.011***	-0.028 ***
	(0.002)	(0.002)	(0.010)
DLR	-1.376***	0.536***	-1.003***
	(0.156)	(0.107)	(0.092)
RET	0.016***	-0.036***	0.030***
	(0.002)	(0.002)	(0.010)
INST	-0.016***	0.078***	-0.100***
	(0.005)	(0.005)	(0.011)
NER	-0.032**	0.022*	-0.033**
	(0.014)	(0.012)	(0.014)
Observations	87,605	87,605	87,605
R^2	0.002	0.007	0.012
Adjusted R^2	0.002	0.007	0.012
F Statistic (df = 5; 87595)	39.378***	121.960***	207.383***
Note:			*p<0.1; **p<0.05; ***p<0.01

TABLE 3.2: Example of monthly results. This table reports results for panel regression with fixed effects of customers trading volume pulled across five most traded future contracts for one months, October 2020, using 30-second time window. **Buy** is a buyer-initiated trading volume, **Sell** is a seller-initiated trading volume, **Order Flow** is the difference between buying and selling trading volumes.

All monthly regression results for retail traders (RET) are significant, with positive signs for buy trades and negative signs for sell trades. Over 70% results for corporate clients ($CORP$) and dealers (DLR) are significant, with negative signs for buy trades and positive signs for sell trades. So, when more retail traders enter the market to buy future contracts, the price returns go up in the next period, while we observe the opposite when corporate clients and dealers enter the market. We must consider the short-term movements as we use 10-, 30-, and 60-second time-frequency windows.

We also check if our intraday results vary using standardised volumes for buyer- and seller-initiated trades and order flows (Appendix 3.8.3). As mentioned above, for trading volume standardisation, we divide the client's trading volume by their standard deviation to remove the difference in absolute volume size across futures. As we analyse intraday, we compute the standard deviation of flows via a rolling scheme over a previous 5-day trading period. For robustness, we also pull regressions for the five most liquid futures and the rest separately. We do not observe significant differences. The results stay the same.

Our results show that the retail traders' buyer- and seller-initiated trading volumes and order flows have predictive power for the returns in the futures market

[11] 20 months, 7 time frequencies for each month give us 140 regressions, then we run for all futures and two subsets, giving us 420 regressions, and then we run for the entire period also 8 regression for tree subsets (7*3=21)—total 441 regressions. Then we run separately for buyer-initiated, seller-initiated volumes and order flows. Totally 441*3=1,323 regressions.

	30 sec		60 sec		5 min		10 min		30 min		60 min	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Panel A. Buyer-initiated volume												
CORP	-	100	-	95	-	70	-	70	-	70	-	35
DLR	-	60	-	75	-	60	-	60	-	45	-	40
RET	+	100	+	100	+	55	-	60	+	50	+	45
INST	-	60	-	30	+	70	+	75	+	60	+	45
NER	-	30	+	25	+	25	+	25	+	10	-	10
Panel B. Seller-initiated volume												
CORP	+	100	+	80	+	60	+	60	+	60	+	50
DLR	+	60	+	65	+	45	+	45	+	30	-	25
RET	-	100	-	90	-	75	-	90	-	80	-	85
INST	+	65	+	60	+	55	-	70	-	50	-	35
NER	+	35	-	25	-	20	-	15	-	15	+	25
Panel C. Order Flow												
CORP	-	50	-	40	-	30	-	30	-	25	-	15
DLR	-	80	-	70	-	65	-	65	-	60	-	60
RET	+	75	+	40	-	30	-	20	+	20	+	20
INST	-	85	-	75	-	40	-	40	-	25	-	10
NER	-	30	-	25	-	30	-	25	-	15	-	10

TABLE 3.3: Summary of the monthly panel regression results with fixed effects (Equation 3.1 and Equation 3.2) pulled across all stocks using 10-, 30-, 60-second and 5-, 10-, 30-, 60-minute time windows. (1) - sign of the coefficient, (2) - percentage of significant coefficients across all regressions.

for the next trading period in a high-frequency time dimension of 30 and 60 seconds. In longer time frequencies, the trading behaviour of retail clients loses its predictive power. Corporate clients tend to trade in the opposite direction to retail traders, and their behaviour does not show informativeness. The negative coefficient of the order flow in all time frequencies tells us that when corporate clients buy futures, the price decreases and vice versa. Retail traders and institutions show the informativeness of future price movements, while corporate clients trade in the opposite direction. Daily dealers' order flows' coefficients are insignificant, while intraday dealers' order flows have negative signs. The higher dealer's order flow, the lower the price returns.

So, we find a positive relationship between retail flows and future prices. Kaniel et al. (2012), Barrot et al. (2016) claim that retail investors are rewarded for providing liquidity to institutional investors. Our analysis reveals that in 30- and 60-second time frequencies, retail traders provide liquidity to institutional investors that need to execute their trades immediately, as suggested by Kaniel et al. (2008). The willingness of retail investors to provide liquidity and autocorrelated order flows (Section 3.3.2) may contribute to the forecasting power for the short-term price pressure (Barber et al. 2008).

Moreover, the retail selling volume anticipates negative returns in the next period (Table 3.4, Panel B.). We can not observe whether the seller-initiated trade was closing the long position or a short sell; our results are most consistent with the information hypothesis that retail short sellers possess and act on unique information beyond that held by other investors. Under this theory, retail short selling predicts negative returns as prices of the underlying assets converge to their fundamental values, just as informed order flow predicts returns in models such as Kyle (1985).

Even if each retail trader has wildly inaccurate information, the resulting signal is relatively precise when the information is aggregated through the trades of many

	Dependent variable: Returns						
	30 seconds	60 seconds	5 minutes	10 minutes	30 minutes	60 minutes	1 day
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Panel A. Buyer-initiated volume				
CORP	-0.039^{***}	-0.023^{***}	-0.017^{***}	-0.017^{***}	-0.014^{***}	-0.008^{***}	-0.028^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.010)
DLR	-0.040^{***}	-0.030^{***}	-0.032^{***}	-0.029^{***}	-0.021^{*}	-0.015	-0.021
	(0.005)	(0.005)	(0.008)	(0.009)	(0.011)	(0.013)	(0.069)
RET	0.026^{***}	0.015^{***}	0.009^{***}	0.008^{***}	0.005^{***}	0.004^{***}	0.013^{***}
	(0.0005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)
INST	-0.006^{***}	-0.001	0.020^{***}	0.028^{***}	0.029^{***}	0.013^{***}	0.057^{**}
	(0.026)	(0.002)	(0.002)	(0.003)	(0.003)	(0.004)	(0.005)
NER	0.003	0.001	-0.0002	-0.0004	-0.001	-0.001	-0.020^{*}
	(0.012)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
Observations	5,993,365	3,623,479	933,469	498,699	168,276	79,713	9,911
R^2	0.001	0.0003	0.0003	0.0004	0.0005	0.0003	0.001
Adjusted R^2	0.001	0.0003	0.0003	0.0004	0.0005	0.0002	0.001
F Statistic	$4{,}375.464^{***}$	$1{,}038.102^{***}$	263.512^{***}	209.330^{***}	81.354^{***}	20.005^{***}	10.172^{*}
			Panel B. Seller-initiated volume				
CORP	0.028^{***}	0.013^{***}	0.007^{***}	0.012^{***}	0.017^{***}	0.008^{***}	0.026^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.010)
DLR	0.080^{***}	0.054^{***}	0.016	0.012	0.009	0.002	-0.086
	(0.007)	(0.008)	(0.011)	(0.012)	(0.015)	(0.016)	(0.080)
RET	-0.026^{***}	-0.013^{***}	-0.005^{***}	-0.007^{***}	-0.011^{***}	-0.004^{***}	-0.012^{**}
	(0.0005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)
INST	0.023^{***}	0.016^{***}	0.0004	-0.007^{**}	-0.012^{***}	-0.005	-0.049^{*}
	(0.002)	(0.002)	(0.003)	(0.003)	(0.004)	(0.005)	(0.026)
NER	-0.016^{***}	-0.014^{***}	-0.005^{**}	-0.001	0.003	0.001	0.0001
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.013)
R^2	0.001	0.0002	0.0001	0.0002	0.001	0.0003	0.001
Adjusted R^2	0.001	0.0002	0.0001	0.0002	0.001	0.0002	0.001
F Statistic	$3{,}531.261^{***}$	733.900^{***}	65.190^{***}	116.234^{***}	163.758^{***}	20.681^{***}	10.704^{*}
			Panel C. Order Flow				
CORP	-0.033^{***}	-0.022^{***}	-0.031^{***}	-0.043^{***}	-0.043^{***}	-0.023^{*}	-0.052
	(0.004)	(0.005)	(0.007)	(0.009)	(0.011)	(0.014)	(0.105)
DLR	-0.070^{***}	-0.055^{***}	-0.057^{***}	-0.065^{***}	-0.062^{***}	-0.038^{*}	0.067
	(0.006)	(0.007)	(0.011)	(0.012)	(0.016)	(0.020)	(0.149)
RET	0.042^{***}	0.027^{***}	0.006	0.003	0.016	0.017	0.102
	(0.004)	(0.005)	(0.007)	(0.009)	(0.011)	(0.014)	(0.106)
INST	-0.070^{***}	-0.053^{***}	-0.041^{***}	-0.049^{***}	-0.053^{***}	-0.036^{***}	-0.071
	(0.004)	(0.005)	(0.008)	(0.009)	(0.011)	(0.014)	(0.106)
NER	0.068^{***}	0.057^{***}	0.040^{***}	0.039^{***}	0.055^{***}	0.050^{**}	0.281^{*}
	(0.005)	(0.006)	(0.009)	(0.011)	(0.016)	(0.021)	(0.170)
R^2	0.002	0.001	0.001	0.001	0.003	0.002	0.006
Adjusted R^2	0.002	0.001	0.001	0.001	0.003	0.002	0.006
F Statistic	$10{,}935.030^{***}$	$3{,}062.981^{***}$	611.306^{***}	623.706^{***}	504.348^{***}	139.725^{***}	60.274^{***}
Note:					$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		

TABLE 3.4: Intraday and daily results for entire period. Table shows panel regression results with fixed effects pulled across all traded future contracts for thebentire period (Equation 3.1 and Equation 3.2). Panels A, B, and C represent the buyer-initiated, seller-initiated trading volumes, and the order flow results, respectively.

individuals. We show that 46% of futures trading volume on MOEX is associated with retail traders. Also, retail traders have fewer constraints than institutional investors, at least with respect to diversification requirements or short-selling. Thus retail traders are better positioned to trade aggressively when they are informed.

3.4.1 Asset classes subsample analysis

We are interested in analysing if there are any differences in the impact of investors' order flows on future returns on asset class subsamples. We divide our dataset into four subsamples based on the underlying asset class. In Table 3.14, we mark each contract with its group number in column (3). The first group includes three futures contracts on currency exchange (SI, Eu, ED), the second group includes five futures on market indices (RTS, MIX, MXI, SPYF, RTSM), the third group includes five futures on commodities (BR, GOLD, SILV, NG, PLT), and the fourth group - fourteen futures on stocks. Then we test each group separately, running panel regressions with fixed effects (Equation 3.1 and Equation 3.2).

	Group 1		**Group 2**		**Group 3**		**Group 4**	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Panel A. Buyer-initiated volume								
CORP	+	83	-	100	-	100	-	100
DLR	-	100	-	67	-	83	-	83
RET	-	100	+	100	+	100	+	100
INST	-	100	+	100	+	50	+	50
NER	+	100	-	100	-	83	-	83
Panel B. Seller-initiated volume								
CORP	-	100	+	100	+	100	+	100
DLR	+	67	+	50	+	33	+	33
RET	+	100	-	83	-	100	-	100
INST	+	100	-	100	-	100	-	100
NER	-	83	+	33	+	50	+	50
Panel C. Order Flow								
CORP	+	100	-	100	-	67	-	67
DLR	-	83	-	100	-	83	-	83
RET	-	83	-	100	+	100	-	83
INST	-	67	-	100	+	83	+	83
NER	+	67	-	100	+	100	+	100

TABLE 3.5: Results across different asset classes. Table is a summary of monthly panel regression results (Equation 3.1 and Equation 3.2) with fixed effects pulled across all stocks for intraday time windows for each group of futures. (1) - sign of the coefficient, (2) - percentage of significant coefficients across all regressions. Group 1: currency exchange futures, Group 2: futures on market indices, Group 3: commodity future, Group 4: stock futures.

Analysing the futures of different asset classes, we reveal dissimilarities in the impact of different client groups' trading activity on market outcomes. Retail traders', institutional investors' and non-residents' order flows are good predictors for intraday returns in commodity futures. In contrast, only institutional inventors' order flow positively predicts returns in futures on stocks. In the high-frequency environment, order flows of corporate clients, retail traders, and non-residence positively predict returns on currency futures, while the order flow of institutional investors does not. Our findings show that non-residents obtain information about price movements of currency and commodity futures and correctly predict intraday returns. None of the customer groups anticipates the price movements of market

	Dependent variable: Returns						
	30 seconds	60 seconds	5 minutes	10 minutes	30 minutes	60 minutes	1 day
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Buyer-initiated volume							
CORP	0.00004	0.005***	0.011***	0.011***	0.010***	0.013***	0.0003
	(0.0003)	(0.0003)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.003)
DLR	−0.081***	−0.075***	−0.058***	−0.045***	−0.036***	−0.026**	−0.052*
	(0.003)	(0.003)	(0.006)	(0.007)	(0.009)	(0.010)	(0.032)
RET	0.012***	0.004***	−0.008***	−0.009***	−0.008***	−0.009***	0.003
	(0.0003)	(0.0004)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)
INST	−0.039***	−0.030***	−0.013***	−0.010***	−0.008***	−0.011***	−0.012
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.009)
NER	0.013***	0.019***	0.023***	0.020***	0.015***	0.015***	−0.004
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.003)	(0.009)
Observations	1,157,351	617,772	124,036	63,233	20,673	9,731	1,212
R^2	0.004	0.003	0.006	0.007	0.009	0.018	0.014
Adjusted R^2	0.004	0.003	0.006	0.007	0.008	0.018	0.010
F Statistic	4,115.694***	1,991.755***	717.435***	432.659***	180.843***	183.173***	16.522***
Panel B. Seller-initiated volume							
CORP	−0.001***	−0.006***	−0.009***	−0.009***	−0.009***	−0.008***	−0.001
	(0.0003)	(0.0003)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.003)
DLR	0.060***	0.050***	0.027***	0.025***	0.010	0.0005	−0.075**
	(0.003)	(0.003)	(0.006)	(0.007)	(0.008)	(0.010)	(0.032)
RET	−0.013***	−0.005***	0.003***	0.004***	0.005***	0.003***	0.002
	(0.0003)	(0.0004)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)
INST	0.045***	0.037***	0.021***	0.015***	0.016***	0.020***	−0.005
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.008)
NER	−0.010***	−0.014***	−0.014***	−0.014***	−0.009***	−0.003	0.002
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.003)	(0.009)
R^2	0.004	0.004	0.005	0.005	0.010	0.014	0.015
Adjusted R^2	0.004	0.004	0.005	0.005	0.010	0.014	0.011
F Statistic	5,019.140***	2,541.279***	669.044***	342.648***	206.231***	140.478***	18.071***
Panel C. Order flow							
CORP	0.015***	0.014***	0.007**	0.006*	0.012***	0.022***	0.023
	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)	(0.005)	(0.024)
DLR	−0.047***	−0.043***	−0.038***	−0.037***	−0.023***	−0.002	0.071
	(0.003)	(0.003)	(0.005)	(0.006)	(0.008)	(0.010)	(0.050)
RET	0.031***	0.020***	−0.003	−0.009***	−0.009**	−0.003	0.024
	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)	(0.005)	(0.025)
INST	−0.029***	−0.024***	−0.013***	−0.008**	0.0004	0.005	0.015
	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)	(0.005)	(0.024)
NER	0.037***	0.036***	0.026***	0.018***	0.007	0.001	0.006
	(0.002)	(0.002)	(0.004)	(0.004)	(0.006)	(0.008)	(0.046)
R^2	0.009	0.007	0.008	0.009	0.016	0.029	0.002
Adjusted R^2	0.009	0.007	0.008	0.009	0.016	0.029	−0.002
F Statistic	10,040.870***	4,506.858***	1,039.535***	604.315***	340.600***	293.800***	2.735
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE 3.6: Intraday and daily results for Group 1 (currency exchange rate futures) of Equation 3.1 and Equation 3.2 for daily and intraday returns. Table shows panel regression results with fixed effects pulled across future contracts in Group 1 for the entire period. Panels A, B, and C represent the buyer-initiated, seller-initiated trading volumes, and the order flow results, respectively.

	Dependent variable: Returns						
	30 seconds	60 seconds	5 minutes	10 minutes	30 minutes	60 minutes	1 day
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Buyer-initiated volume							
CORP	-0.164^{***}	-0.042^{***}	-0.040^{***}	-0.112^{***}	-0.172^{***}	-0.171^{***}	-0.047
	(0.004)	(0.005)	(0.008)	(0.009)	(0.012)	(0.014)	(0.066)
DLR	-0.619^{***}	-0.505^{***}	-0.321^{***}	-0.287^{**}	-0.215	-0.289^{*}	-0.449
	(0.060)	(0.072)	(0.103)	(0.117)	(0.136)	(0.157)	(0.556)
RET	0.166^{***}	0.053^{***}	0.016^{**}	0.057^{***}	0.102^{***}	0.106^{***}	0.013
	(0.004)	(0.004)	(0.006)	(0.007)	(0.009)	(0.011)	(0.053)
INST	-0.110^{***}	-0.051^{***}	0.251^{***}	0.405^{***}	0.379^{***}	0.354^{***}	0.467
	(0.015)	(0.018)	(0.029)	(0.034)	(0.045)	(0.058)	(0.346)
NER	-0.112^{***}	-0.065^{***}	-0.039^{***}	-0.063^{***}	-0.088^{***}	-0.083^{***}	-0.016
	(0.010)	(0.011)	(0.013)	(0.014)	(0.016)	(0.018)	(0.066)
Observations	1,148,178	650,584	138,430	70,769	23,148	10,896	1,357
R^2	0.002	0.0003	0.001	0.003	0.010	0.014	0.003
Adjusted R^2	0.002	0.0003	0.001	0.003	0.010	0.014	-0.001
F Statistic	$2{,}241.885^{***}$	219.540^{***}	112.738^{***}	242.393^{***}	238.013^{***}	156.856^{***}	4.228
Panel B. Seller-initiated volume							
CORP	0.193^{***}	0.079^{***}	0.016^{*}	0.044^{***}	0.093^{***}	0.101^{***}	0.059
	(0.004)	(0.005)	(0.008)	(0.009)	(0.012)	(0.014)	(0.067)
DLR	0.623^{***}	0.458^{***}	0.217^{**}	0.174	0.122	0.011	-0.999
	(0.060)	(0.071)	(0.108)	(0.120)	(0.154)	(0.185)	(0.683)
RET	-0.183^{***}	-0.073^{***}	0.017^{**}	0.002	-0.036^{***}	-0.037^{***}	-0.020
	(0.004)	(0.004)	(0.007)	(0.008)	(0.009)	(0.011)	(0.056)
INST	0.215^{***}	0.113^{***}	-0.282^{***}	-0.393^{***}	-0.473^{***}	-0.537^{***}	-0.176
	(0.015)	(0.018)	(0.028)	(0.034)	(0.045)	(0.058)	(0.333)
NER	0.079^{***}	0.042^{***}	0.003	0.006	0.029^{*}	0.027	0.037
	(0.010)	(0.011)	(0.013)	(0.014)	(0.016)	(0.018)	(0.071)
R^2	0.003	0.001	0.001	0.002	0.006	0.010	0.003
Adjusted R^2	0.003	0.001	0.001	0.002	0.006	0.010	-0.0003
F Statistic	$2{,}896.087^{***}$	373.020^{***}	108.233^{***}	143.443^{***}	148.903^{***}	113.553^{***}	4.544
Panel C. Order flow							
CORP	-0.490^{***}	-0.359^{***}	-0.577^{***}	-0.841^{***}	-1.292^{***}	-1.455^{***}	-1.503
	(0.028)	(0.036)	(0.067)	(0.086)	(0.130)	(0.183)	(1.194)
DLR	-0.969^{***}	-0.971^{***}	-1.167^{***}	-1.155^{***}	-1.108^{***}	-1.020^{***}	0.168
	(0.053)	(0.065)	(0.109)	(0.131)	(0.184)	(0.247)	(1.458)
RET	-0.071^{***}	-0.209^{***}	-0.599^{***}	-0.689^{***}	-0.731^{***}	-0.717^{***}	-1.075
	(0.027)	(0.036)	(0.067)	(0.086)	(0.130)	(0.182)	(1.185)
INST	-0.426^{***}	-0.368^{***}	-0.218^{***}	-0.225^{**}	-0.487^{***}	-0.557^{***}	-0.337
	(0.029)	(0.038)	(0.070)	(0.090)	(0.136)	(0.191)	(1.232)
NER	-0.264^{***}	-0.482^{***}	-1.092^{***}	-1.247^{***}	-1.375^{***}	-1.233^{***}	-1.392
	(0.031)	(0.040)	(0.072)	(0.092)	(0.138)	(0.191)	(1.227)
R^2	0.005	0.001	0.004	0.010	0.035	0.057	0.009
Adjusted R^2	0.005	0.001	0.004	0.010	0.034	0.056	0.005
F Statistic	$5{,}969.441^{***}$	834.985^{***}	596.211^{***}	696.351^{***}	828.613^{***}	657.022^{***}	11.976^{**}
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE 3.7: Intraday and daily results for Group 2 (market index futures) of Equation 3.1 and Equation 3.2 for daily and intraday returns. Table shows panel regression results with fixed effects pulled across future contracts in Group 2 for the entire period. Panels A, B, and C represent the buyer-initiated, seller-initiated trading volumes, and the order flow results, respectively.

	Dependent variable: Returns						
	30 seconds	60 seconds	5 minutes	10 minutes	30 minutes	60 minutes	1 day
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Buyer-initiated volume							
CORP	-0.116^{***}	-0.084^{***}	-0.083^{***}	-0.077^{***}	-0.059^{***}	-0.046^{***}	-0.075^{***}
	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)	(0.005)	(0.027)
DLR	-0.023^{***}	-0.017^{**}	-0.066^{***}	-0.093^{***}	-0.065^{***}	-0.036	-0.136
	(0.007)	(0.008)	(0.012)	(0.014)	(0.020)	(0.023)	(0.193)
RET	0.044^{***}	0.032^{***}	0.031^{***}	0.028^{***}	0.021^{***}	0.019^{***}	0.023^{*}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.013)
INST	0.002	0.001	0.073^{***}	0.126^{***}	0.078^{***}	0.019	0.440^{***}
	(0.007)	(0.008)	(0.013)	(0.015)	(0.019)	(0.023)	(0.165)
NER	-0.004	-0.009^{***}	-0.010^{***}	-0.008^{**}	-0.009^{**}	-0.008^{*}	-0.020
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.020)
Observations	1,334,841	769,737	191,776	102,093	34,102	16,087	2,007
R^2	0.005	0.003	0.006	0.007	0.007	0.006	0.007
Adjusted R^2	0.005	0.003	0.005	0.007	0.007	0.006	0.004
F Statistic	$6{,}921.737^{***}$	$2{,}387.757^{***}$	$1{,}061.144^{***}$	726.953^{***}	257.528^{***}	103.920^{***}	13.449^{**}
Panel B. Seller-initiated volume							
CORP	0.079^{***}	0.051^{***}	0.047^{***}	0.058^{***}	0.068^{***}	0.037^{***}	0.064^{**}
	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)	(0.004)	(0.028)
DLR	0.053^{***}	0.036^{**}	0.011	0.014	0.017	-0.0003	-0.457
	(0.014)	(0.016)	(0.024)	(0.029)	(0.038)	(0.041)	(0.504)
RET	-0.036^{***}	-0.023^{***}	-0.018^{***}	-0.023^{***}	-0.028^{***}	-0.013^{***}	-0.035^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.012)
INST	-0.027^{***}	-0.029^{***}	-0.083^{***}	-0.060^{***}	-0.113^{***}	-0.099^{***}	0.199
	(0.006)	(0.008)	(0.011)	(0.013)	(0.016)	(0.020)	(0.163)
NER	-0.016^{***}	-0.011^{***}	-0.002	0.003	0.007^{*}	0.002	0.012
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.022)
R^2	0.003	0.001	0.002	0.004	0.012	0.005	0.006
Adjusted R^2	0.003	0.001	0.002	0.004	0.012	0.005	0.004
F Statistic	$3{,}875.577^{***}$	$1{,}088.840^{***}$	375.696^{***}	441.705^{***}	407.201^{***}	88.515^{***}	12.189^{**}
Panel C. Order flow							
CORP	-0.079^{***}	-0.051^{***}	-0.040^{*}	-0.049^{*}	-0.038	0.020	-0.064
	(0.013)	(0.015)	(0.022)	(0.027)	(0.035)	(0.045)	(0.563)
DLR	-0.122^{***}	-0.087^{***}	-0.090^{***}	-0.103^{***}	-0.092^{**}	-0.021	0.049
	(0.015)	(0.018)	(0.025)	(0.030)	(0.040)	(0.051)	(0.596)
RET	0.044^{***}	0.038^{**}	0.051^{**}	0.052^{**}	0.064^{*}	0.088^{**}	0.131
	(0.013)	(0.015)	(0.022)	(0.027)	(0.035)	(0.045)	(0.013)
INST	-0.040^{***}	-0.024	0.097^{***}	0.085^{***}	0.099^{**}	0.116^{**}	-0.401
	(0.014)	(0.017)	(0.026)	(0.031)	(0.042)	(0.054)	(0.642)
NER	0.136^{***}	0.120^{***}	0.159^{***}	0.181^{***}	0.225^{***}	0.241^{***}	0.608
	(0.014)	(0.018)	(0.027)	(0.034)	(0.049)	(0.065)	(0.727)
R^2	0.007	0.004	0.008	0.013	0.022	0.015	0.014
Adjusted R^2	0.007	0.004	0.008	0.013	0.022	0.015	0.012
F Statistic	$9{,}851.156^{***}$	$3{,}297.797^{***}$	$1{,}460.179^{***}$	$1{,}310.128^{***}$	781.038^{***}	243.904^{***}	29.034^{***}
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE 3.8: Intraday and daily results for Group 3 (commodity futures) of Equation 3.1 and Equation 3.2 for daily and intraday returns. Table shows panel regression results with fixed effects pulled across future contracts in Group 3 for the entire period. Panels A, B, and C represent the buyer-initiated, seller-initiated trading volumes, and the order flow results, respectively.

	Dependent variable: Returns						
	30 seconds	60 seconds	5 minutes	10 minutes	30 minutes	60 minutes	1 day
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Buyer-initiated volume							
CORP	−0.110***	−0.080***	0.023***	0.022**	0.015	−0.008	0.054
	(0.004)	(0.005)	(0.008)	(0.009)	(0.012)	(0.015)	(0.069)
DLR	−0.023	−0.041	0.107*	0.294***	0.139	0.089	−1.033
	(0.036)	(0.043)	(0.065)	(0.077)	(0.108)	(0.138)	(0.668)
RET	0.097***	0.050***	−0.091***	−0.097***	−0.072***	−0.029*	−0.070
	(0.005)	(0.006)	(0.009)	(0.011)	(0.014)	(0.017)	(0.074)
INST	0.051***	0.145***	0.364***	0.390***	0.369***	0.281***	0.232
	(0.011)	(0.013)	(0.019)	(0.022)	(0.029)	(0.036)	(0.167)
NER	−0.029*	−0.005	0.083***	0.072***	0.086***	0.043	0.112
	(0.015)	(0.017)	(0.022)	(0.024)	(0.030)	(0.035)	(0.155)
Observations	2,352,995	1,585,386	479,227	262,604	90,353	42,999	5,371
R^2	0.0003	0.0002	0.001	0.001	0.002	0.002	0.002
Adjusted R^2	0.0003	0.0002	0.001	0.001	0.002	0.002	0.001
F Statistic	791.882***	376.069***	387.705***	344.252***	180.026***	72.926***	9.137
Panel B. Seller-initiated volume							
CORP	0.110***	0.059***	−0.075***	−0.073***	−0.044***	−0.039**	0.192**
	(0.004)	(0.005)	(0.008)	(0.009)	(0.012)	(0.015)	(0.080)
DLR	0.382***	0.200***	−0.091	−0.162**	−0.192**	−0.152	0.205
	(0.035)	(0.043)	(0.062)	(0.071)	(0.095)	(0.109)	(0.661)
RET	−0.114***	−0.011*	0.171***	0.163***	0.109***	0.088***	−0.161*
	(0.005)	(0.006)	(0.009)	(0.011)	(0.014)	(0.018)	(0.083)
INST	−0.047***	−0.178***	−0.266***	−0.271***	−0.167***	−0.123***	−0.079
	(0.011)	(0.013)	(0.019)	(0.022)	(0.028)	(0.035)	(0.157)
NER	−0.034**	−0.042**	−0.109***	−0.083***	−0.042	−0.041	0.027
	(0.015)	(0.018)	(0.023)	(0.025)	(0.031)	(0.037)	(0.173)
R^2	0.0004	0.0002	0.001	0.	0.001	0.001	0.002
0.002							
Adjusted R^2	0.0004	0.0002	0.001	0.001	0.001	0.001	0.001
F Statistic	1,042.588***	327.165***	409.468***	302.672***	96.220***	45.594***	11.329**
Panel C. Order flow							
CORP	−0.041	0.011	−0.013	−0.051	−0.040	−0.015	−0.590
	(0.045)	(0.053)	(0.077)	(0.088)	(0.111)	(0.134)	(0.986)
DLR	−0.063	−0.025	−0.142	−0.078	−0.076	−0.057	−0.974
	(0.051)	(0.061)	(0.087)	(0.099)	(0.124)	(0.146)	(1.117)
RET	0.141***	0.083	−0.334***	−0.463***	−0.479***	−0.415***	−0.387
	(0.045)	(0.053)	(0.077)	(0.087)	(0.109)	(0.130)	(0.974)
INST	0.062	0.215***	0.402***	0.438***	0.449***	0.414***	−0.394
	(0.046)	(0.054)	(0.079)	(0.091)	(0.115)	(0.140)	(1.016)
NER	0.196***	0.137**	−0.309***	−0.453***	−0.403***	−0.351**	0.025
	(0.047)	(0.056)	(0.083)	(0.096)	(0.121)	(0.148)	(1.018)
R^2	0.001	0.0004	0.002	0.004	0.006	0.005	0.001
Adjusted R^2	0.001	0.0004	0.002	0.004	0.006	0.005	0.0002
F Statistic	1,516.070***	579.164***	1,193.973***	1,173.228***	535.817***	236.016***	6.289
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE 3.9: Intraday and daily results for Group 4 (stock futures) of Equation 3.1 and Equation 3.2 for daily and intraday returns. Table shows panel regression results with fixed effects pulled across future contracts in Group 4 for the entire period. Panels A, B, and C represent the buyer-initiated, seller-initiated trading volumes, and the order flow results, respectively.

index futures (Table 3.7), which could be a question for further investigation by including a market-specific set of control variables.

3.4.2 Contemporaneous analysis

We find a genuine predictive power in customers' order flows. The following logical step is to question whether the returns at time t are reflected by the order flows at time t, so the customers' order flows explain the current price movements. This allows us to investigate the entire picture of the customer's trading activity and its effect on future returns. To do so, we run explanatory regressions with fixed effects as shown below in Equation 3.3. We construct the measure separately for 30- and 60-second, 5-, 10-, 30-, 60-minute and daily time frequencies.

$$R_{i,t} = \beta_0 + \sum_{m=1}^{5} \beta_m OrderFlow^{Z}_{m,i,t} + \epsilon_{i,t}, \tag{3.3}$$

where $Z \in \{CORP, DLR, RET, NER, INST\}$ is a client type, so $OrderFlow^{CORP}_{i,t}$ is a difference in an order flow for the future contract i by the corporate trader in current time window; similarly $OrderFlow^{DLR}_{i,t}$, $OrderFlow^{RET}_{i,t}$, $OrderFlow^{NER}_{i,t}$, and $OrderFlow^{INST}_{i,t}$ are the order flows for the future contract i by dealers, retail traders, non-residents, and institutional investors respectively in the current time window.

	Dependent variable: Returns						
	30 seconds	60 seconds	5 minutes	10 minutes	30 minutes	60 minutes	1 day
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Order Flow				
CORP	-0.018^{***}	-0.024^{***}	-0.045^{***}	-0.033^{***}	-0.033^{***}	-0.013^{*}	0.001
	(0.004)	(0.005)	(0.007)	(0.009)	(0.011)	(0.014)	(0.106)
DLR	0.011	-0.009	-0.055^{***}	-0.039^{***}	-0.056^{***}	-0.064^{*}	0.417
	(0.006)	(0.007)	(0.011)	(0.012)	(0.016)	(0.020)	(0.150)
RET	0.022^{***}	0.031^{***}	0.014^{*}	0.001	-0.012	-0.001	0.035
	(0.004)	(0.005)	(0.007)	(0.009)	(0.011)	(0.014)	(0.106)
INST	0.080^{***}	0.050^{***}	0.026^{***}	0.037^{*}	-0.004	-0.046^{***}	0.011
	(0.004)	(0.005)	(0.008)	(0.009)	(0.011)	(0.014)	(0.106)
NER	0.010^{**}	0.020^{***}	0.006	0.0001	0.028^{*}	-0.019	0.173^{*}
	(0.005)	(0.006)	(0.009)	(0.011)	(0.016)	(0.021)	(0.171)
Observations	5,411,783	3,356,239	901,928	492,186	167,472	79,405	9,911
R^2	0.001	0.003	0.001	0.001	0.003	0.003	0.002
Adjusted R^2	0.001	0.003	0.001	0.001	0.003	0.003	0.001
F Statistic	$3{,}066.183^{***}$	849.619^{***}	399.078^{***}	309.276^{***}	47.719^{***}	23.858^{***}	18.143^{***}
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE 3.10: Intraday and daily contemporaneous regression results. Table shows contemporaneous panel regression results with fixed effects pulled across all traded future contracts for the entire period (Equation 3.3). Table represents the order flow results.

Analysing high-frequency intraday time dimensions from 30 seconds to 10 minutes, we observe that the prices go up in the current period when institutions' order flow rises. So, institutions move the price up in the current time window. However, results in Table 3.3 shows a price decrease in the next period. In contrast, retail traders react to the price momentum in the current period and continue to trade in the same direction in the next period. We may interpret it as the market overreacting to the institutional trades in the current short-term period.

Order flows of corporate clients and dealers have negative signs of the coefficients in explanatory and forecasting regressions, which align with correlations analysis in Section 3.3. Non-residents anticipate price movements the next day but also move prices in the current trading day.

Also, we run contemporaneous panel regressions on each of the four asset class groups separately, similar to Section 3.4.1. Results in Table 3.4.2 show that order flows of corporate clients, dealers, retail traders and non-residents tend to have the same effect on market outcomes for both current and future periods. However, while trading currency exchange rate futures (group 1) in a high-frequency dimension of up to 60 seconds, the market seems to overreact to institutional trades, and prices go up in the current period while dropping in the next period. Retail traders' activity moves the price of currency exchange futures (group 1) in short-term time windows of up to 10 minutes. However, the effect diminishes later, and no relationship is found for the daily returns.

In intraday commodity futures (group 3) trading, non-residents, institutional investors, and retail traders' activity moves the price up, while dealers and corporate clients trade in the opposite direction. In trading, corporate clients and dealers behave differently while trading stock futures (group 4). Institutional investors are the only group that moves the prices in the current time window in all intraday time frequencies.

Overall, contemporaneous analysis reveals that corporate clients and dealers trade in the opposite direction to the market. In a high-frequency time analysis, non-residents, retail traders and institutional investors tend to anticipate price movements. Alternatively, we may interpret these findings as non-residents, retail traders, and institutional investors pushing the prices up in the current period, especially knowing that the market share of these three groups is almost 90% of all market trading volume. We observe a more detailed picture while analysing futures on different asset classes. With institutional investors' order flow, we observe the impact on the current price movements. However, the market overreacts to large institutional trades, and thus prices decrease in the next time window. All the above discussion is valid for the intraday breakdown, while results for daily time windows do not show any effect, except a negative sign for corporate clients in the group 2 asset class analysis.

	Dependent variable: Returns						
	30 seconds (1)	60 seconds (2)	5 minutes (3)	10 minutes (4)	30 minutes (5)	60 minutes (6)	1 day (7)
Group 1. Order Flow							
CORP	0.011***	0.010***	0.009***	0.009***	0.014***	0.017***	0.045
	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)	(0.005)	(0.033)
DLR	−0.043***	−0.033***	−0.027***	−0.027***	−0.023***	−0.019***	0.095
	(0.003)	(0.003)	(0.005)	(0.006)	(0.008)	(0.010)	(0.068)
RET	−0.014***	−0.026***	−0.019***	−0.006*	0.003	0.006	0.029
	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)	(0.005)	(0.034)
INST	0.009***	0.014***	0.008***	−0.004	−0.007	−0.008	−0.002
	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)	(0.005)	(0.033)
NER	0.006***	−0.003	−0.010***	−0.009**	−0.024***	−0.037***	−0.102
	(0.002)	(0.002)	(0.004)	(0.004)	(0.006)	(0.008)	(0.062)
Observations	1,117,750	611,446	122,811	63,224	20,669	9,730	1,208
R^2	0.007	0.009	0.005	0.007	0.016	0.027	0.021
Adjusted R^2	0.007	0.009	0.005	0.007	0.015	0.027	0.017
F Statistic	7,900.977***	5,694.822***	574.920***	449.129***	329.097***	273.931***	26.181***
Group 2. Order Flow							
CORP	0.203***	0.049	−0.815***	−0.989***	−0.824***	−0.789***	−2.312*
	(0.028)	(0.037)	(0.067)	(0.086)	(0.132)	(0.188)	(1.277)
DLR	−0.620***	−0.690***	−0.869***	−0.744***	−0.499***	−0.493*	0.457
	(0.055)	(0.066)	(0.109)	(0.131)	(0.186)	(0.253)	(1.573)
RET	−0.352***	−0.421***	−0.586***	−0.541***	−0.444***	−0.410**	−0.951
	(0.028)	(0.036)	(0.067)	(0.086)	(0.131)	(0.187)	(1.268)
INST	−0.032	−0.090**	−0.268***	−0.472***	−0.531***	−0.588***	−1.558
	(0.030)	(0.039)	(0.071)	(0.090)	(0.138)	(0.195)	(1.319)
NER	−0.657***	−0.804***	−1.018***	−0.929***	−0.520***	−0.481**	−0.187
	(0.031)	(0.040)	(0.073)	(0.092)	(0.139)	(0.196)	(1.304)
Observations	1,074,400	631,446	136,871	70,747	23,145	10,896	1,352
R^2	0.009	0.008	0.006	0.012	0.010	0.011	0.041
Adjusted R^2	0.009	0.008	0.006	0.012	0.010	0.010	0.038
F Statistic	9,664.535***	4,847.369***	844.463***	847.782***	230.194***	119.797***	57.946***
Group 3. Order Flow							
CORP	−0.024*	−0.044***	−0.069***	−0.057**	0.034	0.045	0.276
	(0.012)	(0.015)	(0.022)	(0.027)	(0.036)	(0.045)	(0.565)
DLR	−0.039***	−0.073***	−0.113***	−0.086***	−0.076***	−0.068***	0.373
	(0.014)	(0.017)	(0.025)	(0.030)	(0.040)	(0.051)	(0.598)
RET	0.033***	0.028***	0.042***	0.063***	0.066***	0.079***	0.109
	(0.012)	(0.015)	(0.022)	(0.027)	(0.035)	(0.045)	(0.563)
INST	0.021***	0.063***	0.084***	0.071**	0.180***	0.081***	0.365
	(0.014)	(0.017)	(0.026)	(0.032)	(0.043)	(0.054)	(0.644)
NER	0.007	0.041**	0.086***	0.067**	0.079**	0.087**	0.261
	(0.014)	(0.018)	(0.027)	(0.034)	(0.049)	(0.066)	(0.741)
Observations	1,241,340	723,302	185,771	101,176	34,060	16,084	2,001
R^2	0.006	0.005	0.011	0.011	0.014	0.010	0.010
Adjusted R^2	0.006	0.005	0.011	0.015	0.014	0.010	0.010
F Statistic	700.608***	84.291***	1,123.794***	605.409***	59.766***	36.766***	21.037***
Group 4. Order Flow							
CORP	0.181***	0.101*	−0.281***	−0.336***	−0.024	−0.065	−1.175
	(0.046)	(0.054)	(0.079)	(0.089)	(0.111)	(0.134)	(1.734)
DLR	−0.038	−0.166***	−0.325***	−0.353***	−0.147	−0.213	0.115
	(0.053)	(0.062)	(0.088)	(0.100)	(0.124)	(0.146)	(1.836)
RET	−0.073	−0.265***	−0.528***	−0.495***	−0.209*	−0.198*	−1.958
	(0.046)	(0.054)	(0.078)	(0.088)	(0.109)	(0.130)	(1.736)
INST	0.301***	0.300***	0.225***	0.341***	0.320***	0.352***	−0.160
	(0.047)	(0.055)	(0.080)	(0.091)	(0.115)	(0.140)	(1.743)
NER	−0.114**	−0.368***	−0.680***	−0.593***	−0.378***	−0.305**	−1.392
	(0.048)	(0.057)	(0.084)	(0.096)	(0.121)	(0.148)	(1.710)
Observations	1,978,343	1,390,172	456,746	257,298	89,832	42,899	5,348
R^2	0.001	0.002	0.001	0.001	0.001	0.001	0.007
Adjusted R^2	0.001	0.002	0.001	0.001	0.001	0.001	0.006
F Statistic	2,441.944***	3,261.386***	626.538***	169.127***	89.178***	27.080***	36.202***
Note:							*p<0.1; **p<0.05; ***p<0.01

TABLE 3.11: Contemporaneous regression results for different asset classes. Table shows panel regression results with fixed effects pulled across future contracts on different asset classes for the entire period (Equation 3.3). Table represents the order flow results disaggregated by asset class groups; Group 1: future contracts on currency exchange; Group 2: futures on market indices; Group 3: commodity futures; Group 4: stock future.

3.5 Portfolio analysis

3.5.1 Portfolio formation

In this section, we further investigate how trading activity and order flows predict future returns and rely on a portfolio approach. We seek to prove or question our previous findings using an alternative yet not contradictory methodology.

Before sorting futures into portfolios, we need to ensure that order flows are comparable across different future contracts. As the absolute size of trading volume differs across futures on MOEX, it is not sensible to form portfolios based on raw order flows. To allow meaningful cross-asset comparison, we standardise order flows as shown above in Section 3.3.1.

We imitate the returns to investor trading by conditioning on lagged standardised order flows. We first sort futures into portfolios on lagged total order flows for each futures contract. We sort futures into seven portfolios (P_1, P_2, ..., P_7) depending on their total order flows on the period t and compute portfolio returns for the following period. Portfolio P_1 includes futures with the highest order flow, while portfolio P_7 - with the lowest. Then, we construct a long-short portfolio that goes long the top portfolio P_1 with the highest order flow and short the bottom portfolio P_7 with the lowest order flow. Our further analysis in this section reports results of investing in long-short portfolio (long P_1 portfolio and short P_7 portfolio).[12]

The investing period into a long-short portfolio could be explained using the following example. While analysing daily returns, we observe information on the trading volumes of different clients on the 20th of January 2020. For each client group, we sort the futures contracts by their standardised order flows and then put four contracts with the highest order flows in the first portfolio P_1 and three contracts with the lowest order flows in the last portfolio P_7. Thus, we form a long-short portfolio via buying portfolio P_1 and shorting the portfolio P_7. The first day of holding the portfolio is the 21st of January 2020. We follow the same logic when sorting futures inside the day. For example, the portfolio formation period is a trading hour from 10:00 to 11:00, and the first holding period is the following trading hour from 11:00 to 12:00. We stop at the end of the trading day and do not hold portfolios overnight as a common strategy for intraday trading.

3.5.2 Post-formation portfolio returns

We want to investigate whether the order flow forecasts returns because it signals temporary short-term movements or permanent shifts in the future market. This analysis is interested in daily time windows to show the existence of such shifts for the investment purposes. Order flow signals inside the day based on 60-second and 10-minute time windows give us information about intraday patterns' informativeness of short-term traders. So, we examine the results separately for daily and intraday forecasts.

A permanent price impact of daily signals would indicate that the order flow is related to changes in expectations about fundamentals. We form our portfolios as mentioned above and track cumulative future returns of the position of long portfolio P_1 with the highest buying power and short portfolio P_7. Our long-short portfolio has overlapping periods of 30 trading days after portfolio formation. This approach

[12] We made a sensitivity analysis with the different portfolio number choices, forming three and five portfolios from 27 futures. We do not observe meaningful differences in the final results. We show cumulative post-portfolio formation returns for three, five and seven numbers of portfolio in Appendix 3.8.8.

estimates how future prices move after experiencing intense buying or selling pressure from different client groups. Figure 3.12 illustrates the persistence of the predictive content of the order flow. The solid lines show the cumulative annual returns. The plots start with the first period after the portfolio formation.

Figure 3.10 and 3.11 illustrate 60-second and 10-minute portfolio returns, giving us the intraday predictive content of the order flow. Flows of retail traders and institutional investors forecast the 30-minute change in future prices. For the 5 hours (thirty 10-minute periods) in the day, retail traders' order flows do not show consistency in predictive power; however, order flows of institutional investors and non-residents predict positive future returns. Figure 3.12 shows that the order flow of institutional investors predicts only a temporary shift in future prices and loses its predictive power after four trading days. However, non-residents also do not show informativeness over a longer horizon. Institutional investors have a short-term impact and lose their power after four trading days; corporate clients and non-residents do not consistently impact prices, while dealers have a negative impact. Figure in Appendix 3.8.8 represents cumulative returns after portfolio formation on the same scale grouped by time windows.

Relying on the portfolio approach, we find the impact of institutional investors on futures intraday return, knowing that this client group has a relatively significant mean volume per trade (Figure 3.3). It is interesting to find that order flows by retail traders are indeed associated with futures price changes, while neither dealers nor corporate clients process the fundamental information. Retail traders have price informativeness in very short-term intraday trading; alternatively, the results may be the momentum trading patterns in their behaviour. Non-resident clients show mixed results, which is explained by the nature of this group identification, e.g., non-residents could be retail, corporate, or institutional traders.

3.5.3 Analysis in different market volatility environments

We further divide our sample into three sub-periods: periods with low, high, and ultra-high market volatilities, with cutoffs at the 50^{th} and 90^{th} percentiles of the time series distribution. Our main proxy for market volatility is the RVI index (the New Russian Volatility Index), which measures the market's expectation of the 30-day volatility, calculated from the real price of near- and next-series RTS index options. Results are shown in Appendix 3.8.4.

For 60-second time frequencies, the results for all customer groups and all volatility periods are similar. The only difference is an ultra-high market volatility environment when dealers' order flow shows opposite results. For 10-minute time frequencies, we observe the same picture when only in ultra-high market volatility environment results are mixed or opposite. Retail investors' order flow negatively predicts future price returns, and dealers and corporate traders positively predict returns for the first 100 minutes. Daily order flows show mixed results in different volatility environments. Daily retail traders' order flow positively predicts futures prices in ultra-high market volatility while showing negative predictions during high market volatility. Daily dealers' order flows also show mixed results with negative stable predictions during high market volatility. Other customer groups' results are mixed.

In sharp contrast to our earlier finding, short-term intraday order flow prediction in ultra-high market volatility environments are mixed or even opposite to the results of less volatile periods.

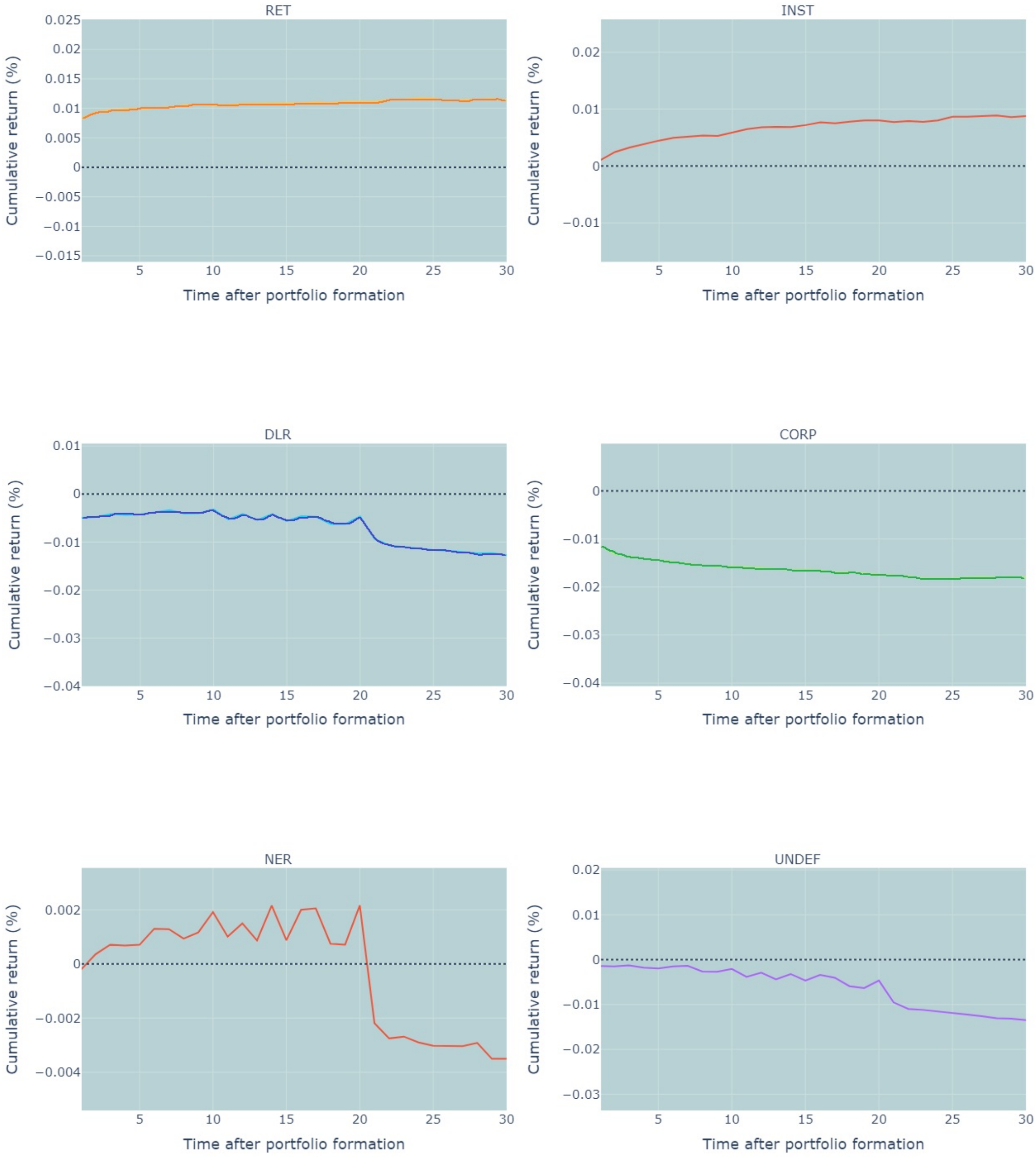


FIGURE 3.10: 60-second cumulative post-formation portfolio returns. This figure shows average cumulative returns for the position of long P_1 portfolio and short P_7 portfolio based on disaggregated order flows over the first 30 time windows or 30 minutes after the portfolio formation inside the day; periods overlap.

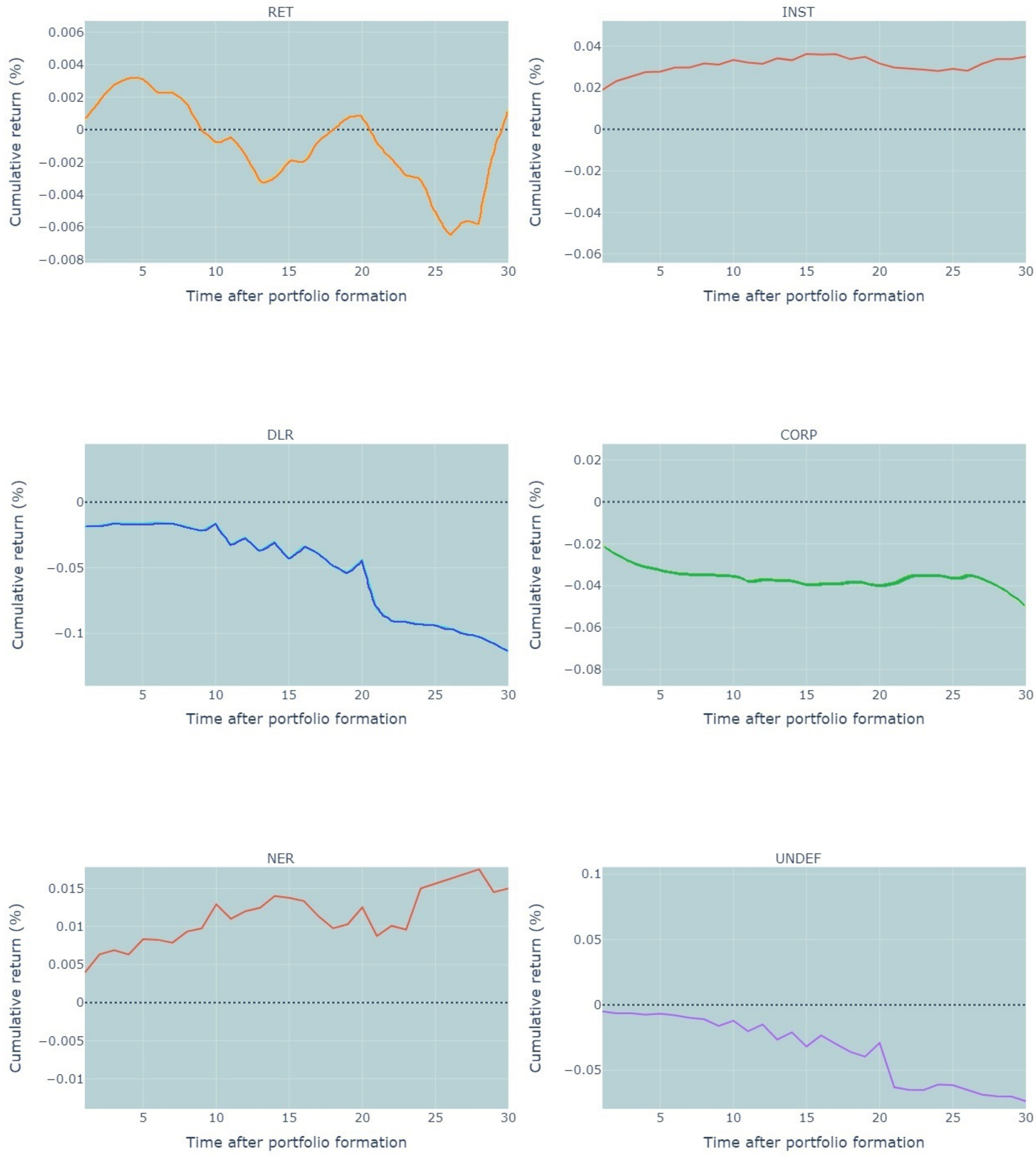


FIGURE 3.11: 10-minute cumulative post-formation portfolio returns. This figure shows average cumulative returns for the position of long P_1 portfolio and short P_7 portfolio based on disaggregated order flows over the first 30 time windows or 5 hours after the portfolio formation inside the day; periods overlap.

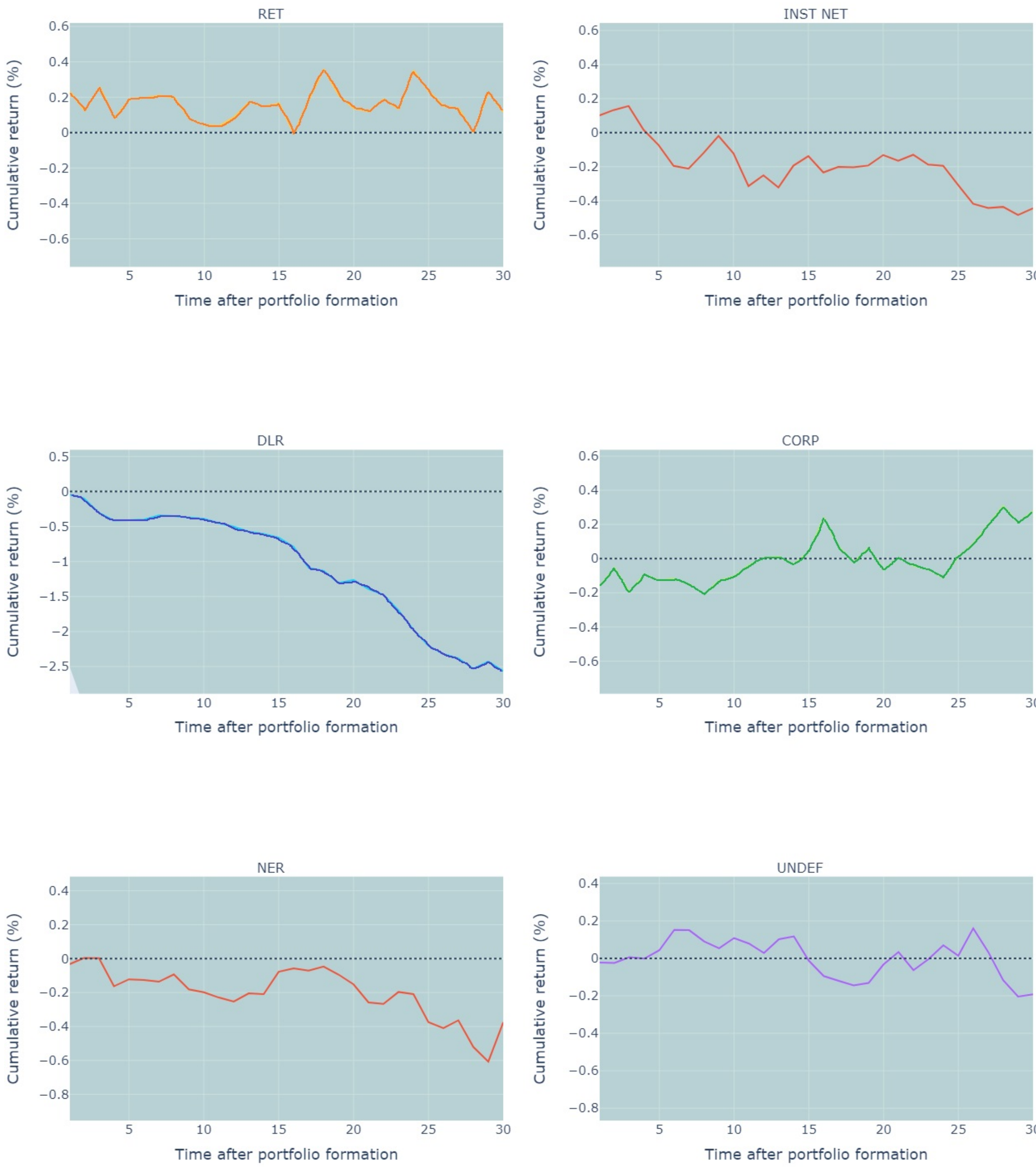


FIGURE 3.12: Daily cumulative post-formation portfolio returns. This figure shows average cumulative returns for the position of long P_1 portfolio and short P_7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation; periods overlap.

3.5.4 Predictive content of order flows at longer horizons

Our analysis has examined the connection between order flows and returns in the following trading period. However, we are also interested in whether the information contained in the order flow is helpful in forecasting returns over more than one period for daily and intraday time horizons. Therefore, we form the portfolios as before but analyse the predictive power with the time lags. So, we check how long the order flow signal has its predictive power. For example, we have information on the order flow in the time window t. We show the portfolio return with $lag = 1$ if the portfolio is formed in the last tick of the period t and is held until the end of the period $t + 1$. If $lag = 5$, we form the portfolios based on the order flow signal at the time window t; however, we enter the market in the last tick of the period $t + 4$ and hold the portfolio until the end of the period $t + 5$. In all cases, we hold the portfolios for only one trading period and check the predictive power of the signal from the order flow at longer horizons.

We analyse the average returns of holding the long-short portfolio (P_1-P_7) sorted on lagged order flow. So, we long portfolio P_1 and short portfolio P_7 and then hold for one period. The number of lags shows the time lag between the order flow signal and portfolio formation (60-second, 10-minute, and daily time periods). We allow longer lags up to ten periods after the order flow signal. For intraday trading, we do not allow portfolios to be held overnight.

Table 3.12 shows that the order flow signal differs for daily and intraday trading. If we work with daily periods, the order flow appears most informative for the first day after portfolio formation. The order flow information becomes insignificant after the first day for every client type. Hence, the information in daily flows is short-lived and incorporated into returns relatively quickly. The findings align with the literature on FX order flow, even though we run the analysis on different markets and asset classes. Menkhoff et al. (2016) find that the order flow is informative for the first two-three days, while our results show only one-day forecasting power. Our results are associated with the increased trading speed over the last decade.

Nevertheless, we observe longer forecast performance intraday. 60-second period analysis shows that order flow signals from corporate investors, retail traders, and institutional investors are informative for future returns for the following ten periods; however, the first period has the highest return. 10-minute order flows also show longer horizon forecast performance, up to eight periods after the signal formation. However, this is only true for corporate clients, retail traders and institutional investors.

Table 3.12 reports average returns of holding the portfolio (P_1-P_7) sorted on lagged order flows. t-statistics based on Newey-West standard errors are reported in brackets. The number of lags shows the time window lag between the order flow signal and portfolio formation (60-second, 10-minutes and daily time windows). We sort the portfolios by order flows of the previous day and by lags up to ten periods after the order flow signal. We show daily returns and intraday returns of 60-second and 10-minute periods. We do not allow portfolios to be held overnight.

Panel A. 60-second Order Flow						
N. of lags	CORP	DLR	RET	INST	NER	UNDEF
1	-2.721 (-83.09)	-1.226 (-28.897)	1.993 (62.957)	0.235 (6.902)	-0.105 (-2.508)	-0.365 (-8.653)
2	-0.222 (-15.035)	0.017 (0.876)	0.131 (9.248)	0.188 (12.038)	0.055 (2.934)	-0.032 (-1.52)
3	-0.051 (-5.179)	0.008 (0.629)	0.031 (3.294)	0.07 (6.703)	0.007 (0.578)	-0.022 (-1.486)
4	-0.028 (-3.779)	0.002 (0.236)	0.016 (2.347)	0.045 (5.858)	-0.001 (-0.211)	-0.012 (-1.175)
5	-0.021 (-3.517)	-0.004 (-0.384)	0.007 (1.263)	0.032 (5.242)	0.006 (0.89)	-0.016 (-1.715)
6	-0.016 (-3.314)	-0.002 (-0.335)	0.007 (1.588)	0.019 (3.766)	0.005 (0.812)	-0.006 (-0.932)
7	-0.012 (-2.856)	-0.001 (-0.171)	0.005 (1.15)	0.013 (2.829)	-0.005 (-0.859)	-0.01 (-1.662)
8	-0.012 (-3.153)	-0.005 (-1.09)	0.01 (2.735)	0.005 (1.4)	-0.005 (-1.191)	-0.016 (-3.328)
9	-0.007 (-1.946)	-0.008 (-1.517)	0.003 (1.055)	0.006 (2.165)	-0.003 (-0.449)	-0.011 (-2.004)
10	-0.0 (-0.116)	-0.007 (-1.359)	-0.003 (-1.202)	0.014 (4.389)	-0.005 (-1.188)	-0.011 (-2.399)
Panel B. 10-minute Order Flow						
N. of lags	CORP	DLR	RET	INST	NER	UNDEF
1	-4.949 (-16.421)	-4.772 (-14.592)	0.127 (0.447)	4.508 (15.242)	0.788 (2.77)	-1.089 (-3.462)
2	-0.621 (-4.729)	0.056 (0.369)	0.263 (2.046)	0.542 (4.321)	0.299 (2.135)	-0.293 (-1.965)
3	-0.392 (-4.501)	-0.023 (-0.206)	0.121 (1.462)	0.221 (2.636)	0.002 (-0.038)	-0.089 (-0.945)
4	-0.105 (-1.622)	0.03 (0.397)	0.058 (0.938)	0.117 (1.898)	-0.005 (-0.134)	-0.015 (-0.149)
5	-0.118 (-2.298)	-0.007 (-0.006)	0.0 (0.001)	0.03 (0.574)	0.081 (1.479)	0.008 (0.269)
6	-0.025 (-0.577)	-0.082 (-1.455)	-0.008 (-0.194)	0.078 (1.814)	-0.014 (-0.416)	-0.102 (-1.798)
7	-0.045 (-1.277)	-0.038 (-0.629)	0.014 (0.404)	0.005 (0.063)	0.009 (0.543)	-0.061 (-1.21)
8	-0.023 (-0.718)	-0.084 (-2.017)	0.004 (0.112)	0.029 (0.829)	0.029 (0.921)	-0.074 (-1.98)
9	-0.026 (-0.936)	-0.075 (-1.17)	0.008 (0.255)	-0.016 (-0.543)	-0.004 (-0.415)	-0.073 (-1.978)
10	-0.031 (-1.248)	-0.053 (-0.893)	0.0 (0.014)	0.091 (3.73)	0.036 (1.816)	-0.058 (-1.565)
Panel C. Daily Order Flow						
N. of lags	CORP	DLR	RET	INST	NER	UNDEF
1	-25.705 (-1.456)	-39.516 (-1.791)	40.955 (2.494)	18.338 (1.115)	-1.971 (-0.133)	14.062 (0.919)
2	15.493 (1.633)	-1.769 (-0.166)	-11.686 (-1.24)	0.733 (0.1)	3.776 (0.447)	-1.043 (-0.134)
3	-12.703 (-2.225)	-16.835 (-2.015)	8.653 (1.404)	1.808 (0.36)	-0.485 (-0.087)	4.339 (0.733)
4	5.686 (1.225)	-6.621 (-1.187)	-7.608 (-1.521)	-6.132 (-1.298)	-9.237 (-2.42)	1.871 (0.448)
5	-1.421 (-0.41)	-1.839 (-0.435)	4.438 (1.3)	-3.905 (-1.125)	2.873 (0.868)	-0.381 (-0.117)
6	0.333 (0.098)	-0.602 (-0.203)	2.144 (0.692)	-4.349 (-1.535)	-0.848 (-0.307)	2.887 (1.098)
7	0.283 (0.122)	0.309 (0.091)	-0.66 (-0.257)	-1.516 (-0.618)	0.41 (0.16)	1.224 (0.514)
8	-1.441 (-0.599)	-1.743 (-0.667)	-0.606 (-0.241)	2.791 (1.384)	0.989 (0.506)	-2.016 (-0.906)
9	2.734 (1.509)	0.443 (0.199)	-3.438 (-1.865)	2.364 (1.251)	-3.083 (-1.653)	-1.742 (-0.909)
10	0.613 (0.362)	-3.279 (-1.578)	-1.551 (-0.805)	-0.838 (-0.438)	-0.004 (-0.003)	-0.131 (-0.078)

TABLE 3.12: Order flow portfolios forecasting performance for longer horizons. The table reports average returns of holding the portfolio difference (P_1-P_7) sorted on lagged order flow. *t*-statistics based on Newey-West standard errors are reported in brackets. Number of lags (N. of lags) show the time-window lag between order flow signal and portfolio formation. We allow longer lags up to ten periods after the order flow signal. We show daily returns, and intraday returns of 60-second and 10-minute periods or time windows. We do not allow portfolios to be held overnight.

3.6 Drivers of Flows

We seek to provide a better understanding of the drivers of clients' order flows and shed light on the source of the negative correlations discussed earlier. First, we examine whether some client groups' order flows systematically lead to other groups' order flows. Second, we study whether the investors' order flows differ in their response to lagged index futures returns. We run panel regressions of order flows, buyer- and seller-initiated trading volume, and further explanatory variables to investigate the question. We run fixed effects panel regressions for 60 seconds, 10 minutes time periods inside the day and daily periods. We divide the volume by 10,000.

We estimate the model on intraday time frequencies as following:

$$\begin{aligned} OF^{c}_{i,t+1} &= \beta_0 + \beta_1 OF^{CORP}_{i,t} + \beta_2 OF^{DLR}_{i,t} + \beta_3 OF^{RET}_{i,t} + \beta_4 OF^{NER}_{i,t} + \beta_5 OF^{INST}_{i,t} + \\ &+ \beta_6 R_{i,t} + \beta_7 R_{RTS,t} + \beta_8 R_{MIX,t} + \beta_9 R_{i,day} + \beta_{10} R_{RTS,day} + \beta_{11} R_{MIX,day} + \epsilon_{i,t}, \end{aligned} \tag{3.4}$$

where OF is a customer order flow, c denotes one of five trading client groups, i denotes futures contract, $R_{i,t}$ is a return of the futures contract i for the period t, RTS and MIX denote RTS and MIX futures contracts as market index benchmarks, day stays for the previous trading day.

Similarly, we conduct the analysis on daily time frequencies as following:

$$\begin{aligned} OF^{c}_{i,t+1} &= \beta_0 + \beta_1 OF^{CORP}_{i,t} + \beta_2 OF^{DLR}_{i,t} + \beta_3 OF^{RET}_{i,t} + \beta_4 OF^{NER}_{i,t} + \beta_5 OF^{INST}_{i,t} + \\ &+ \beta_6 R_{i,t} + \beta_7 R_{RTS,t} + \beta_8 R_{MIX,t} + \beta_9 R_{i,week} + \beta_{10} R_{RTS,week} + \beta_{11} R_{MIX,week} + \epsilon_{i,t}, \end{aligned} \tag{3.5}$$

where OF is a customer order flow, c denotes one of five trading client groups, i denotes futures contract, $R_{i,t}$ is a daily return of the futures contract i for the period t, RTS and MIX denote RTS and MIX future contracts as market index benchmarks, $week$ stays for the previous five trading days.

Similarly we run the regressions for buyer- and seller initiated trading volume and show the results in Appendix 3.8.5.

Order flow results are reported in Tables 3.13, 3.17, 3.18. For each client group, we only include lagged order flows, lagged returns of the asset, and lagged returns of the market index futures. Using more than one lag of order flows in the regressions generally yields insignificant coefficient estimates, so we restrict the regressions to include one lag.

We find that intraday order flows of dealers, retail traders and institutional investors are significantly positively related to the order flows of other trading client groups. Order flows of corporate clients and non-residents are also significantly related to the order flows of the other groups. Our results show that we should consider intraday and daily order flow analyses separately due to the opposite coefficient signs. For daily results, we observe negative adjusted R2 meaning the insignificance of explanatory variables. The results may be improved with the increase in the sample size.

Order flows, buyer- and seller-initiated trading volumes are positively driven by their own lagged intraday and daily flows. Overall, there are numerous interrelationships between clients' order flows and their lags, but we suggest running further deeper analysis to find structural relationships between them.

As we do not find stable results of the lagged returns for intraday order flows, we execute an additional analysis for different groups depending on the futures' underlying assets. We run the same models (Equation 3.4, 3.5) for subcategories of

	Dependent variable:				
	OF^{CORP}_{t+1}	OF^{DLR}_{t+1}	OF^{RET}_{t+1}	OF^{INST}_{t+1}	OF^{NER}_{t+1}
	(1)	(2)	(3)	(4)	(5)
OF^{CORP}	0.287^{***}	0.004^{***}	0.029^{***}	-0.002	-0.010^{***}
	(0.003)	(0.0005)	(0.003)	(0.001)	(0.001)
OF^{DLR}	-0.154^{***}	0.357^{***}	0.073^{***}	0.020^{***}	0.021^{***}
	(0.005)	(0.001)	(0.004)	(0.002)	(0.001)
OF^{RET}	0.034^{***}	0.005^{***}	0.233^{***}	0.048^{***}	-0.012^{***}
	(0.003)	(0.0005)	(0.003)	(0.001)	(0.001)
OF^{INST}	-0.109^{***}	0.004^{***}	0.083^{***}	0.349^{***}	-0.014^{***}
	(0.003)	(0.0005)	(0.003)	(0.001)	(0.001)
OF^{NER}	0.036^{***}	0.029^{***}	0.067^{***}	0.040^{***}	0.134^{***}
	(0.004)	(0.001)	(0.004)	(0.002)	(0.001)
R	-0.005^{***}	0.0001	0.003^{***}	0.001^{***}	0.001^{***}
	(0.0004)	(0.0001)	(0.0003)	(0.0002)	(0.0001)
R_{RTS}	0.001	-0.0001	0.001^{**}	-0.001^{***}	-0.001^{***}
	(0.001)	(0.0001)	(0.001)	(0.0003)	(0.0002)
R_{MIX}	-0.0005	0.0001	-0.001	0.002^{***}	-0.0004^{*}
	(0.001)	(0.0001)	(0.001)	(0.0004)	(0.0002)
R_{day}	-0.00002^{***}	0.00000^{*}	0.00000	0.00002^{***}	0.00000
	(0.00001)	(0.00000)	(0.00001)	(0.00000)	(0.00000)
$R_{RTS,day}$	0.0003^{***}	-0.00000	-0.0001^{***}	-0.0002^{***}	0.00000
	(0.00003)	(0.00000)	(0.00003)	(0.00001)	(0.00001)
$R_{MIX,day}$	-0.0003^{***}	-0.00000	0.0001^{**}	0.0002^{***}	-0.00001
	(0.00004)	(0.00001)	(0.00003)	(0.00001)	(0.00001)
Observations	3,181,271	3,181,271	3,181,271	3,181,271	3,181,271
R^2	0.088	0.124	0.043	0.142	0.022
Adjusted R^2	0.087	0.124	0.043	0.142	0.022
F Statistic (df = 11; 3181035)	$27{,}748.470^{***}$	$41{,}017.680^{***}$	$13{,}037.230^{***}$	$47{,}693.440^{***}$	$6{,}517.826^{***}$
Note:					$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

TABLE 3.13: Drivers of customer order flow, 60-second time dimension. This table reports results for panel regressions of customer order flows (OF) on lagged customer order flow (CORP for corporate clients, DLR, dealres, RET, retail traders, INST, institutional investors, and non-residents, NER). The regressions also consider lagged returns as additional regressors (lagged future returns and lagged market indices returns over the previous trading day, R_{day}, $R_{RTS,day}$ and $R_{MIX,day}$ respectively.

the underlying assets. Our results in Table 3.25 reveal an essential factor affecting the intraday order flow: its lag, which indicates that the customers' order flow measures are persistent. However, we observe this momentum pattern only for retail traders for the daily order flow. In contrast to Boehmer et al. (2021) we do not observe past returns as an essential factor affecting order flows. The results are mixed across different time dimensions, investor groups and asset classes being insignificant in many cases.

3.7 Conclusion and further research

Concluding remarks

How do different trading practices by client types affect market outcomes? Do investors systematically trade in opposite directions, or is their trading activity positively correlated? Does trader's order flow information differs depending on the investment horizon? What characterises different investor groups' futures trading? We address these questions using data on MOEX futures. The granularity and completeness of our dataset enable us to analyse the extent to which any group of investors have a comprehensive advantage in information relevant to future returns.

Our results reveal that the retail traders' order flows have predictive power for the returns in the futures market for the next trading period in 30 and 60 seconds time frequencies. In longer-time windows, the trading behaviour of retail clients loses its predictive power. Corporate clients tend to trade in the opposite direction to retail and institutional traders, and their behaviour does not show informativeness. Retail and institutional investors trade in the same direction, which could be explained as retail traders following large institutional order flow signals. Retail investors trade in the opposite direction to dealers inside the day, while their order flows correlate positively in a longer daily perspective.

Contemporaneous analysis reveals that non-residents, retail traders, and institutional investors pushing the prices up in the current period a high-frequency time dimensions, corporate clients and dealers trade in the opposite direction to the market. With institutional investors' order flow, our research shows its impact on the current price movements. However, the market overreacts to large institutional trades, and we observe prices decrease in the next period.

Analysing longer intraday forecast performance, we find that 60-second order flow signals are informative for future returns, with the highest return in the first minute. 10-minute order flows also show longer horizon forecast performance, up to eight minutes after the signal formation.

The results suggest that the buying power and order flow are highly informative about daily and intraday future returns. We also find differences in customers' intraday and longer trading practices. The order flow information becomes insignificant after the first day for every client type. Hence, the information in daily flows is short-lived and incorporated into returns relatively quickly.

Splitting the data into different underlying asset classes reveals dissimilarities in the impact of different client groups on market outcomes. Retail traders', institutional investors' and non-residents' order flows are good predictors for intraday return in commodity futures. In contrast, only institutional inventors' order flow positively predicts returns in futures on stocks. In a high-frequency environment, order flows of corporate clients, retail traders, and non-residence positively predict returns on currency futures, while the order flow of institutional investors does not.

Our findings show that non-residents obtain information about price movements of currency and commodity futures and correctly predict intraday returns.

In this chapter, we show empirically in the three-fold investigation (correlation, regression and portfolio analyses) that the order flow signal differs for daily and intraday trading, and the study needs to be built separately for intraday frequencies such as high-frequency trading and algorithmic trading, and daily trading. Also, we find the major differences while splitting the data into sub-samples based on the underlying asset classes and in different volatility environments. We discovered that non-residents anticipate future returns in currency and commodity futures. This makes economic sense due to the interest in the Russian commodity futures by non-resident clients. In contrast to our other findings, corporate clients' order flows positively predicts future returns in currency exchange rate futures. Russian corporate clients trade currency futures to hedge their exchange rate risk exposure. Moreover, we find that order flows have different predictive power in the ultra-high volatility environment.

Limitations of current study and further research directions

There are some areas in the current research that could be developed in further work and some limitations that we have imposed could be relaxed.

In Section 3.4, our results show small R-squares by analysing the predictive power of flows using panel regressions with fixed effects. We know from the related literature that it is tough to explain returns, even with the dataset of costumer's order flows. Based on asset class subsample analysis, we observe higher R-squares for currency exchange futures than stock futures. The limitation of a small number of control variables could be relaxed in further research by focusing on the futures of the same asset class. For example, to explain stock returns, scholars control for important new announcements, management changes, specificity of the industry sector, reporting period, dividend payment periods, size of the company, returns in highly correlated assets and many others. Another set of control variables could be used separately for commodity futures or foreign exchange futures analysis.

In Section 3.5, we investigate the sensitivity of our results to the number of portfolios and show the robustness of our main finding. Specifically, we show that no matter how many portfolios we form from 27 initial futures (three, five or seven, retail and institutional traders tend to have positive cumulative returns intraday and also via investing for up to five trading days, while corporate clients and dealers tend to trade in the opposite direction (Appendix 3.8.8).

Researchers can further verify the feasibility of our results for futures on different asset classes traded in other markets. Future research may also explore the possibility of incorporating the morning and evening trading session data to provide further insights into the informativeness of traders outside regular trading hours.

To continue our discussion in Section 3.6, a potential future research direction is to look at drivers of order flows in more detail based on industry or firm-level characteristics. The question of whether order flow has a permanent or transitory effect on prices is a central one in microstructure literature (Hasbrouck 1991*a*,*b*). Since we find substantial heterogeneity about the forecasting power of different client groups' order flows, the question of whether order flows signal information relevant for permanent asset price changes is of interest. Moreover, we could look into the details, whereas a permanent movement in price changes would indicate that order flow conveys information about fundamentals. Moreover, the results show that factors that drive flows differ across client groups. A deeper investigation of other factors and drivers could be considered in further research.

Finally, it is worth further examining whether order flows are related to future macro fundamentals as suggested, for example, by Evans & Lyons (2008). Scholars investigate this question in a cross-sectional setting, looking at growth rates in macroeconomic fundamentals. Again, it is worth conducting an analysis of the futures of different asset classes separately.

Note that the main finding in this paper relies on the data from the developing market. If a similar predictive relationship exists in other markets, it is more convincing that a general economic mechanism exists behind the phenomenon. We also leave this for future research.

3.8 Appendices

3.8.1 Appendix 1. Descriptive statistics

	Ticker	Description	Trades number (1)	Ruble volume (2)	Group (3)
1	Si	USD/RUB Exchange Rate	663.74	3914.85	1
2	RTS	RTS Index	371.56	800.10	2
3	BR	Brent oil	262.66	2831.58	3
4	SBRF	Sberbank ordinary shares	142.09	506.71	4
5	GAZR	Gazprom ordinary shares	63.88	272.09	4
6	GOLD	Gold	62.41	212.51	3
7	Eu	EUR/RUB Exchange Rate	43.12	258.82	1
8	SILV	Silver	31.00	348.62	3
9	VTBR	VTB BANK ordinary shares	27.51	212.96	4
10	MIX	MOEX Russia Index	22.29	37.61	2
11	ED	EUR/USD Exchange Rate	19.82	187.69	1
12	MXI	MOEX Russia Index (mini)	15.30	46.56	2
13	LKOH	LUKoil ordinary shares	13.20	35.93	4
14	GMKN	MMC Norilsk Nickel ordinary shares	7.99	27.74	4
15	MGNT	Magnit ordinary shares	7.84	60.84	4
16	SBPR	Sberbank preferred shares	5.80	16.08	4
17	ROSN	Rosneft ordinary shares	5.40	11.76	4
18	SNG	Surgutneftegas preferred shares	3.97	9.98	4
19	NG	Natural Gas	3.93	18.92	3
20	AFLT	Aeroflot ordinary shares	3.66	20.17	4
21	ALRS	ALROSA ordinary shares	2.94	11.10	4
22	PLT	Platinum	2.45	7.31	3
23	YNDF	Yandex N.V. ordinary shares	1.95	3.86	4
24	TATN	Tatneft ordinary shares	1.79	3.86	4
25	SPYF	SPDR SandP 500 ETF Trust	1.56	13.31	2
26	NLMK	NLMK ordinary shares	1.06	2.67	4
27	RTSM	RTS Index (mini)	0.49	1.26	2

TABLE 3.14: Descriptive statistics of futures on MOEX. Table shows the average number of trades per minute (1) and the average trading ruble volume per minute (2) for the most liquid twenty seven future contracts traded on MOEX from January 2020 to September 2021. Futures are sorted by the number of trades (1). We sort futures by the underlying asset class into four groups (3): Group 1: futures on currency exchange; Group 2: futures on market indices; Group 3: commodity futures; Group 4: stock futures.

3.8.2 Appendix 2. Correlation matrices

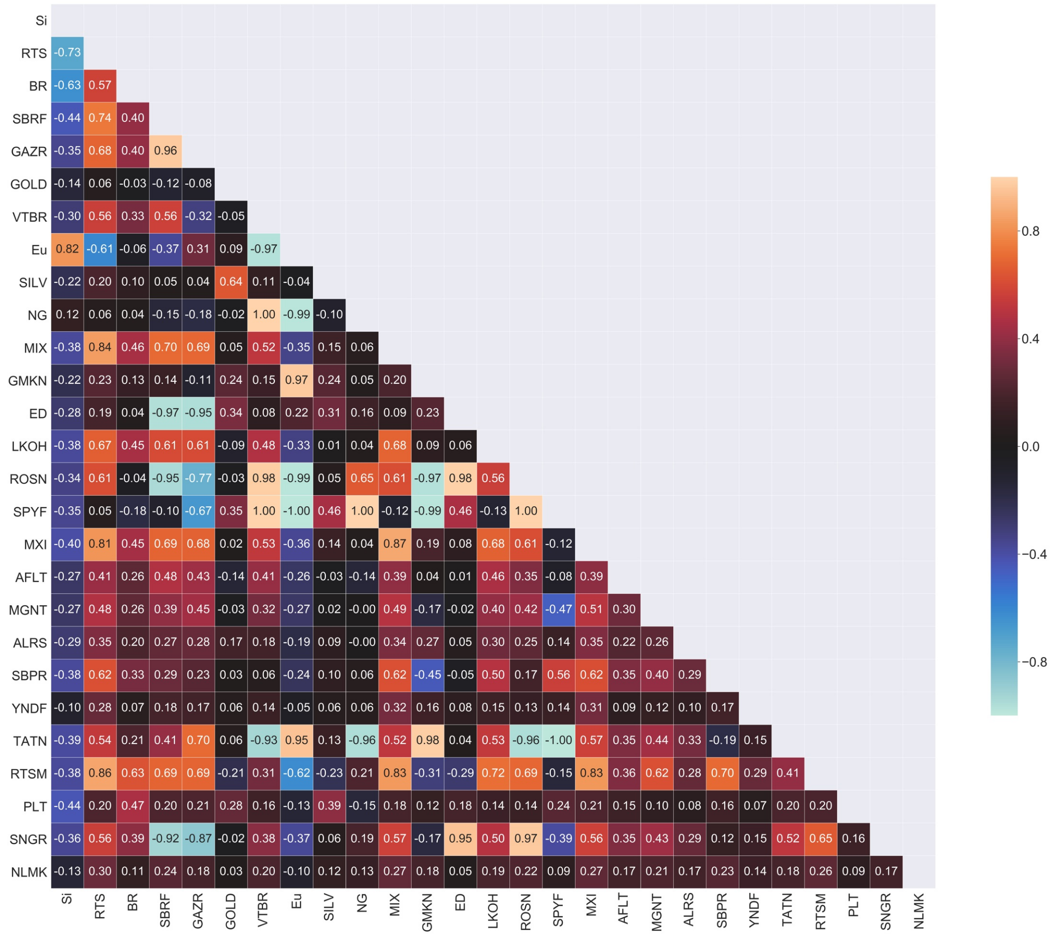


FIGURE 3.13: Correlation matrix between 60-second futures returns

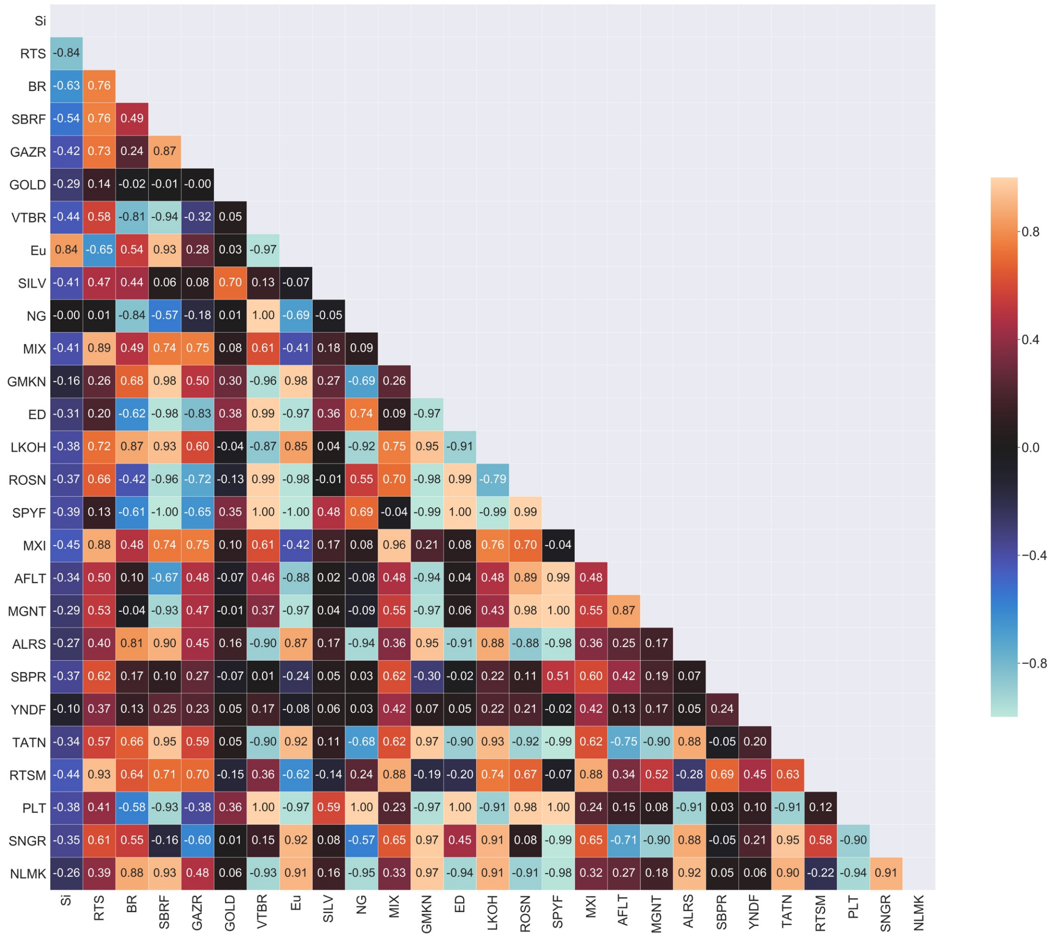


FIGURE 3.14: Correlation matrix between 5-minute futures returns

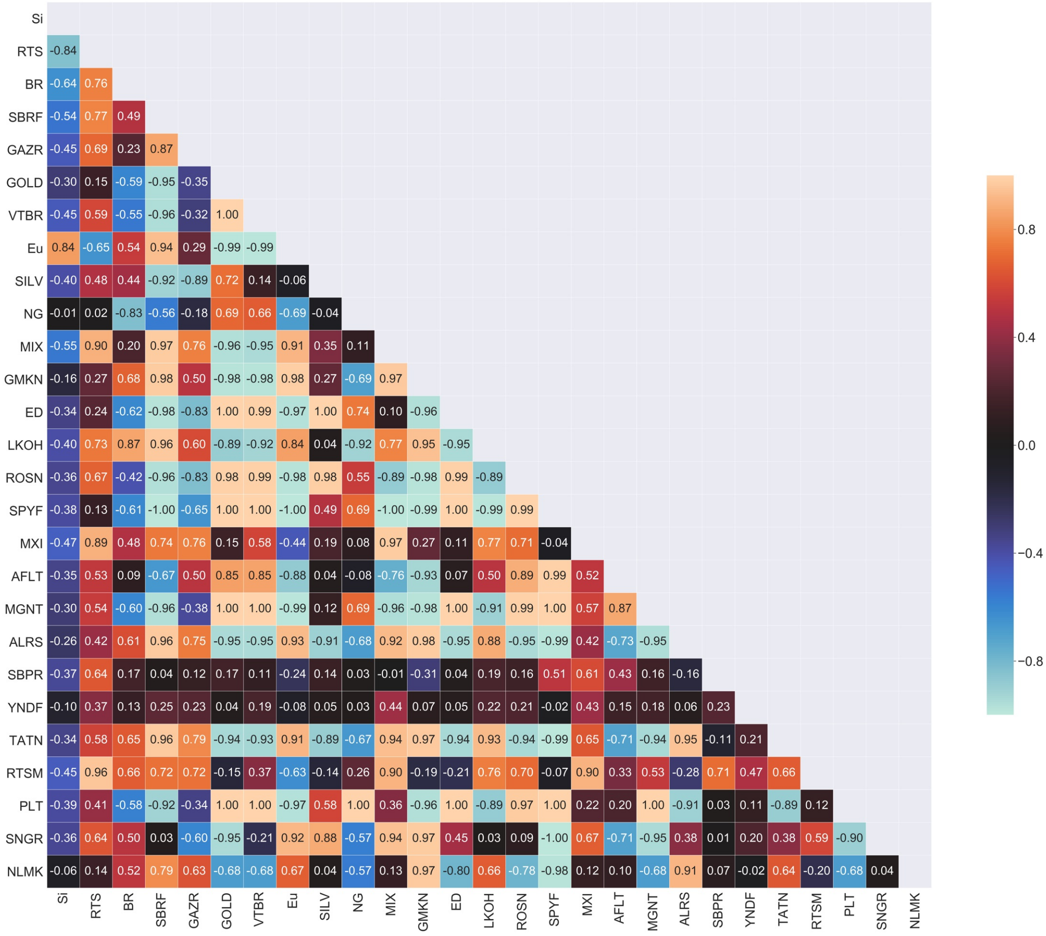


FIGURE 3.15: Correlation matrix between 10-minute futures returns

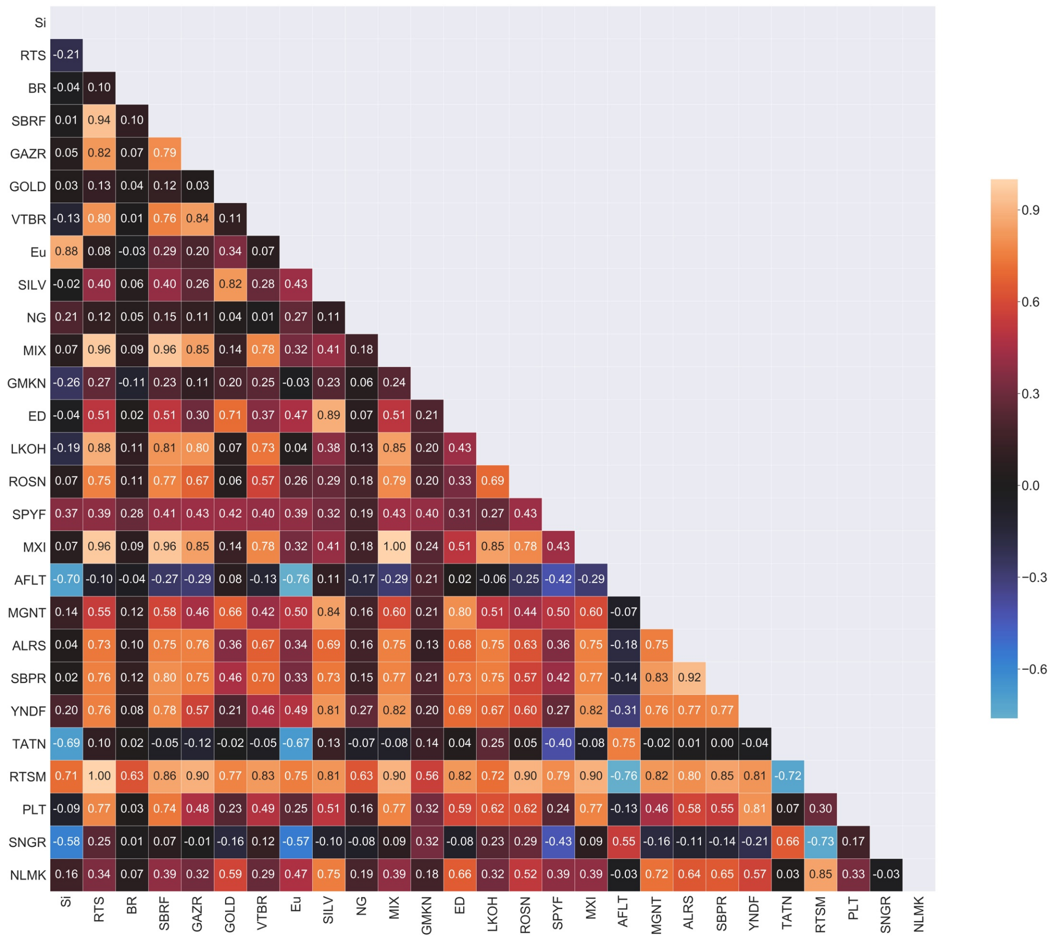


FIGURE 3.16: Correlation matrix between daily futures returns

3.8.3 Appendix 3. Predictive power of flows

Panel A: buyer-initiated trading volume

	Dependent variable: Returns					
	30 seconds	60 seconds	5 minutes	10 minutes	30 minutes	60 minutes
	(1)	(2)	(3)	(4)	(5)	(6)
$St.Volume^{CORP}$	−0.00003***	−0.00002***	−0.00001***	−0.00001***	−0.00001***	−0.00001***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
$St.Volume^{DLR}$	−0.0001***	−0.0001***	−0.00004***	−0.00003***	−0.00002	−0.00001
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00002)
$St.Volume^{RET}$	0.00001***	0.00001***	0.00000**	0.00000	−0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
$St.Volume^{INST}$	0.00000	0.00002***	0.0001***	0.0001***	0.0001***	0.00005***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00001)	(0.00001)
$St.Volume^{NER}$	0.00000**	0.00000**	0.00000**	0.00000*	0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Constant	−0.0001	−0.00002	−0.0001	−0.00004	−0.0002	−0.003
	(0.00005)	(0.0001)	(0.0002)	(0.0005)	(0.001)	(0.003)
Observations	5,933,010	3,576,957	919,488	491,395	165,879	78,585
R^2	0.001	0.0002	0.0003	0.001	0.001	0.001
Adjusted R^2	0.001	0.0002	0.0003	0.001	0.001	0.001
F Statistic	4,003.466***	854.717***	319.787***	339.549***	164.692***	68.221***

Panel B: seller-initiated trading volume

	Dependent variable: Returns					
	30 seconds	60 seconds	5 minutes	10 minutes	30 minutes	60 minutes
	(1)	(2)	(3)	(4)	(5)	(6)
$St.Volume^{CORP}$	0.00003***	0.00002***	0.00000***	0.00001***	0.00001***	0.00000***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
$St.Volume^{DLR}$	0.0001***	0.0001***	0.00001	0.00000	−0.00000	−0.00001
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00002)
$St.Volume^{RET}$	−0.00002***	−0.00001***	−0.00000***	−0.00000***	−0.00000***	−0.00000***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
$St.Volume^{INST}$	0.00003***	0.00001***	−0.00001***	−0.00001	−0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00001)	(0.00001)
$St.Volume^{NER}$	−0.00001***	−0.00001***	−0.00000	0.00000	0.00000*	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Constant	0.00004	0.00005	−0.0002	−0.0003	−0.001	−0.003
	(0.00005)	(0.0001)	(0.0002)	(0.0005)	(0.001)	(0.003)
Observations	5,933,010	3,576,957	919,488	491,395	165,879	78,585
R^2	0.001	0.0002	0.00004	0.0001	0.0005	0.0001
Adjusted R^2	0.001	0.0002	0.00003	0.0001	0.0004	0.0001
F Statistic	4,279.276***	793.042***	35.354***	54.635***	75.982***	11.232**

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE 3.15: Panel regression results for standardized intraday buyer-initiated (Panel A) and seller-initiated (Panel B) trading volume. The table shows panel regression results with fixed effects pulled across all future contracts for the entire period using Equation 3.1.

	Dependent variable:					
	Returns					
	30 seconds	60 seconds	5 minutes	10 minutes	30 minutes	60 minutes
	(1)	(2)	(3)	(4)	(5)	(6)
$St.Volume^{CORP}$	-0.0001^{***}	-0.00004^{***}	-0.00001^{***}	-0.00002^{***}	-0.00004^{***}	-0.00005^{***}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00001)
$St.Volume^{DLR}$	-0.0001^{***}	-0.0001^{***}	-0.00004^{***}	-0.00003^{***}	-0.00003^{*}	-0.00004^{*}
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00002)	(0.00002)
$St.Volume^{RET}$	0.00001^{***}	0.00000^{***}	0.00000	0.00000	0.00000^{*}	-0.00001^{**}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
$St.Volume^{INST}$	-0.0001^{***}	-0.00003^{***}	0.00004^{***}	0.00004^{***}	0.00001	-0.00001
	(0.00000)	(0.00000)	(0.00000)	(0.00001)	(0.00001)	(0.00001)
$St.Volume^{NER}$	0.00002^{***}	0.00002^{***}	0.00001^{***}	0.00001^{*}	0.00001	-0.00001
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00001)	(0.00001)
Constant	-0.0001	-0.00004	-0.0001	-0.0002	-0.001	-0.002
	(0.00005)	(0.0001)	(0.0002)	(0.0004)	(0.001)	(0.003)
Observations	5,933,010	3,576,957	919,488	491,395	165,879	78,585
R^2	0.002	0.001	0.0004	0.001	0.002	0.002
Adjusted R^2	0.002	0.001	0.0004	0.001	0.002	0.002
F Statistic	$10{,}761.960^{***}$	$2{,}426.608^{***}$	358.012^{***}	415.394^{***}	373.547^{***}	158.900^{***}
Note:						$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

TABLE 3.16: Panel regression results for standardized intraday customer order flow. The table shows panel regression results with fixed effects pulled across all future contracts for the entire period using Equation 3.2.

3.8.4 Appendix 4. Portfolio analysis

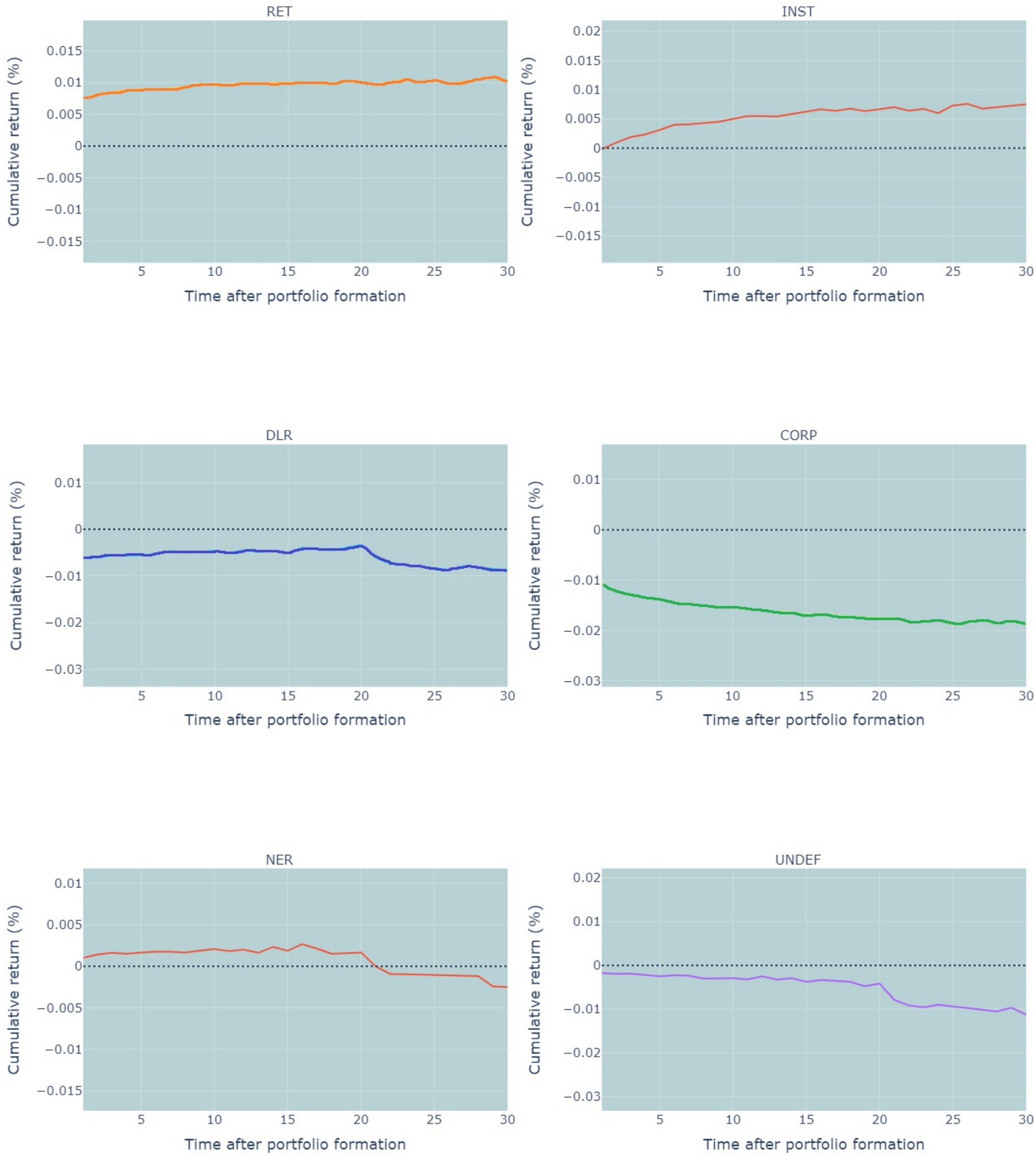


FIGURE 3.17: 60-seconds cumulative post-formation portfolio returns in low volatility environment. This figure shows average cumulative returns for the position of long P1 portfolio and short P7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation inside the day; periods overlap.

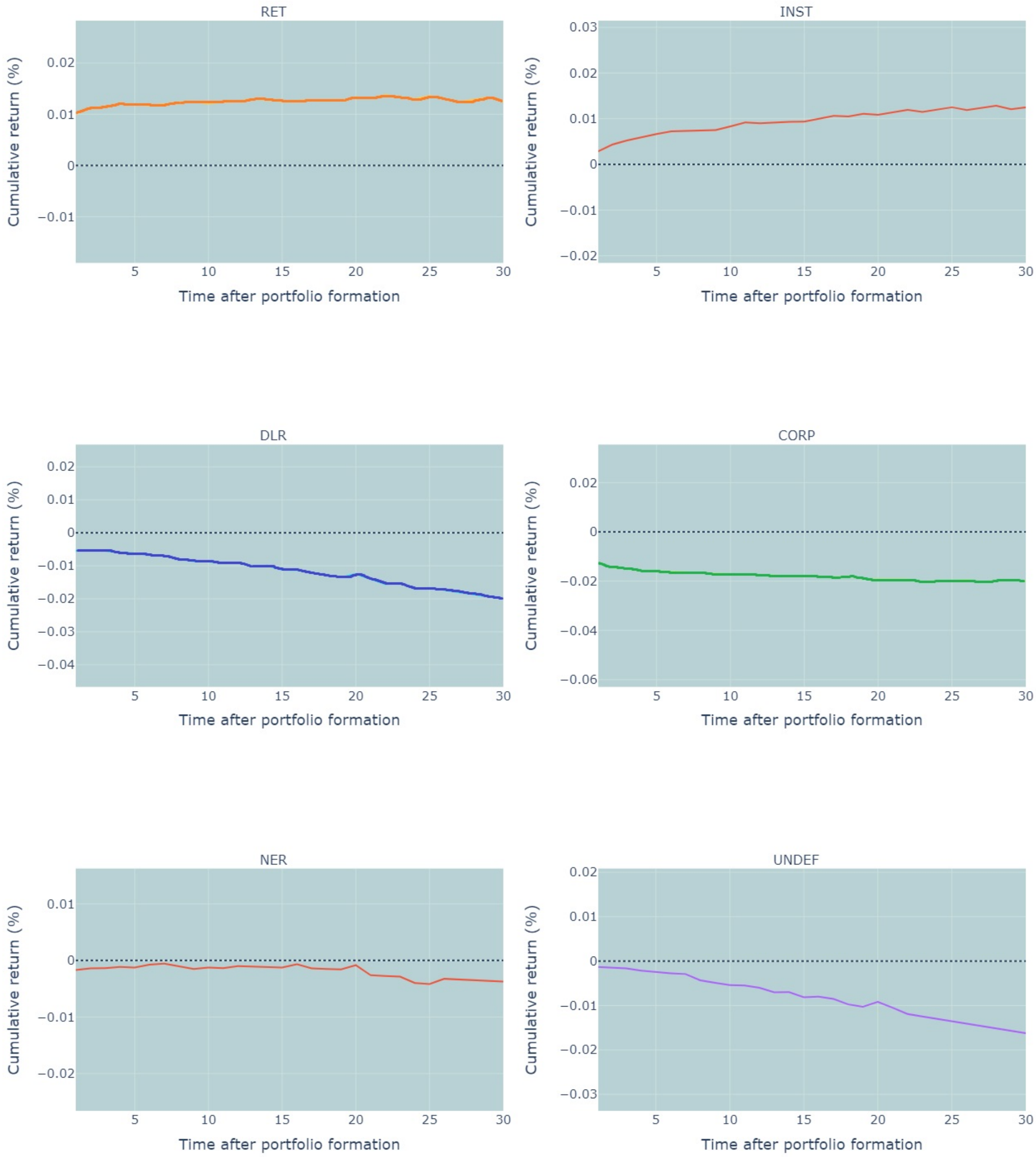


FIGURE 3.18: 60-seconds cumulative post-formation portfolio returns in high volatility environment. This figure shows average cumulative returns for the position of long P1 portfolio and short P7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation inside the day; periods overlap.

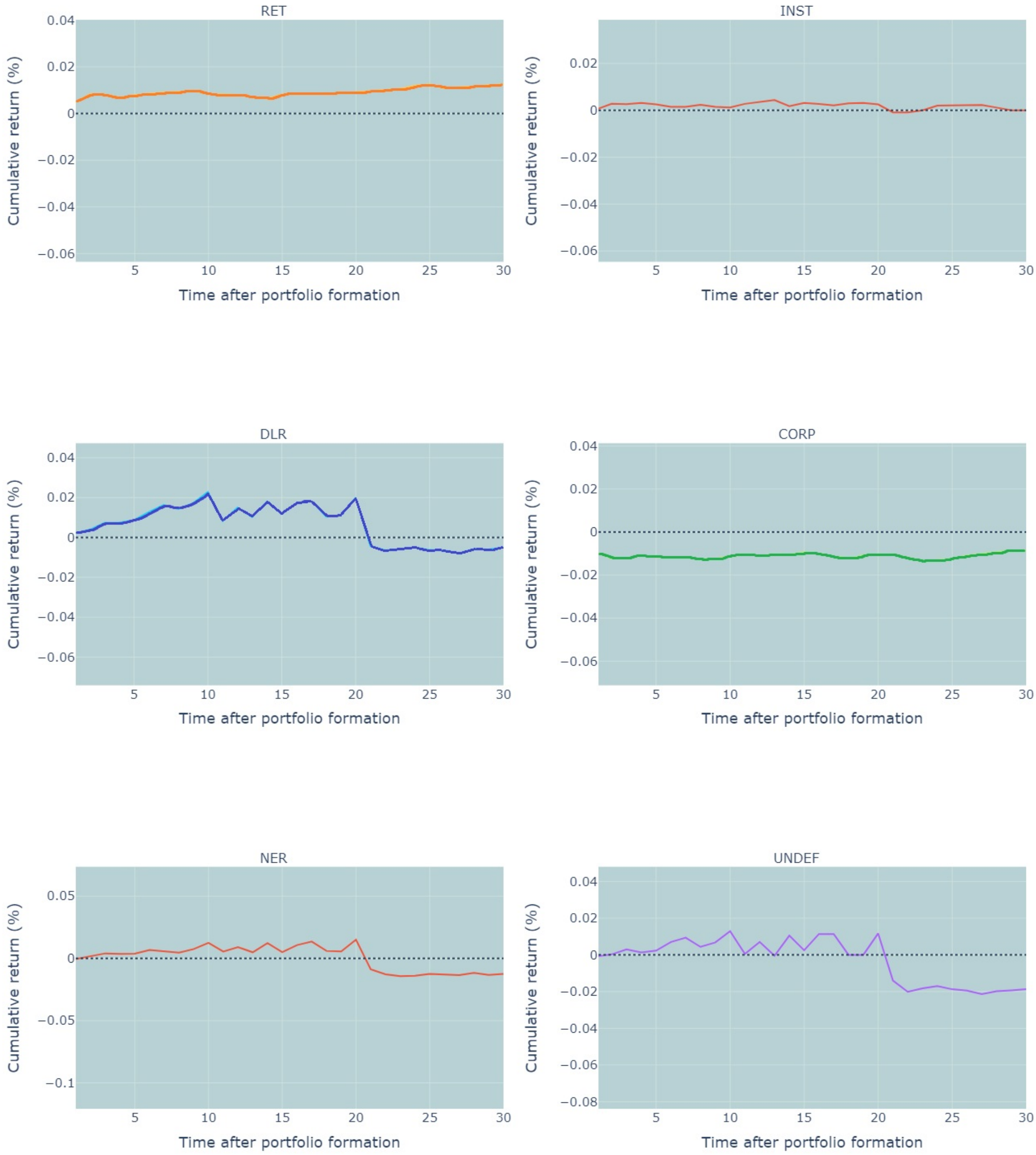


FIGURE 3.19: 60-seconds cumulative post-formation portfolio returns in ultra-high volatility environment. This figure shows average cumulative returns for the position of long P1 portfolio and short P7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation inside the day; periods overlap.

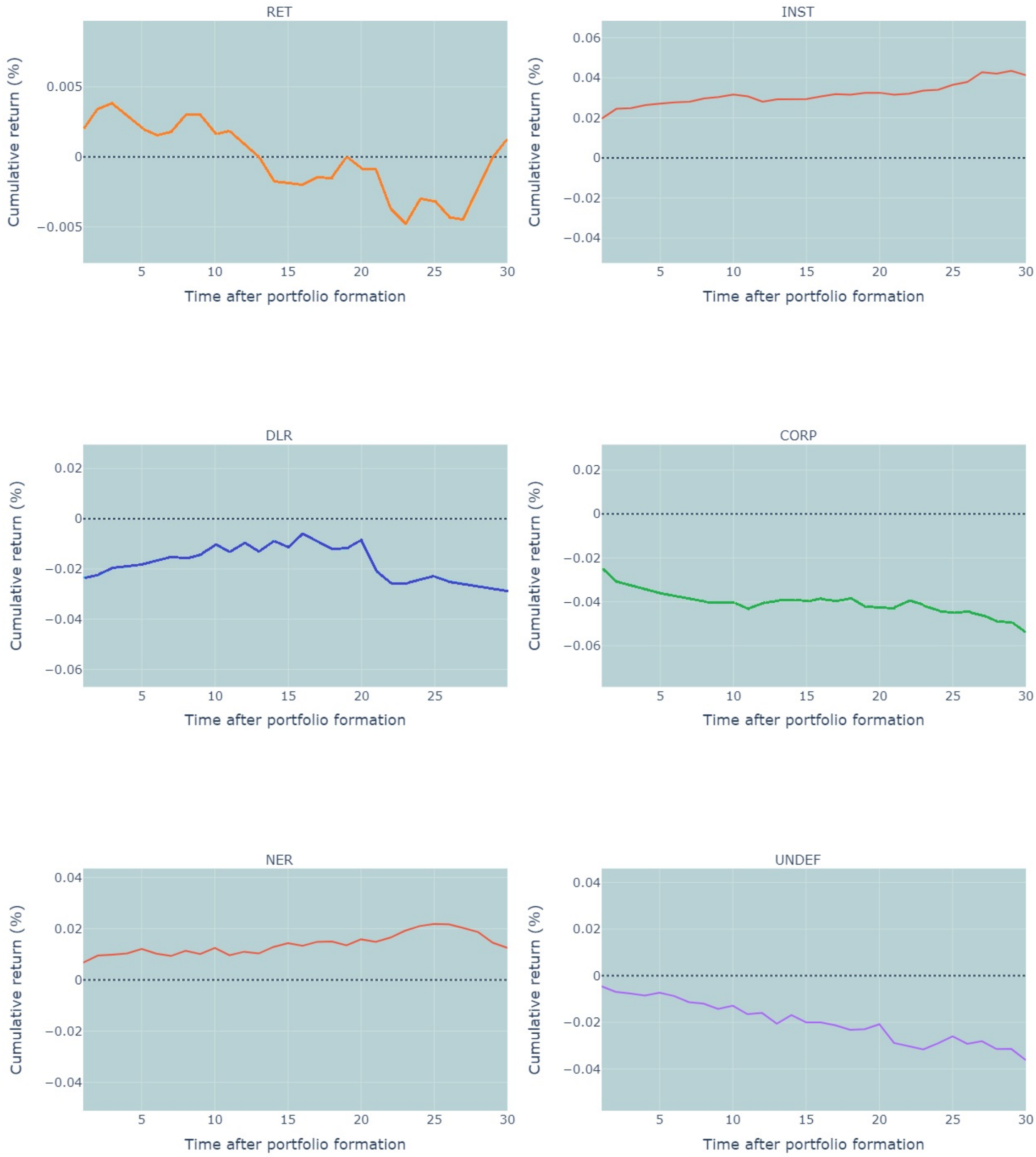


FIGURE 3.20: 10-minutes cumulative post-formation portfolio returns in low volatility environment. This figure shows average cumulative returns for the position of long P1 portfolio and short P7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation inside the day; periods overlap.

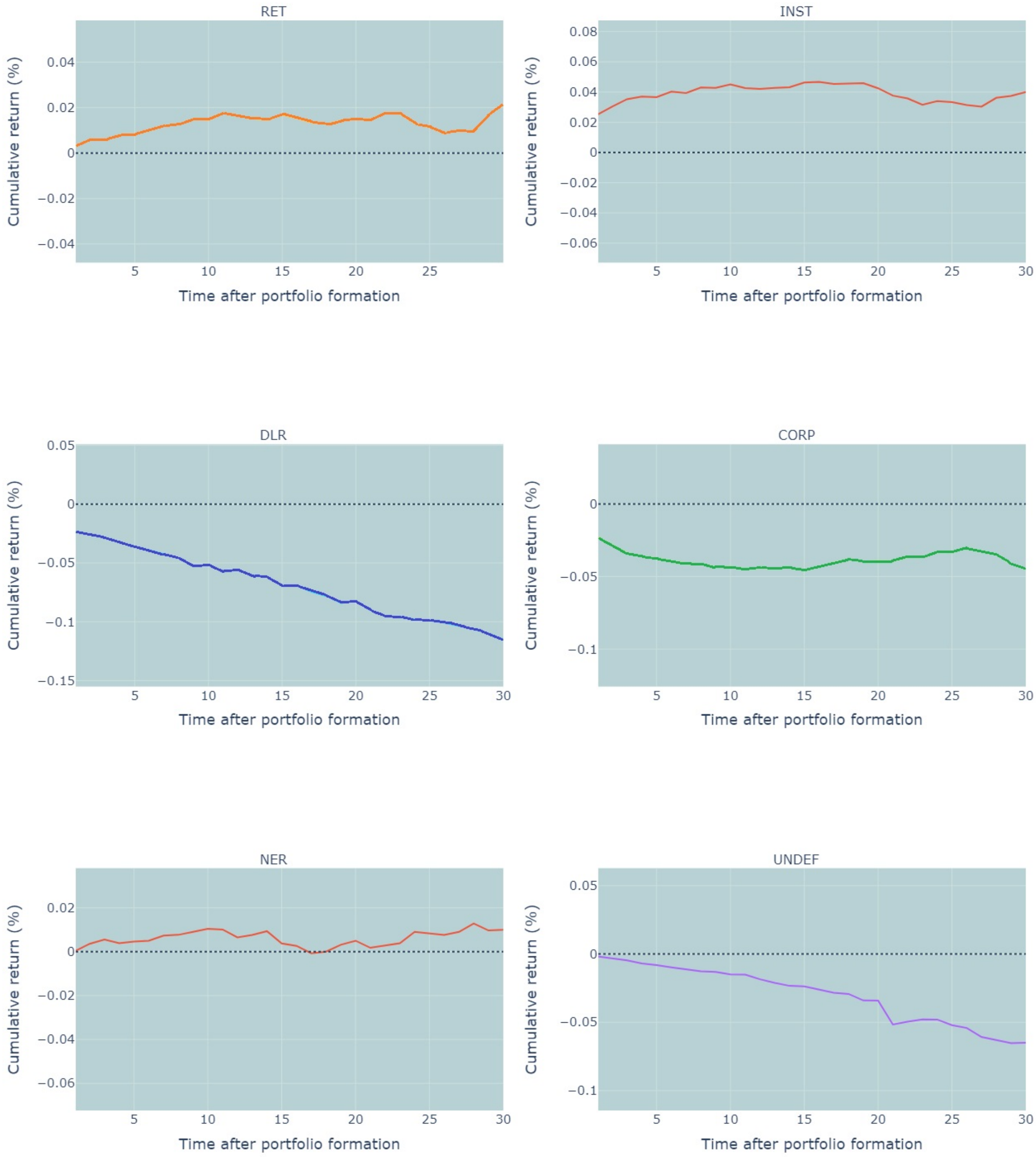


FIGURE 3.21: 10-minutes cumulative post-formation portfolio returns in high volatility environment. This figure shows average cumulative returns for the position of long P1 portfolio and short P7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation inside the day; periods overlap.

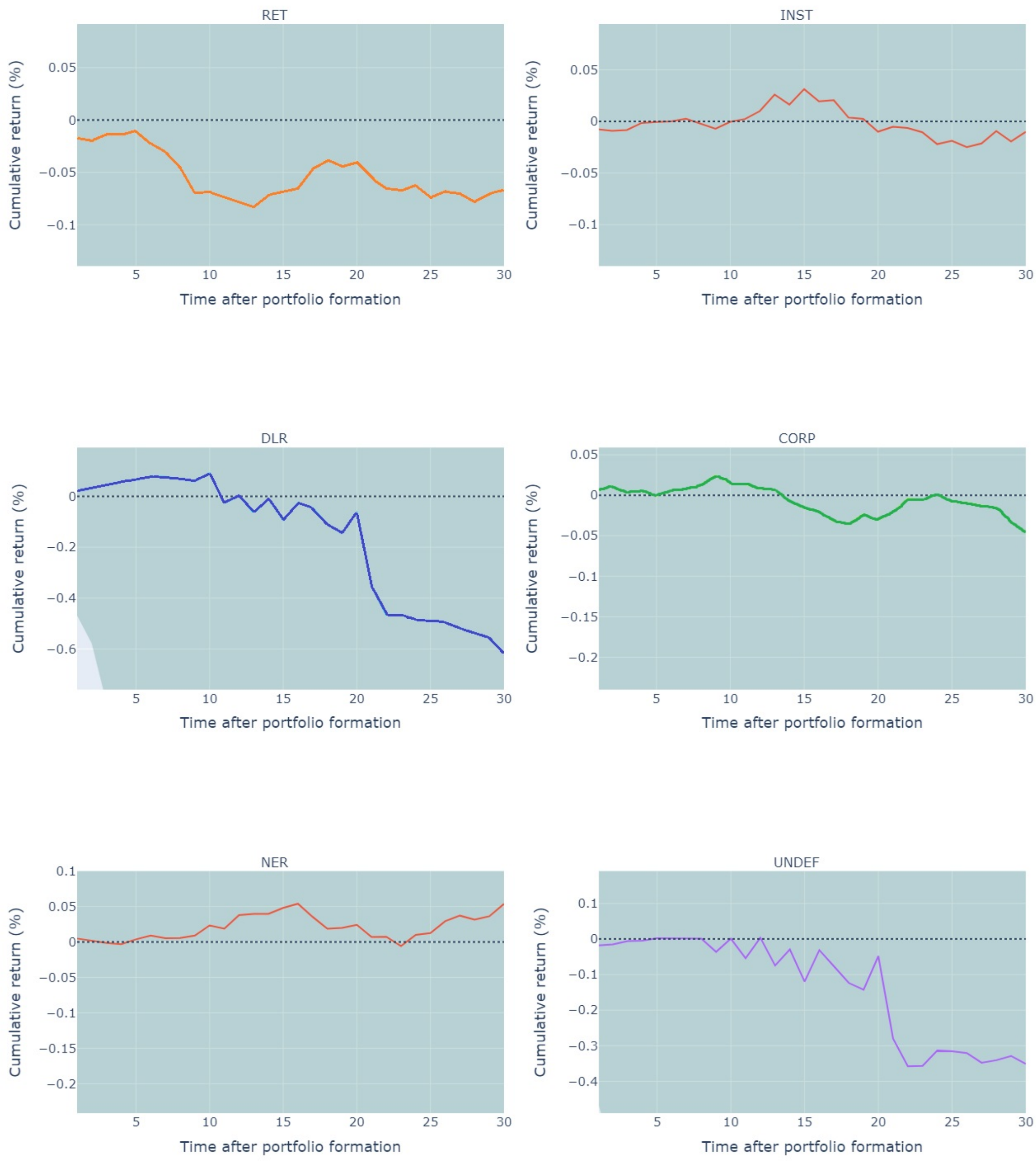


FIGURE 3.22: 10-minutes cumulative post-formation portfolio returns in ultra-high volatility environment. This figure shows average cumulative returns for the position of long P1 portfolio and short P7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation inside the day; periods overlap.

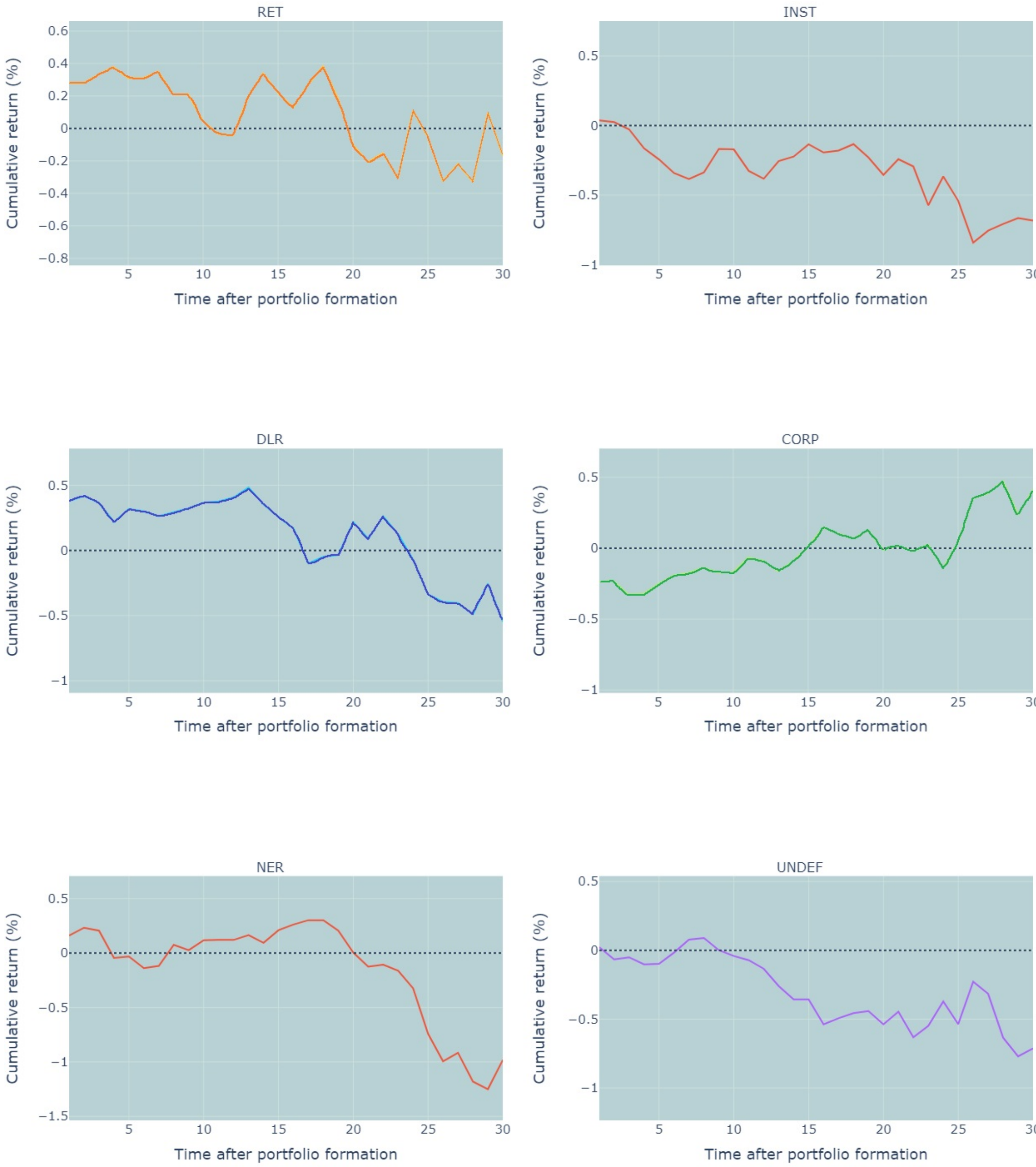


FIGURE 3.23: Daily cumulative post-formation portfolio returns in low volatility environment. This figure shows average cumulative returns for the position of long P1 portfolio and short P7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation inside the day; periods overlap.

FIGURE 3.24: Daily cumulative post-formation portfolio returns in high volatility environment. This figure shows average cumulative returns for the position of long P1 portfolio and short P7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation inside the day; periods overlap.

FIGURE 3.25: Daily cumulative post-formation portfolio returns in ultra-high volatility environment. This figure shows average cumulative returns for the position of long P1 portfolio and short P7 portfolio based on disaggregated order flows over the first 30 time windows or 30 days after the portfolio formation inside the day; periods overlap.

3.8.5 Appendix 5. Drivers of flows

	Dependent variable:				
	OF^{CORP}_{t+1}	OF^{DLR}_{t+1}	OF^{RET}_{t+1}	OF^{INST}_{t+1}	OF^{NER}_{t+1}
	(1)	(2)	(3)	(4)	(5)
OF^{CORP}	0.246***	0.013***	0.129***	0.092***	−0.077***
	(0.008)	(0.001)	(0.006)	(0.004)	(0.002)
OF^{DLR}	−0.072***	0.303***	0.158***	0.107***	−0.081***
	(0.011)	(0.002)	(0.009)	(0.005)	(0.002)
OF^{RET}	−0.070***	0.014***	0.409***	0.130***	−0.081***
	(0.008)	(0.001)	(0.006)	(0.004)	(0.002)
OF^{INST}	−0.013	0.011***	0.050***	0.430***	−0.073***
	(0.008)	(0.001)	(0.007)	(0.004)	(0.002)
OF^{NER}	−0.053***	0.044***	0.211***	0.115***	0.083***
	(0.010)	(0.002)	(0.009)	(0.005)	(0.002)
R	0.007***	−0.0001	−0.004***	−0.003***	0.0001
	(0.001)	(0.0002)	(0.001)	(0.0005)	(0.0002)
R_{RTS}	−0.012***	0.0003	0.004**	0.006***	0.001
	(0.002)	(0.0004)	(0.002)	(0.001)	(0.001)
R_{MIX}	0.005	−0.0005	0.0005	−0.003*	−0.001*
	(0.004)	(0.001)	(0.003)	(0.002)	(0.001)
R_{day}	−0.0002**	0.00001	0.00003	0.0001***	0.00001
	(0.0001)	(0.00001)	(0.0001)	(0.00003)	(0.00001)
$R_{RTS,day}$	0.002***	−0.00002	−0.001**	−0.001***	0.00003
	(0.0003)	(0.0001)	(0.0003)	(0.0002)	(0.0001)
$R_{MIX,day}$	−0.002***	0.00000	0.001*	0.001***	−0.00004
	(0.0003)	(0.0001)	(0.0003)	(0.0002)	(0.0001)
Observations	496,893	496,893	496,893	496,893	496,893
R^2	0.090	0.088	0.075	0.133	0.033
Adjusted R^2	0.090	0.088	0.075	0.133	0.032
F Statistic (df = 11; 496657)	4,484.356***	4,358.231***	3,670.950***	6,943.976***	1,519.479***
Note:					*p<0.1; **p<0.05; ***p<0.01

TABLE 3.17: Drivers of customer order flow, 10-minute time dimension. This table reports results for panel regressions of customer order flows (OF) on lagged customer order flow (CORP for corporate clients, DLR, dealres, RET, retail traders, INST, institutional investors, and non-residents, NER). The regressions also consider lagged returns as additional regressors (lagged futures returns and lagged market indices returns over the previous trading day, R_{day}, $R_{RTS,day}$ and $R_{MIX,day}$, respectively.

	Dependent variable:				
	OF^{CORP}_{t+1}	OF^{DLR}_{t+1}	OF^{RET}_{t+1}	OF^{INST}_{t+1}	OF^{NER}_{t+1}
	(1)	(2)	(3)	(4)	(5)
OF^{CORP}	-0.200^{***}	0.003	0.167^{***}	0.095^{***}	-0.018^{**}
	(0.064)	(0.011)	(0.052)	(0.030)	(0.008)
OF^{DLR}	-0.149^{*}	-0.057^{***}	0.166^{**}	0.158^{***}	-0.047^{***}
	(0.089)	(0.015)	(0.073)	(0.042)	(0.012)
OF^{RET}	-0.205^{***}	0.005	0.169^{***}	0.098^{***}	-0.015^{*}
	(0.064)	(0.011)	(0.052)	(0.030)	(0.009)
OF^{INST}	0.077	0.001	0.072	-0.073^{**}	-0.026^{***}
	(0.064)	(0.011)	(0.052)	(0.030)	(0.009)
OF^{NER}	-0.137	-0.050^{***}	0.485^{***}	0.013	-0.244^{***}
	(0.102)	(0.017)	(0.084)	(0.048)	(0.014)
R	-0.006	0.001	-0.001	0.007^{**}	0.001
	(0.007)	(0.001)	(0.006)	(0.003)	(0.001)
R_{RTS}	0.127^{***}	-0.002	-0.034	-0.094^{***}	0.005
	(0.037)	(0.006)	(0.030)	(0.018)	(0.005)
R_{MIX}	-0.121^{***}	0.0001	0.032	0.090^{***}	-0.005
	(0.039)	(0.007)	(0.032)	(0.018)	(0.005)
R_{week}	0.035^{***}	0.004^{**}	-0.046^{***}	0.008^{*}	-0.001
	(0.010)	(0.002)	(0.008)	(0.005)	(0.001)
$R_{RTS,week}$	0.054^{**}	-0.005	-0.018	-0.027^{**}	-0.001
	(0.026)	(0.004)	(0.021)	(0.012)	(0.003)
$R_{MIX,week}$	-0.066^{*}	0.004	0.033	0.022	0.004
	(0.039)	(0.007)	(0.032)	(0.018)	(0.005)
Observations	9,771	9,771	9,771	9,771	9,771
R^2	0.022	0.007	0.011	0.033	0.059
Adjusted R^2	-0.002	-0.017	-0.013	0.010	0.035
F Statistic (df = 11; 9535)	19.312^{***}	6.160^{***}	9.637^{***}	29.904^{***}	53.903^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE 3.18: Drivers of customer order flow, daily time dimension. This table reports results for panel regressions of customer order flows (OF) on lagged customer order flow (CORP for corporate clients, DLR, dealres, RET, retail traders, INST, institutional investors, and non-residents, NER). The regressions also consider lagged returns as additional regressors (lagged futures returns and lagged market indices returns over the previous five trading days, R_{week}, $R_{RTS,week}$ and $R_{MIX,week}$, respectively.

	Dependent variable:				
	$Volume^{CORP}_{t+1}$	$Volume^{DLR}_{t+1}$	$Volume^{RET}_{t+1}$	$Volume^{INST}_{t+1}$	$Volume^{NER}_{t+1}$
	(1)	(2)	(3)	(4)	(5)
$Volume^{CORP}$	0.396***	−0.001***	0.096***	0.019***	−0.005***
	(0.001)	(0.0001)	(0.001)	(0.0002)	(0.0002)
$Volume^{DLR}$	−0.145***	0.370***	−0.064***	0.026***	0.006***
	(0.004)	(0.001)	(0.006)	(0.002)	(0.001)
$Volume^{RET}$	0.075***	0.001***	0.505***	0.013***	0.005***
	(0.0005)	(0.0001)	(0.001)	(0.0002)	(0.0001)
$Volume^{INST}$	0.108***	0.001***	−0.082***	0.401***	−0.004***
	(0.002)	(0.0002)	(0.002)	(0.001)	(0.0004)
$Volume^{NER}$	−0.045***	0.015***	0.035***	−0.014***	0.469***
	(0.002)	(0.0002)	(0.002)	(0.001)	(0.001)
R	−0.002***	0.00004	0.002***	0.0003**	0.0002**
	(0.0004)	(0.0001)	(0.001)	(0.0001)	(0.0001)
R_{RTS}	0.001	−0.0001	−0.0005	−0.0002	−0.0001
	(0.001)	(0.0001)	(0.001)	(0.0003)	(0.0002)
R_{MIX}	−0.002**	0.00005	−0.00004	0.001*	−0.0002
	(0.001)	(0.0001)	(0.001)	(0.0003)	(0.0003)
R_{day}	0.0001***	0.00000***	0.0002***	0.00002***	0.00000
	(0.00001)	(0.00000)	(0.00001)	(0.00000)	(0.00000)
$R_{RTS,day}$	−0.0005***	0.00001	−0.001***	−0.0002***	0.00001
	(0.00004)	(0.00000)	(0.00005)	(0.00001)	(0.00001)
$R_{MIX,day}$	0.0003***	−0.00002***	0.001***	0.0001***	0.00000
	(0.00004)	(0.00001)	(0.0001)	(0.00001)	(0.00001)
Observations	3,181,271	3,181,271	3,181,271	3,181,271	3,181,271
R^2	0.230	0.142	0.296	0.210	0.224
Adjusted R^2	0.230	0.142	0.296	0.210	0.224
F Statistic (df = 11; 3181035)	86,602.950***	47,904.400***	121,398.500***	76,788.380***	83,635.250***
Note:					*p<0.1; **p<0.05; ***p<0.01

TABLE 3.19: Drivers of customer buyer-initiated volume, 60-second time dimension. This table reports results for panel regressions of customer buyer-initiated volume (Volume) on lagged customer buyer-initiated volume (CORP for corporate clients, DLR, dealres, RET, retail traders, INST, institutional investors, and non-residents, NER). The regressions also consider lagged returns as additional regressors (lagged futures returns and lagged market indices returns over the previous trading day, R_{day}, $R_{RTS,day}$ and $R_{MIX,day}$, respectively.

	Dependent variable:				
	$Volume^{CORP}_{t+1}$	$Volume^{DLR}_{t+1}$	$Volume^{RET}_{t+1}$	$Volume^{INST}_{t+1}$	$Volume^{NER}_{t+1}$
	(1)	(2)	(3)	(4)	(5)
$Volume^{CORP}$	0.355***	−0.003***	0.036***	0.012***	−0.010***
	(0.002)	(0.0002)	(0.003)	(0.001)	(0.0005)
$Volume^{DLR}$	−0.286***	0.369***	−0.337***	−0.021***	0.046***
	(0.013)	(0.001)	(0.017)	(0.004)	(0.003)
$Volume^{RET}$	0.082***	−0.0001	0.553***	0.009***	0.003***
	(0.001)	(0.0001)	(0.002)	(0.0004)	(0.0003)
$Volume^{INST}$	−0.026***	−0.0001	−0.217***	0.377***	−0.007***
	(0.005)	(0.001)	(0.007)	(0.002)	(0.001)
$Volume^{NER}$	−0.114***	0.032***	−0.004	−0.023***	0.578***
	(0.005)	(0.001)	(0.007)	(0.002)	(0.001)
R	0.004***	−0.0001	−0.006***	−0.001**	0.00004
	(0.002)	(0.0002)	(0.002)	(0.001)	(0.0004)
R_{RTS}	−0.013***	0.0001	−0.007	−0.001	−0.0002
	(0.004)	(0.0004)	(0.005)	(0.001)	(0.001)
R_{MIX}	0.013**	−0.0002	0.014*	0.004**	0.001
	(0.006)	(0.001)	(0.007)	(0.002)	(0.001)
R_{day}	0.001***	0.00003***	0.001***	0.0002***	0.00001
	(0.0001)	(0.00001)	(0.0001)	(0.00003)	(0.00002)
$R_{RTS,day}$	−0.004***	−0.00000	−0.007***	−0.001***	−0.00002
	(0.001)	(0.0001)	(0.001)	(0.0002)	(0.0001)
$R_{MIX,day}$	0.003***	−0.0001	0.006***	0.001***	0.0001
	(0.001)	(0.0001)	(0.001)	(0.0002)	(0.0001)
Observations	496,893	496,893	496,893	496,893	496,893
R^2	0.195	0.165	0.299	0.177	0.343
Adjusted R^2	0.194	0.164	0.299	0.177	0.343
F Statistic (df = 11; 496657)	10,926.130***	8,912.509***	19,267.380***	9,736.351***	23,593.740***
Note:					*p<0.1; **p<0.05; ***p<0.01

TABLE 3.20: Drivers of customer buyer-initiated volume, 10-minute time dimension. This table reports results for panel regressions of customer buyer-initiated volume (Volume) on lagged customer buyer-initiated volume (CORP for corporate clients, DLR, dealres, RET, retail traders, INST, institutional investors, and non-residents, NER). The regressions also consider lagged returns as additional regressors (lagged futures returns and lagged market indices returns over the previous trading day, R_{day}, $R_{RTS,day}$ and $R_{MIX,day}$, respectively.

	Dependent variable:				
	$Volume^{CORP}_{t+1}$	$Volume^{DLR}_{t+1}$	$Volume^{RET}_{t+1}$	$Volume^{INST}_{t+1}$	$Volume^{NER}_{t+1}$
	(1)	(2)	(3)	(4)	(5)
$Volume^{CORP}$	0.444***	−0.021***	0.127***	0.064***	−0.043***
	(0.017)	(0.002)	(0.028)	(0.005)	(0.005)
$Volume^{DLR}$	−1.288***	0.294***	−2.179***	−0.334***	−0.023
	(0.105)	(0.011)	(0.181)	(0.030)	(0.033)
$Volume^{RET}$	−0.061***	0.007***	0.316***	−0.026***	0.015***
	(0.009)	(0.001)	(0.016)	(0.003)	(0.003)
$Volume^{INST}$	0.153***	−0.041***	−0.141**	0.213***	−0.068***
	(0.041)	(0.004)	(0.070)	(0.011)	(0.013)
$Volume^{NER}$	−0.172***	0.073***	0.350***	−0.033***	0.747***
	(0.034)	(0.004)	(0.059)	(0.010)	(0.011)
R	0.068***	0.003**	0.124***	0.014***	0.004
	(0.014)	(0.001)	(0.024)	(0.004)	(0.004)
R_{RTS}	−0.221***	−0.001	−0.546***	−0.108***	−0.004
	(0.074)	(0.008)	(0.127)	(0.021)	(0.023)
R_{MIX}	0.135*	−0.005	0.422***	0.087***	0.002
	(0.078)	(0.008)	(0.134)	(0.022)	(0.024)
R_{week}	0.076***	0.004*	0.012	0.013**	0.005
	(0.020)	(0.002)	(0.034)	(0.006)	(0.006)
$R_{RTS,week}$	−0.121**	0.001	−0.190**	−0.016	−0.003
	(0.052)	(0.005)	(0.090)	(0.015)	(0.016)
$R_{MIX,week}$	−0.037	−0.004	0.019	−0.014	0.002
	(0.077)	(0.008)	(0.133)	(0.022)	(0.024)
Observations	9,771	9,771	9,771	9,771	9,771
R^2	0.236	0.292	0.165	0.154	0.479
Adjusted R^2	0.217	0.275	0.144	0.133	0.466
F Statistic (df = 11; 9535)	267.934***	357.684***	170.904***	157.381***	795.785***
Note:					*p<0.1; **p<0.05; ***p<0.01

TABLE 3.21: Drivers of customer buyer-initiated volume, daily time dimension. This table reports results for panel regressions of customer buyer-initiated volume (Volume) on lagged customer buyer-initiated volume (CORP for corporate clients, DLR, dealres, RET, retail traders, INST, institutional investors, and non-residents, NER). The regressions also consider lagged returns as additional regressors (lagged futures returns and lagged market indices returns over the previous five trading days, R_{week}, $R_{RTS,week}$ and $R_{MIX,week}$, respectively.

	Dependent variable:				
	$Volume^{CORP}_{t+1}$	$Volume^{DLR}_{t+1}$	$Volume^{RET}_{t+1}$	$Volume^{INST}_{t+1}$	$Volume^{NER}_{t+1}$
	(1)	(2)	(3)	(4)	(5)
$Volume^{CORP}$	0.393***	−0.001***	0.090***	0.018***	−0.004***
	(0.001)	(0.0001)	(0.001)	(0.0002)	(0.0002)
$Volume^{DLR}$	−0.215***	0.326***	−0.274***	−0.013***	0.085***
	(0.007)	(0.001)	(0.008)	(0.002)	(0.002)
$Volume^{RET}$	0.073***	0.00003	0.503***	0.011***	0.004***
	(0.0005)	(0.00004)	(0.001)	(0.0002)	(0.0001)
$Volume^{INST}$	0.117***	−0.001***	−0.046***	0.439***	−0.007***
	(0.002)	(0.0001)	(0.002)	(0.001)	(0.0004)
$Volume^{NER}$	−0.043***	0.021***	0.035***	−0.012***	0.459***
	(0.002)	(0.0002)	(0.002)	(0.001)	(0.001)
R	0.003***	−0.00004	−0.002***	−0.0003**	−0.0004***
	(0.0004)	(0.00004)	(0.001)	(0.0001)	(0.0001)
R_{RTS}	0.001	0.00004	−0.0003	0.0004	0.0002
	(0.001)	(0.0001)	(0.001)	(0.0003)	(0.0002)
R_{MIX}	−0.002*	−0.0001	−0.0001	−0.001**	0.0003
	(0.001)	(0.0001)	(0.001)	(0.0003)	(0.0003)
R_{day}	0.0002***	0.00000***	0.0002***	0.00000*	−0.00000
	(0.00001)	(0.00000)	(0.00001)	(0.00000)	(0.00000)
$R_{RTS,day}$	−0.001***	0.00001**	−0.001***	0.00004***	0.00000
	(0.00004)	(0.00000)	(0.00005)	(0.00001)	(0.00001)
$R_{MIX,day}$	0.001***	−0.00002***	0.001***	−0.0001***	0.00001
	(0.00004)	(0.00000)	(0.0001)	(0.00001)	(0.00001)
Observations	3,181,271	3,181,271	3,181,271	3,181,271	3,181,271
R^2	0.225	0.124	0.294	0.238	0.222
Adjusted R^2	0.225	0.124	0.294	0.238	0.222
F Statistic (df = 11; 3181035)	84,107.290***	40,913.760***	120,541.900***	90,275.770***	82,570.610***
Note:					*p<0.1; **p<0.05; ***p<0.01

TABLE 3.22: Drivers of customer seller-initiated volume, 10-minute time dimension. This table reports results for panel regressions of customer seller-initiated volume (Volume) on lagged customer seller-initiated volume (CORP for corporate clients, DLR, dealres, RET, retail traders, INST, institutional investors, and non-residents, NER). The regressions also consider lagged returns as additional regressors (lagged futures returns and lagged market indices returns over the previous trading day, R_{day}, $R_{RTS,day}$ and $R_{MIX,day}$, respectively.

	Dependent variable:				
	$Volume^{CORP}_{t+1}$	$Volume^{DLR}_{t+1}$	$Volume^{RET}_{t+1}$	$Volume^{INST}_{t+1}$	$Volume^{NER}_{t+1}$
	(1)	(2)	(3)	(4)	(5)
$Volume^{CORP}$	0.370***	−0.002***	0.024***	0.007***	−0.010***
	(0.002)	(0.0001)	(0.003)	(0.001)	(0.0005)
$Volume^{DLR}$	−0.540***	0.544***	−0.404***	−0.099***	0.077***
	(0.020)	(0.001)	(0.026)	(0.007)	(0.005)
$Volume^{RET}$	0.074***	0.0005***	0.565***	0.009***	0.003***
	(0.001)	(0.0001)	(0.002)	(0.0004)	(0.0003)
$Volume^{INST}$	−0.055***	−0.004***	−0.215***	0.397***	−0.006***
	(0.005)	(0.0003)	(0.006)	(0.002)	(0.001)
$Volume^{NER}$	−0.079***	0.010***	−0.003	−0.015***	0.564***
	(0.005)	(0.0003)	(0.007)	(0.002)	(0.001)
R	−0.002	−0.00005	0.0001	0.001**	0.0001
	(0.002)	(0.0001)	(0.002)	(0.001)	(0.0004)
R_{RTS}	−0.003	−0.0002	−0.011**	−0.007***	−0.001
	(0.004)	(0.0003)	(0.005)	(0.001)	(0.001)
R_{MIX}	0.008	0.0003	0.015**	0.007***	0.0005
	(0.006)	(0.0004)	(0.007)	(0.002)	(0.001)
R_{day}	0.001***	0.00001**	0.001***	0.0001**	0.00001
	(0.0001)	(0.00001)	(0.0001)	(0.00003)	(0.00003)
$R_{RTS,day}$	−0.006***	0.00001	−0.006***	−0.0001	−0.00005
	(0.001)	(0.00003)	(0.001)	(0.0002)	(0.0001)
$R_{MIX,day}$	0.004***	−0.00004	0.005***	−0.0001	0.0001
	(0.001)	(0.00004)	(0.001)	(0.0002)	(0.0001)
Observations	496,893	496,893	496,893	496,893	496,893
R^2	0.194	0.316	0.304	0.186	0.330
Adjusted R^2	0.194	0.315	0.304	0.186	0.330
F Statistic (df = 11; 496657)	10,871.890***	20,836.140***	19,759.610***	10,350.040***	22,248.370***
Note:					*p<0.1; **p<0.05; ***p<0.01

TABLE 3.23: Drivers of customer seller-initiated volume, 10-minute time dimension. This table reports results for panel regressions of customer seller-initiated volume (Volume) on lagged customer seller-initiated volume (CORP for corporate clients, DLR, dealres, RET, retail traders, INST, institutional investors, and non-residents, NER). The regressions also consider lagged returns as additional regressors (lagged futures returns and lagged market indices returns over the previous trading day, R_{day}, $R_{RTS,day}$ and $R_{MIX,day}$, respectively.

	Dependent variable:				
	$Volume^{CORP}_{t+1}$	$Volume^{DLR}_{t+1}$	$Volume^{RET}_{t+1}$	$Volume^{INST}_{t+1}$	$Volume^{NER}_{t+1}$
	(1)	(2)	(3)	(4)	(5)
$Volume^{CORP}$	0.614***	−0.019***	0.504***	0.028***	−0.046***
	(0.016)	(0.001)	(0.028)	(0.004)	(0.005)
$Volume^{DLR}$	−1.617***	0.476***	−1.161***	−0.609***	−0.093*
	(0.153)	(0.012)	(0.268)	(0.044)	(0.049)
$Volume^{RET}$	−0.128***	0.007***	0.145***	−0.016***	0.015***
	(0.009)	(0.001)	(0.016)	(0.003)	(0.003)
$Volume^{INST}$	−0.157***	−0.016***	−0.546***	0.149***	−0.045***
	(0.041)	(0.003)	(0.072)	(0.012)	(0.013)
$Volume^{NER}$	−0.034	0.050***	0.404***	−0.026**	0.737***
	(0.037)	(0.003)	(0.065)	(0.011)	(0.012)
R	0.098***	0.001	0.116***	0.009**	0.001
	(0.014)	(0.001)	(0.024)	(0.004)	(0.004)
R_{RTS}	−0.402***	−0.0002	−0.554***	0.018	−0.004
	(0.073)	(0.006)	(0.128)	(0.021)	(0.023)
R_{MIX}	0.295***	−0.004	0.439***	−0.037*	0.004
	(0.077)	(0.006)	(0.134)	(0.022)	(0.024)
R_{week}	0.039**	−0.0004	0.015	0.018***	0.003
	(0.020)	(0.002)	(0.034)	(0.006)	(0.006)
$R_{RTS,week}$	−0.097*	0.005	−0.127	−0.008	−0.005
	(0.052)	(0.004)	(0.090)	(0.015)	(0.016)
$R_{MIX,week}$	−0.033	−0.007	−0.018	−0.031	0.004
	(0.076)	(0.006)	(0.133)	(0.022)	(0.024)
Observations	9,771	9,771	9,771	9,771	9,771
R^2	0.278	0.414	0.162	0.109	0.456
Adjusted R^2	0.260	0.399	0.142	0.087	0.443
F Statistic					
(df = 11; 9535)	334.167***	611.349***	168.027***	105.902***	727.737***
Note:					*p<0.1; **p<0.05; ***p<0.01

TABLE 3.24: Drivers of customer seller-initiated volume, daily time dimension. This table reports results for panel regressions of customer seller-initiated volume (Volume) on lagged customer seller-initiated volume (CORP for corporate clients, DLR, dealres, RET, retail traders, INST, institutional investors, and non-residents, NER). The regressions also consider lagged returns as additional regressors (lagged futures returns and lagged market indices returns over the previous five trading days, R_{week}, $R_{RTS,week}$ and $R_{MIX,week}$, respectively.

3.8.6 Appendix 6. Drivers of flows for different groups of futures

	Panel A: Group 1					Panel B: Group 2				
	Dependent variable: Customer order flow					*Dependent variable: Customer order flow*				
	OF^{CORP}_{t+1}	OF^{DLR}_{t+1}	OF^{RET}_{t+1}	OF^{INST}_{t+1}	OF^{NER}_{t+1}	OF^{CORP}_{t+1}	OF^{DLR}_{t+1}	OF^{RET}_{t+1}	OF^{INST}_{t+1}	OF^{NER}_{t+1}
	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)	(2.1)	(2.2)	(2.3)	(2.4)	(2.5)
OF^{CORP}	0.297***	0.005***	0.029***	−0.023***	−0.005***	0.018	0.270***	−0.071***	−0.071***	−0.018***
	(0.007)	(0.001)	(0.005)	(0.003)	(0.002)	(0.011)	(0.001)	(0.012)	(0.006)	(0.005)
OF^{DLR}	−0.130***	0.295***	−0.002	0.087***	0.075***	0.018	0.270***	−0.071***	−0.071***	−0.018***
	(0.013)	(0.001)	(0.011)	(0.006)	(0.004)	(0.011)	(0.001)	(0.012)	(0.006)	(0.005)
OF^{RET}	0.042***	0.004***	0.198***	0.068***	−0.009***	0.055***	0.0001	0.086***	−0.025***	0.005**
	(0.006)	(0.001)	(0.005)	(0.003)	(0.002)	(0.006)	(0.001)	(0.007)	(0.003)	(0.002)
OF^{INST}	−0.115***	0.006***	0.088***	0.340***	−0.009***	−0.052***	−0.001	−0.036***	0.198***	0.008***
	(0.007)	(0.001)	(0.006)	(0.003)	(0.002)	(0.007)	(0.001)	(0.007)	(0.003)	(0.003)
OF^{NER}	0.043***	0.038***	0.023***	0.049***	0.147***	0.028***	0.002**	−0.032***	−0.018***	0.145***
	(0.008)	(0.001)	(0.007)	(0.004)	(0.002)	(0.007)	(0.001)	(0.007)	(0.003)	(0.003)
R	0.010**	0.0003	−0.035***	0.015***	0.005***	−0.004***	−0.00003	0.001	0.004***	−0.001***
	(0.005)	(0.001)	(0.004)	(0.003)	(0.001)	(0.0005)	(0.0001)	(0.001)	(0.0002)	(0.0002)
R_{RTS}	0.009***	−0.0001	−0.004	−0.005***	−0.001	−0.0002	−0.00003	−0.0004	0.001***	0.0001
	(0.003)	(0.0004)	(0.003)	(0.002)	(0.001)	(0.0004)	(0.00004)	(0.0004)	(0.0002)	(0.0001)
R_{MIX}	−0.028***	0.0004	0.022***	0.009***	−0.002*	0.003***	0.00004	−0.0003	−0.002***	−0.0002
	(0.004)	(0.0004)	(0.003)	(0.002)	(0.001)	(0.0005)	(0.0001)	(0.001)	(0.0002)	(0.0002)
R_{day}	−0.001***	0.0001***	0.0004**	0.001***	−0.00001	0.00004*	0.00000	−0.0001***	−0.00000	0.00003***
	(0.0002)	(0.00002)	(0.0001)	(0.0001)	(0.00005)	(0.00002)	(0.00000)	(0.00002)	(0.00001)	(0.00001)
$R_{RTS,day}$	0.001***	−0.00002	−0.0003*	−0.001***	−0.00001	−0.00000	0.00000*	−0.00000	0.00001	0.00000
	(0.0002)	(0.00002)	(0.0001)	(0.0001)	(0.00005)	(0.00002)	(0.00000)	(0.00002)	(0.00001)	(0.00001)
$R_{MIX,day}$	−0.001***	0.00001	0.0002	0.001***	0.00001	−0.00004*	−0.00001***	0.0001***	−0.00001	−0.00004***
	(0.0002)	(0.00002)	(0.0002)	(0.0001)	(0.0001)	(0.00002)	(0.00000)	(0.00002)	(0.00001)	(0.00001)
Obs.	572,508	572,508	572,508	572,508	572,508	605,922	605,922	605,922	605,922	605,922
R^2	0.103	0.091	0.035	0.169	0.026	0.016	0.073	0.010	0.054	0.021
Adj. R^2	0.103	0.091	0.035	0.169	0.026	0.016	0.073	0.010	0.054	0.021
	Panel C: Group 3					Panel D: Group 4				
	(3.1)	(3.2)	(3.3)	(3.4)	(3.5)	(4.1)	(4.2)	(4.3)	(4.4)	(4.5)
OF^{CORP}	0.266***	−0.0004	0.141***	−0.008***	−0.050***	0.112***	0.023***	0.205***	−0.021***	−0.052***
	(0.012)	(0.002)	(0.012)	(0.002)	(0.003)	(0.008)	(0.001)	(0.007)	(0.004)	(0.002)
OF^{DLR}	−0.191***	0.395***	0.206***	−0.014***	−0.046***	−0.032***	0.275***	0.091***	0.012**	−0.068***
	(0.013)	(0.002)	(0.013)	(0.003)	(0.003)	(0.010)	(0.001)	(0.008)	(0.005)	(0.003)
OF^{RET}	0.008	0.002	0.388***	0.001	−0.051***	0.014*	0.022***	0.253***	0.021***	−0.044***
	(0.012)	(0.002)	(0.012)	(0.002)	(0.003)	(0.008)	(0.001)	(0.007)	(0.004)	(0.002)
OF^{INST}	−0.023*	0.013***	0.264***	0.144***	−0.050***	−0.145***	0.021***	0.239***	0.198***	−0.047***
	(0.013)	(0.002)	(0.013)	(0.003)	(0.003)	(0.009)	(0.001)	(0.007)	(0.004)	(0.003)
OF^{NER}	0.008	0.007***	0.271***	0.001	0.072***	−0.040***	0.040***	0.142***	0.010**	0.105***
	(0.014)	(0.002)	(0.014)	(0.003)	(0.003)	(0.009)	(0.001)	(0.008)	(0.005)	(0.003)
R	−0.017***	0.00000	0.014***	0.002***	0.001***	−0.00001	0.00005**	−0.001***	0.001***	−0.00003
	(0.001)	(0.0001)	(0.001)	(0.0002)	(0.0002)	(0.0001)	(0.00002)	(0.0001)	(0.0001)	(0.00004)
R_{RTS}	−0.0003	0.0001	0.001	−0.001	−0.0002	0.002***	−0.00004	−0.001***	−0.001***	−0.0004***
	(0.002)	(0.0003)	(0.002)	(0.0004)	(0.0004)	(0.0003)	(0.00004)	(0.0002)	(0.0001)	(0.0001)
R_{MIX}	0.004	−0.0002	−0.004*	0.001*	0.0002	0.002***	0.0001	−0.002***	0.0001	−0.00002
	(0.003)	(0.0004)	(0.003)	(0.0005)	(0.001)	(0.0003)	(0.0001)	(0.0003)	(0.0002)	(0.0001)
R_{day}	0.00001	0.00000	−0.00001	−0.00000	0.00000	−0.00003***	0.00000	0.00000	0.00002***	0.00000***
	(0.00001)	(0.00000)	(0.00001)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
$R_{RTS,day}$	0.00000	0.00003	0.00004	−0.0001***	−0.00000	0.00000	0.00001**	0.00000	−0.00001*	0.00001
	(0.0001)	(0.00002)	(0.0001)	(0.00002)	(0.00002)	(0.00001)	(0.00000)	(0.00001)	(0.00001)	(0.00000)
$R_{MIX,day}$	−0.00003	−0.00002	−0.00002	0.0001***	0.00000	−0.00000	−0.00001***	0.00000	0.00001*	−0.00001**
	(0.0001)	(0.00002)	(0.0001)	(0.00002)	(0.00002)	(0.00002)	(0.00000)	(0.00001)	(0.00001)	(0.00000)
Obs.	679,387	679,387	679,387	679,387	679,387	1,323,451	1,323,451	1,323,451	1,323,451	1,323,451
R^2	0.074	0.153	0.061	0.024	0.017	0.027	0.066	0.006	0.056	0.025
Adj. R^2	0.074	0.153	0.060	0.024	0.017	0.027	0.066	0.006	0.056	0.024
Note:										*p<0.1; **p<0.05; ***p<0.01

TABLE 3.25: Drivers of customer order flow for different groups of futures: panel regressions, 60 seconds. This table reports results for panel regressions of customer order flows (OF) on lagged customer order flows pulled across future contracts groupped by asset classes for the period 21 months. The regressions also consider lagged future returns (R_{day}) and lagged market indices returns over the previous trading day ($R_{RTS,day}$ and $R_{MIX,day}$). Group 1: futures on currency exchange; Group 2: futures on market indices; Group 3: commodity futures; Group 4: stock futures.

3.8.7 Appendix 7. Non-overlapping correlation of the order flows

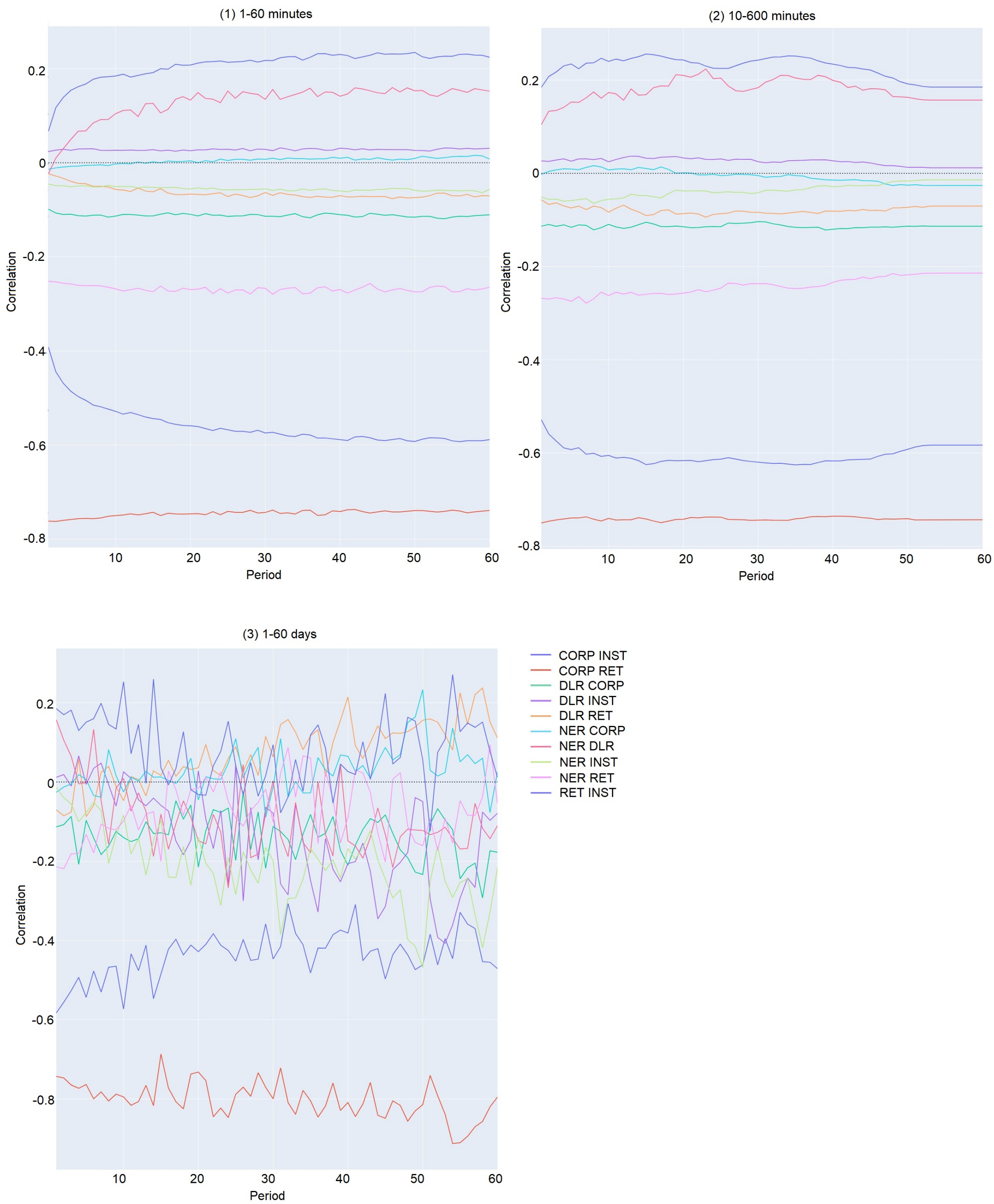


FIGURE 3.26: Correlation of customers' order flows over long horizon. The figures plot contemporaneous Pearson correlations between standardised order flows of different investor groups for horizons up to 60 periods, where one period is (1) one minute, (2) ten minutes, and (3) one day; periods do not overlap.

3.8.8 Appendix 8. Cumulative post-formation portfolio returns

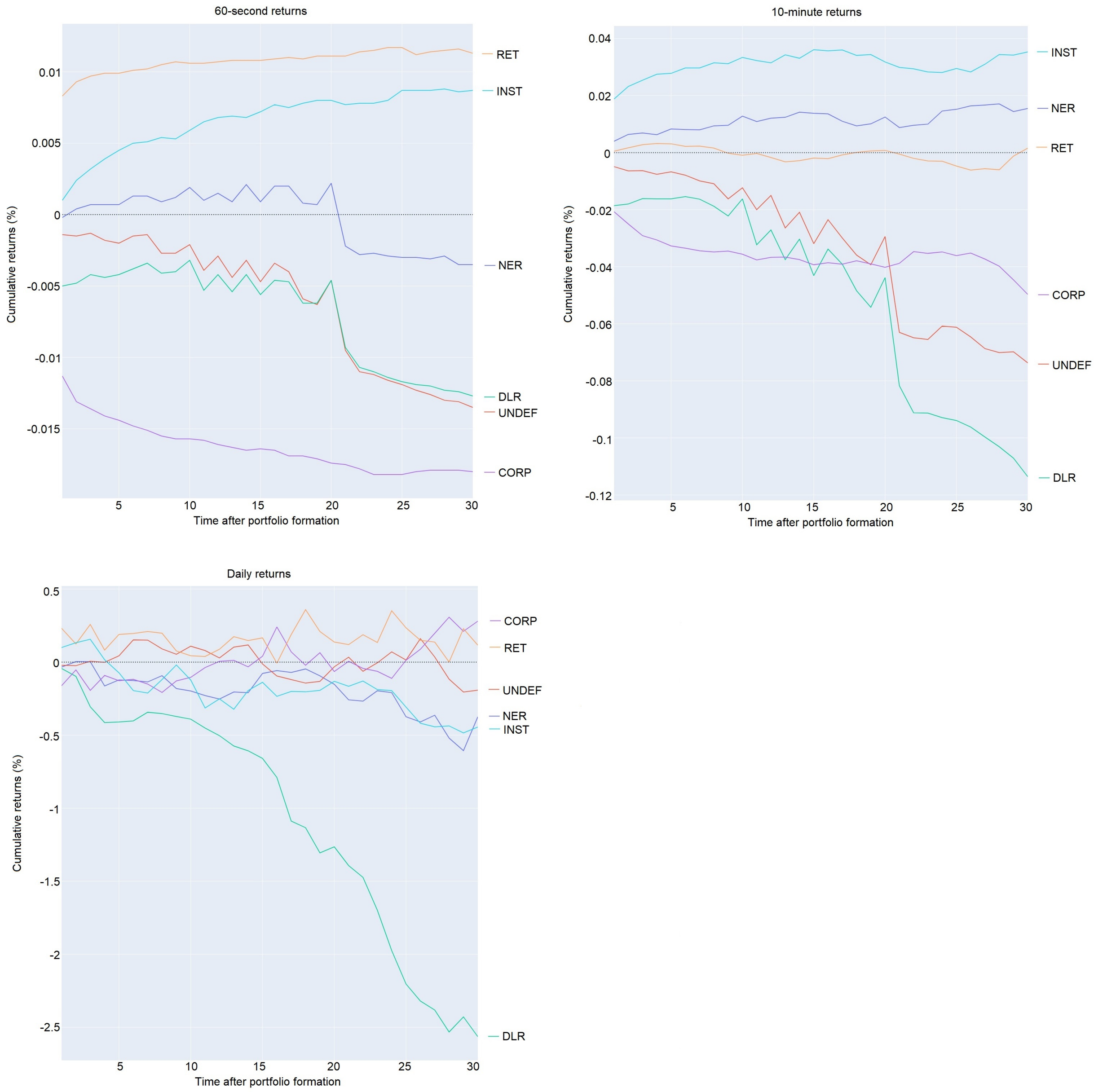


FIGURE 3.27: Cumulative post-formation portfolio returns (P_1 less P_7). This figure shows average cumulative returns for the position of long P_1 portfolio and short P_7 portfolio based on disaggregated order flows over the first 30 time windows after the portfolio formation, where time window of 60-second, 10-minutes and one day; periods overlap. We form 27 futures into seven portfolios.

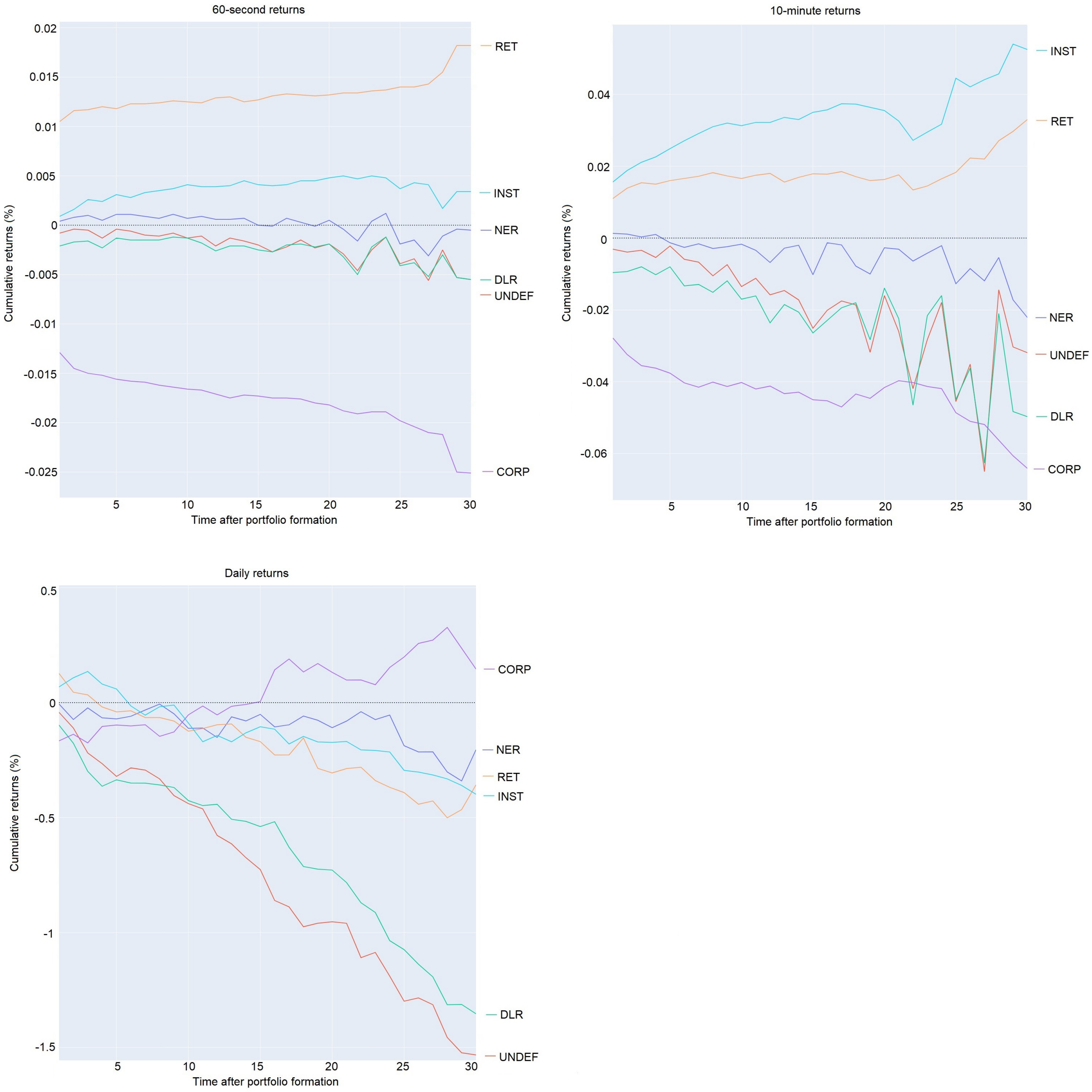


FIGURE 3.28: Cumulative post-formation portfolio returns (P_1 less P_3). This figure shows average cumulative returns for the position of long P_1 portfolio and short P_3 portfolio based on disaggregated order flows over the first 30 time windows after the portfolio formation, where time window of 60-second, 10-minutes and one day; periods overlap. We form 27 futures into three portfolios.

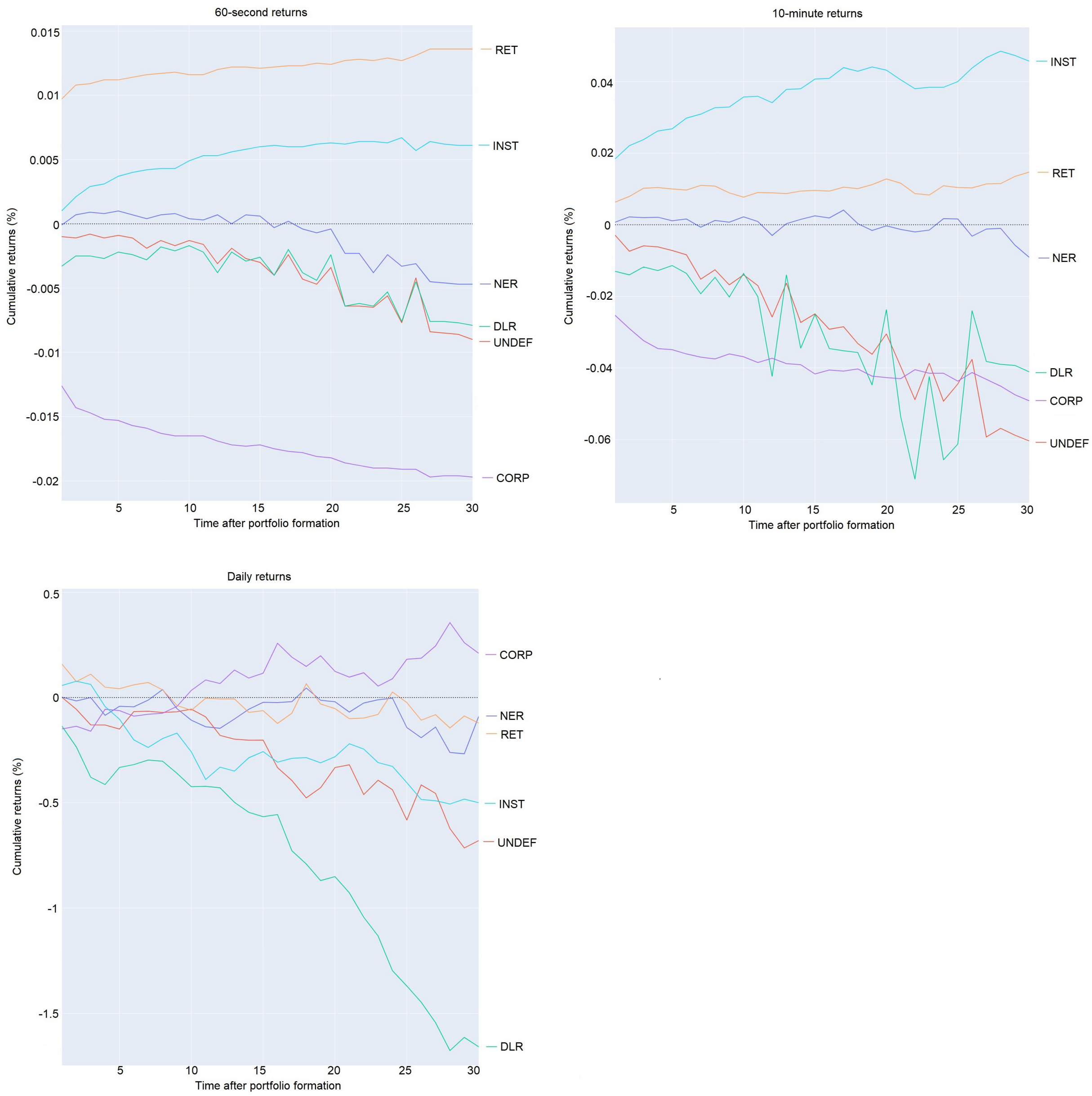


FIGURE 3.29: Cumulative post-formation portfolio returns (P_1 less P_5). This figure shows average cumulative returns for the position of long P_1 portfolio and short P_5 portfolio based on disaggregated order flows over the first 30 time windows after the portfolio formation, where time window of 60-second, 10-minutes and one day; periods overlap. We form 27 futures into five portfolios.

Chapter 4

Price Discovery between Bitcoin Spot Markets and Exchange Traded Products

This chapter has been published as

Gemayel, R., Franus, T., Bowden, J. (2023). Price discovery between Bitcoin spot markets and exchange traded products. *Economics Letters, 228*, 111152.

4.1 Introduction

The process through which new information is efficiently incorporated into asset prices is less clear when trading an asset is fragmented across multiple venues or markets. In such a scenario, it is of interest to identify where price discovery takes place (Hasbrouck 1995).

Crypto spot exchanges have attracted significant interest from both retail and institutional investors. As regulations constrained the ability of traditional funds and banks to participate in these exchanges, an opportunity arose to create a more traditional product allowing exposure to Bitcoin and other cryptocurrencies. Thus, Bitcoin Exchange-Traded Products (ETPs) allow investors on traditional equity exchanges to gain exposure to the underlying asset without the need to hold Bitcoin.

Evidence suggests that these products have witnessed significant fund flows, with over 180 active crypto ETFs, ETPs, and trusts in existence. Approximately half of these have been launched since late 2021, during which time the total value of underlying crypto assets dropped by 70%, from $84 billion to $24 billion[1]. With traditional investors and institutions now able to access crypto markets, we aim to examine the extent to which the ETP market offers a venue for Bitcoin price discovery.

Previous literature has mainly focused on the lead-lag relationship between futures and spot markets, with the overarching hypothesis that price discovery predominantly occurs in futures markets. Studies have presented evidence in support of this across markets including equities (Kawaller et al. 1987, Chan 1992, Wahab & Lashgari 1993, Koutmos & Tucker 1996, Booth et al. 1999, Tse 1999, Hasbrouck 2003, Covrig et al. 2004, So & Tse 2004, Bohl et al. 2011, Theissen 2012, Yang et al. 2012, Ahn et al. 2019, Fassas & Siriopoulos 2019), commodities (Kuiper et al. 2002, Peri et al. 2013, Dolatabadi et al. 2015, Hauptfleisch et al. 2016, Dimpfl et al. 2017), and foreign exchange (Chen & Gau 2010).

Another branch of literature has investigated whether equity exchange-traded funds (ETFs) enhance price discovery in the underlying securities. The evidence presented is mixed. On the one hand, prior studies such as Lettau & Madhavan (2018), Madhavan (2016), and Madhavan & Sobczyk (2016) indicate that ETFs offer a supplementary layer of liquidity on top of the underlying securities, which can improve price discovery in the latter. This is because ETFs are a cost-effective tool for investors to make directional bets on the index, consequently reflecting new information before the underlying securities. This hypothesis is corroborated by several empirical studies (Richie et al. 2008, Marshall et al. 2013, Glosten et al. 2021). On the other hand, several studies have presented evidence showing that non-fundamental trades in the ETF may propagate to the underlying securities, causing mispricing and degrading informational efficiency (Broman 2016, Israeli et al. 2017, Da & Shive 2018, Brown et al. 2021).

In the cryptocurrency space, studies have largely focused on price discovery in Bitcoin markets. One branch of literature has investigated Bitcoin price dynamics within spot markets to determine which exchanges (Brandvold et al. 2015) and factors (Balcilar et al. 2017, Jang & Lee 2017, Brauneis & Mestel 2018, Beneki et al. 2019) explain price dynamics. For instance, Brandvold et al. (2015) found that the Mt.Gox and BTC-e exchanges are market leaders with the highest information share, and that the latter changes significantly over time. Others including Balcilar et al. (2017),

[1]See https://www.coindesk.com (last accessed February 26, 2023)

Jang & Lee (2017), Brauneis & Mestel (2018), and Beneki et al. (2019) have focused on the impact of volume, liquidity, and volatility on return dynamics.

A second branch of literature has examined price discovery between Bitcoin spot and futures markets, making use of four popular cross-market metrics: Information Share (IS) (Hasbrouck 1995), Component Share (CS) (Gonzalo & Granger 1995), Information Leadership (IL) (Yan & Zivot 2010), and the Information Leadership Share (ILS) (Putniņš 2013). These studies have produced mixed results. On the one hand, Corbet et al. (2018) apply the above-mentioned measures to one-minute CME, CBOE, and spot market data and find that price discovery is focused on the spot market. Similar evidence is found by Baur & Dimpfl (2019) using five-minute sampled data. On the other hand, Kapar & Olmo (2019) use daily-sampled data and find that the CME futures market dominates price discovery. Similarly, Fassas et al. (2020), Akyildirim et al. (2020), and Alexander et al. (2020) show that Bitcoin futures play a leading role in price discovery.

According to Hu et al. (2020), a key reason for these mixed findings is that cointegration relationships may go undetected if the underlying model formulation is constrained to be time-invariant. By applying time-varying cointegrating coefficients (Park & Hahn 1999, Shi et al. 2018), the authors conclude that futures prices Granger-cause spot prices.

Our study contributes to the literature on price discovery in Bitcoin markets as – to the best of our knowledge – it is the first to empirically examine the price dynamics of Bitcoin ETPs in relation to spot markets. We apply four popular measures of price discovery to Bitcoin ETP and spot exchange data and show that spot markets dominate the price discovery process, suggesting that ETPs tend to lag in terms of informational efficiency.

The remainder of this paper is organised as follows. Section 4.2 presents the data. Section 4.3 describes the price discovery metrics. Section 4.4 discusses the results. Finally, Section 4.5 concludes our analysis.

4.2 Data

We use two data sources that span from August 2021 to July 2022. The first, CryptoCompare, provides spot transaction data on leading centralised exchanges. We obtain data on the BTC/USD and BTC/USDT[2] markets from the ten leading exchanges by volume: Binance, Bitfinex, Bitstamp, Coinbase, Gemini, Huobi, itBit, Kraken, Kucoin, and OKX. Descriptive statistics for the exchanges using daily sampled data are presented in Table 4.1. Moreover, we present the illiquidity measure of Amihud (2002), which is calculated as

$$AI_i = \frac{1}{T}\sum_{t=1}^{T}\frac{|R_{i,t}|}{V_{i,t}} \tag{4.1}$$

where T is the number of days in the period of analysis, $|R_{i,t}|$ is the absolute daily return in percentage of asset i, and $V_{i,t}$ is the volume in millions of notional of the quote currency. The larger the value of AI, the greater the degree of illiquidity of the asset.

[2]Some exchanges offer a BTC/USD market while others offer BTC/USDT. To account for potential movements in USDT/USD, we convert all BTC/USDT markets to BTC/USD using a market aggregate USDT/USD rate. The results for all analyses discussed in the paper remain the same, thus we do not report them due to spatial limitations.

Exchanges are ranked by decreasing mean daily volume traded. Generally, the mean and standard deviation of returns are found to be the same across exchanges at -0.12% and 3.6%, respectively. In addition, the Bitcoin market on Binance is found to be the most liquid across exchanges according to the AI measure.

We use Bloomberg to obtain transaction data for the most popular Bitcoin ETPs issued by 21Shares, Coinshares, ETC Group, Iconic Funds, SEBA Bank AG, and VanEck. These ETPs trade on several stock exchanges located in the Eurozone. While the price dynamics of an ETP issued by a particular issuer can differ across exchanges due to varying market activity on these exchanges, we focus on the exchanges with the most volume traded for each ETP. We present descriptive statistics for the ETPs using daily data in Table 4.2.

The majority of Bitcoin ETPs have a mean daily return of -0.09%, with the exception of SBTCU (-0.14%), and ABTC (0.06%) due to variations in launch date. The AI measure indicates that BTCE is the most liquid Bitcoin ETP. Nonetheless, the AI value for BTCE (0.0674) is around 56 times larger than that of Binance (0.0012), which suggests that even the most liquid Bitcoin ETP is approximately 56 times less liquid than the most liquid spot exchange. In unreported results, we estimate the AI for the SPDR equity ETF and individual stocks, including TSLA, AMZN, MSFT, and AAPL, to be between 0.0001 and 0.00078. This similarly highlights that the most liquid Bitcoin ETP is, on average, around 250 times less liquid than the largest tech stocks. To put things further into perspective, the aggregate market capitalization of the Bitcoin ETPs in our sample reached a maximum of around \$3.2 billion. This is significantly lower than the market cap of large tech stocks, including Apple Inc. (AAPL) and Microsoft Corporation (MSFT), which have consistently been valued at over \$1 trillion during the period of analysis.

4.3 Methodology

There are two price discovery measures, which assume a common implicit efficient price that can be estimated using a vector error correction model (VECM). The Information Share (IS) (Hasbrouck 1995), estimates the proportion of the efficient price innovation variance explained by innovations stemming from different markets. Alternatively, the Component Share (CS) approach (Booth et al. 1999, Chu et al. 1999, Harris et al. 2002) adopts the permanent-transitory decomposition technique in Gonzalo & Granger (1995). Specifically, the permanent component represents the common efficient price, while the temporary component reflects deviations from the efficient price caused by trading fractions. Despite their disparate focus points, both measures adopt cointegration to constrain multiple price series to share a common efficient price.

Consider an asset that is trading on two venues, where $p_{i,t}$ denotes the log price of the asset on venue i at time t. We assume that the two price series are closely linked due to arbitrage and that they contain a random-walk element rendering them non-stationary. Following Hauptfleisch et al. (2016) and Corbet et al. (2018), we write the VECM representation for the two venues as

$$\Delta p_{1,t} = \alpha_1(p_{1,t-1} - p_{2,t-1}) + \sum_{i=1}^{200} \gamma_i \Delta p_{1,t-i} + \sum_{j=1}^{200} \delta_j \Delta p_{2,t-j} + \varepsilon_{1,t} \tag{4.2}$$

$$\Delta p_{2,t} = \alpha_2(p_{1,t-1} - p_{2,t-1}) + \sum_{k=1}^{200} \varphi_k \Delta p_{1,t-k} + \sum_{m=1}^{200} \phi_m \Delta p_{2,t-m} + \varepsilon_{2,t} \tag{4.3}$$

Exchange	Base	Quote	Param	Mean	StDev	Min	Median	Max
Binance	BTC	USDT	price	41,871	11,969	18,970	42,380	67,525
			return	-0.12	3.6	-15.38	-0.03	14.49
			volume	2,279,970,333	1,114,931,810	615,314,010	2,057,841,593	8,776,020,939
			AI	0.0012				
Coinbase	BTC	USD	price	41,880	11,984	18,948	42,415	67,554
			return	-0.12	3.6	-15.42	-0.01	14.52
			volume	673,327,112	337,238,581	163,733,089	617,922,091	2,087,246,485
			AI	0.0040				
Huobi	BTC	USDT	price	41,870	11,968	18,972	42,380	67,514
			return	-0.12	3.6	-15.41	-0.04	14.51
			volume	555,473,649	293,159,527	137,115,072	500,812,178	2,504,865,354
			AI	0.0049				
OKX	BTC	USDT	price	41,872	11,969	18,971	42,380	67,525
			return	-0.12	3.6	-15.4	-0.03	14.52
			volume	527,982,434	325,255,083	63,741,448	450,255,151	1,897,742,838
			AI	0.0060				
Kucoin	BTC	USDT	price	41,870	11,968	18,979	42,394	67,509
			return	-0.12	3.6	-15.4	-0.04	14.51
			volume	393,177,711	194,085,241	106,053,933	371,383,549	1,679,772,958
			AI	0.0070				
Bitfinex	BTC	USD	price	41,888	11,976	18,965	42,418	67,526
			return	-0.12	3.59	-15.53	-0.02	14.49
			volume	232,810,308	168,573,656	32,082,332	185,490,745	1,085,086,080
			AI	0.0131				
Kraken	BTC	USD	price	41,881	11,985	18,950	42,419	67,559
			return	-0.12	3.6	-15.47	-0.02	14.55
			volume	134,371,690	78,327,255	23,260,223	116,053,303	447,377,859
			AI	0.0204				
Bitstamp	BTC	USD	price	41,886	11,986	18,956	42,420	67,559
			return	-0.12	3.61	-15.55	-0.01	14.49
			volume	97,023,649	68,360,816	14,036,016	80,542,267	479,685,055
			AI	0.0309				
Gemini	BTC	USD	price	41,884	11,986	18,948	42,415	67,552
			return	-0.12	3.6	-15.4	-0.03	14.54
			volume	62,068,086	43,104,267	10,112,372	50,670,616	283,009,135
			AI	0.0474				
Bitfinex	BTC	USDT	price	41,872	11,969	18,979	42,377	67,517
			return	-0.12	3.6	-15.48	-0.01	14.61
			volume	50,601,341	38,119,450	2,124,543	40,282,227	223,731,253
			AI	0.0783				
Coinbase	BTC	USDT	price	41,872	11,970	18,977	42,379	67,530
			return	-0.12	3.59	-15.41	-0.03	14.59
			volume	27,309,774	15,056,375	2,316,301	24,842,512	111,971,110
			AI	0.1065				
itBit	BTC	USD	price	41,884	11,985	18,948	42,412	67,554
			return	-0.12	3.6	-15.49	-0.04	14.53
			volume	13,261,508	10,081,651	1,584,938	10,296,224	70,668,139
			AI	0.2455				
Kraken	BTC	USDT	price	41,874	11,971	19,001	42,379	67,512
			return	-0.12	3.6	-15.54	-0.08	14.52
			volume	12,958,329	8,882,283	1,440,347	10,785,588	62,099,533
			AI	0.2334				
Bitstamp	BTC	USDT	price	41,894	11,958	19,022	42,562	67,634
			return	-0.12	3.61	-16.11	0.0	14.39
			volume	750,717	1,096,083	1,936	446,162	11,475,061
			AI	8.6200				

TABLE 4.1: Descriptive Statistics of Bitcoin Spot Exchanges. This table presents descriptive statistics for daily Bitcoin prices, returns, and volumes over the period August 2021 to July 2022. We report the mean (**Mean**), standard deviation (**StDev**), minimum (**Min**), median (**Median**), and maximum (**Max**) values. Moreover, we report the Amihud illiquidity measure (AI) where the volume parameter in the denominator is in the millions of notional of the quote currency.

Issuer	Ticker	Exchange	Quote	Param	Mean	StDev	Min	Median	Max
ETC Group	BTCE	Xetra	EUR	price	36.33	9.55	17.52	36.65	57.76
				return	-0.09	3.43	-20.41	0	9.92
				volume	2,865,877	11,214,308	0	908,034	152,048,456
				market cap	785,922,300	345,048,300	310,831,400	720,241,400	1,667,034,000
				AI	0.0674				
VanEck	VBTC	Xetra	EUR	price	20.68	5.4	10.02	20.88	32.82
				return	-0.09	3.43	-20.04	0	10.21
				volume	1,856,566	5,062,089	0	423,666	44,544,488
				market cap	209,036,200	56,614,410	98,231,690	211,778,500	335,896,700
				AI	0.2354				
SEBA Bank AG	SBTCU	SIX	USD	price	4.16	1.2	1.85	4.24	6.75
				return	-0.14	3.53	-19.88	0	10.58
				volume	681,313	3,288,609	0	20,000	35,076,564
				market cap	73,278,340	15,955,170	41,906,200	73,732,000	110,194,000
				AI	36.5259				
21Shares	ABTC	SIX	USD	price	15.01	4.44	6.73	15.25	24.5
				return	0.06	7.08	-36.06	0	46.25
				volume	605,740	1,872,248	0	152,574	23,237,304
				market cap	314,024,400	97,922,430	156,472,000	303,231,000	577,426,000
				AI	1.9264				
Iconic Funds	XBTI	SIX	CHF	price	3.73	0.97	1.81	3.76	5.92
				return	-0.09	3.45	-20.51	0	10.14
				volume	245,743	1,235,972	0	30,594	21,164,432
				market cap	6,427,968	1,825,737	3,428,631	6,617,954	11,593,840
				AI	182.8244				
Coinshares	BITC	SIX	USD	price	41.71	12.01	18.56	42.6	68.07
				return	-0.09	4.58	-32.19	0	35.86
				volume	26,485	125,588	0	2,742	1,817,852
				market cap	336,187,200	92,189,760	173,105,000	323,565,000	569,732,000
				AI	62.1941				

TABLE 4.2: Descriptive Statistics of Bitcoin Exchange-Traded Products (ETPs). This table presents descriptive statistics for daily Bitcoin ETP prices, returns, volumes, and market cap over the period August 2021 to July 2022. We report the mean (**Mean**), standard deviation (**StDev**), minimum (**Min**), median (**Median**), and maximum (**Max**) values. Moreover, we report the Amihud illiquidity measure (AI) where the volume parameter in the denominator is in the millions of notional of the quote currency.

where $\Delta P_{i,t}$ represents the change in the log price series $p_{i,t}$ of venue i at time t.

We estimate CS from the normalised orthogonal coefficients to the vector of error correction as

$$CS_1 = \gamma_1 = \frac{\alpha_2}{\alpha_2 - \alpha_1} \quad \text{and} \quad CS_2 = \gamma_2 = \frac{\alpha_1}{\alpha_1 - \alpha_2}. \tag{4.4}$$

Using the covariance matrix of the reduced form VECM error terms, given as

$$M = \begin{pmatrix} m_{1,1} & 0 \\ m_{1,2} & m_{2,2} \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2(1-\rho^2)^{\frac{1}{2}} \end{pmatrix} \tag{4.5}$$

we compute IS as

$$IS_1 = \frac{(\gamma_1 m_{1,1} + \gamma_2 m_{1,2})^2}{(\gamma_1 m_{1,1} + \gamma_2 m_{1,2})^2 + (\gamma_2 m_{2,2})^2} \quad \text{and} \quad IS_2 = \frac{(\gamma_2 m_{2,2})^2}{(\gamma_1 m_{1,1} + \gamma_2 m_{1,2})^2 + (\gamma_2 m_{2,2})^2}. \tag{4.6}$$

The literature highlights that IS and CS are sensitive to the relative level of noise in each market. Hence, on their own, these measures are likely to overstate the contribution to price discovery of the less noisy market. Yan & Zivot (2010) and Putniņš (2013) show that a combination of the two measures can remove dependence on noise and liquidity shocks. Specifically, the Information Leadership (IL) metric of Yan & Zivot (2010) is expressed as

$$IL_1 = \left| \frac{IS_1}{IS_2} \frac{CS_2}{CS_1} \right| \quad \text{and} \quad IL_2 = \left| \frac{IS_2}{IS_1} \frac{CS_1}{CS_2} \right|. \tag{4.7}$$

Unlike IS and CS, the IL measure does not represent a proportion, whereby the sum of IL_1 and IL_2 do not necessarily equal unity. Instead, IL_1 ranges from $[0, \infty)$, where values over (under) one imply that p_1 leads (lags) in the process of price discovery. To standardise IL, Putniņš (2013) proposes the Information Leadership Share (ILS), written as

$$ILS_1 = \frac{IL_1}{IL_1 + IL_2} \quad \text{and} \quad ILS_2 = \frac{IL_2}{IL_1 + IL_2}. \tag{4.8}$$

Values of ILS range between zero and one, with numbers higher (lower) than 0.5, suggesting that the corresponding market leads (lags) in price discovery.

4.4 Results

We calculate the above-mentioned metrics for all combinations of Bitcoin spot exchanges and ETPs using 1-minute, 5-minute, 60-minute, and 1-day sampling frequencies. The results we obtain for all exchange-ETP combinations are broadly consistent. Due to spatial limitations, Table 4.3 only shows the results for the top three exchanges and ETPs by average daily traded volume.

TABLE 4.3: Price Discovery Metrics between Bitcoin Exchange-Traded Products (ETPs) and Spot Markets. This table presents the values for the Component Share (**CS**), Information Share (**IS**), Information Leadership (**IL**), and Information Leadership Share (**ILS**) between Bitcoin ETP and spot markets based on 1-minute, 5-minute, 60-minute, and 1-day sampled price data.

Freq	**ETP**	**Exchange**	**Market**	**CS**	**IS**	**IL**	**ILS**
1 min	BTCE	$\text{Binance}_{\text{BTC/USDT}}$	ETP	0.065	0.004	0.060	0.004
			Exchange	0.935	0.996	16.762	0.996
		$\text{Coinbase}_{\text{BTC/USD}}$	ETP	0.069	0.005	0.066	0.004
			Exchange	0.931	0.995	15.148	0.996
		$\text{Huobi}_{\text{BTC/USDT}}$	ETP	0.065	0.004	0.062	0.004
			Exchange	0.935	0.996	16.045	0.996
	SBTCU	$\text{Binance}_{\text{BTC/USDT}}$	ETP	0.014	0.002	0.145	0.020
			Exchange	0.986	0.998	6.916	0.980
		$\text{Coinbase}_{\text{BTC/USD}}$	ETP	0.012	0.001	0.118	0.014
			Exchange	0.988	0.999	8.448	0.986
		$\text{Huobi}_{\text{BTC/USDT}}$	ETP	0.014	0.002	0.143	0.020
			Exchange	0.986	0.998	6.998	0.980
	VBTC	$\text{Binance}_{\text{BTC/USDT}}$	ETP	0.073	0.007	0.085	0.007
			Exchange	0.927	0.993	11.81	0.993
		$\text{Coinbase}_{\text{BTC/USD}}$	ETP	0.075	0.007	0.089	0.008
			Exchange	0.925	0.993	11.271	0.992
		$\text{Huobi}_{\text{BTC/USDT}}$	ETP	0.072	0.007	0.086	0.007
			Exchange	0.928	0.993	11.564	0.993
5 min	BTCE	$\text{Binance}_{\text{BTC/USDT}}$	ETP	0.065	0.016	0.232	0.051
			Exchange	0.935	0.984	4.311	0.949
		$\text{Coinbase}_{\text{BTC/USD}}$	ETP	0.071	0.015	0.203	0.040
			Exchange	0.929	0.985	4.925	0.960
		$\text{Huobi}_{\text{BTC/USDT}}$	ETP	0.064	0.016	0.233	0.051
			Exchange	0.936	0.984	4.294	0.949
	SBTCU	$\text{Binance}_{\text{BTC/USDT}}$	ETP	0.017	0.003	0.190	0.035
			Exchange	0.983	0.997	5.265	0.965
		$\text{Coinbase}_{\text{BTC/USD}}$	ETP	0.010	0.001	0.131	0.017
			Exchange	0.990	0.999	7.640	0.983
		$\text{Huobi}_{\text{BTC/USDT}}$	ETP	0.017	0.003	0.191	0.035
			Exchange	0.983	0.997	5.226	0.965
	VBTC	$\text{Binance}_{\text{BTC/USDT}}$	ETP	0.072	0.009	0.123	0.015
			Exchange	0.928	0.991	8.151	0.985
		$\text{Coinbase}_{\text{BTC/USD}}$	ETP	0.086	0.009	0.098	0.009
			Exchange	0.914	0.991	10.23	0.991
		$\text{Huobi}_{\text{BTC/USDT}}$	ETP	0.071	0.009	0.124	0.015
			Exchange	0.929	0.991	8.080	0.985
60 min	BTCE	$\text{Binance}_{\text{BTC/USDT}}$	ETP	0.130	0.032	0.221	0.046
			Exchange	0.870	0.968	4.533	0.954
		$\text{Coinbase}_{\text{BTC/USD}}$	ETP	0.138	0.033	0.213	0.043
			Exchange	0.862	0.967	4.700	0.957

Continued on next page

Freq	**ETP**	**Exchange**	**Market**	**CS**	**IS**	**IL**	**ILS**
		$\text{Huobi}_{BTC/USDT}$	ETP	0.129	0.032	0.221	0.046
			Exchange	0.871	0.968	4.533	0.954
	SBTCU	$\text{Binance}_{BTC/USDT}$	ETP	0.024	0.005	0.201	0.039
			Exchange	0.976	0.995	4.978	0.961
		$\text{Coinbase}_{BTC/USD}$	ETP	0.021	0.005	0.216	0.044
			Exchange	0.979	0.995	4.640	0.956
		$\text{Huobi}_{BTC/USDT}$	ETP	0.024	0.005	0.201	0.039
			Exchange	0.976	0.995	4.973	0.961
	VBTC	$\text{Binance}_{BTC/USDT}$	ETP	0.139	0.029	0.185	0.033
			Exchange	0.861	0.971	5.405	0.967
		$\text{Coinbase}_{BTC/USD}$	ETP	0.147	0.030	0.183	0.032
			Exchange	0.853	0.970	5.478	0.968
		$\text{Huobi}_{BTC/USDT}$	ETP	0.138	0.029	0.186	0.034
			Exchange	0.862	0.971	5.368	0.966
1 day	BTCE	$\text{Binance}_{BTC/USDT}$	ETP	0.369	0.142	0.283	0.074
			Exchange	0.631	0.858	3.530	0.926
		$\text{Coinbase}_{BTC/USD}$	ETP	0.372	0.143	0.282	0.073
			Exchange	0.628	0.857	3.552	0.927
		$\text{Huobi}_{BTC/USDT}$	ETP	0.368	0.142	0.283	0.074
			Exchange	0.632	0.858	3.529	0.926
	SBTCU	$\text{Binance}_{BTC/USDT}$	ETP	0.046	0.213	5.659	0.970
			Exchange	0.954	0.787	0.177	0.030
		$\text{Coinbase}_{BTC/USD}$	ETP	0.046	0.212	5.631	0.969
			Exchange	0.954	0.788	0.178	0.031
		$\text{Huobi}_{BTC/USDT}$	ETP	0.047	0.213	5.544	0.968
			Exchange	0.953	0.787	0.180	0.032
	VBTC	$\text{Binance}_{BTC/USDT}$	ETP	0.371	0.141	0.279	0.072
			Exchange	0.629	0.859	3.580	0.928
		$\text{Coinbase}_{BTC/USD}$	ETP	0.373	0.142	0.278	0.072
			Exchange	0.627	0.858	3.598	0.928
		$\text{Huobi}_{BTC/USDT}$	ETP	0.370	0.141	0.279	0.072
			Exchange	0.630	0.859	3.579	0.928

For all sampling frequencies and metrics considered, the spot market across all exchanges leads in price discovery[3]. The ILS across spot exchanges is above 90%, implying that most information impacting Bitcoin prices stems from spot markets. This may be due to (i) the greater degree of liquidity on spot exchanges as indicated by the AI measure, (ii) more established continuously traded spot markets on crypto exchanges compared to limited market-hours trading on equity exchanges, (iii) a greater degree of anonymity on crypto exchanges, which may attract informed investors, and (iv) the fact that ETP creations is preceded by a hedge transaction in spot markets. Additionally, ILS is larger for higher frequency data, which supports the notion that information is more quickly reflected in spot markets due to their dynamic and liquid nature — as indicated by the smaller AI values for crypto spot exchanges relative to ETPs.

4.5 Conclusion

As an emerging innovation in recent years, Bitcoin has received much attention due to its unique features. Design of cryptocurrency Exchange Traded Products opens to gain exposure to the underlying asset without the need to hold Bitcoin, which could be attractive for investors. This study investigates price discovery between Bitcoin ETPs and spot markets using four popular metrics from the literature.

Using CS, IS, IL and ILS in our analysis, we find that the spot market is a leading market and has stayed the same over the time of our observations from August 2021 till July 2022. In particular, our robustness analysis with 1-minute, 5-minute, 60-minute, and 1-day sampling frequencies shows a clear price leadership of the bitcoin spot market.

Our empirical results show that spot markets dominate this process due to their deeper liquidity, continuous trading hours, and greater degree of anonymity. Nonetheless, ETPs may play a more significant role in the future as this market matures and complies with regulatory frameworks, thus gaining popularity among institutional investors.

Our findings underscore the importance of trading activities on centralised crypto exchanges in determining crypto prices despite regulators broadly dismissing these markets in favour of more traditional regulated venues. Our research shows that, at least, the analysis of determinants on price discovery leads to economically reasonable results, which can also be found in other asset classes.

[3]The sole exception is the SBTCU ETP sampled at daily intervals, a potential anomaly given that higher data frequencies for this ETP suggest that spot markets lead.

Conclusion

Chapter 1 of this thesis is dedicated to trade-based manipulation strategy, spoofing, its determination and its effect on market quality on MOEX. We find a negative relationship between spoofing activity and the change in market quality for the period just before the manipulation while demonstrating a positive relationship after the end of the spoofing strategy execution. We observe that spoofing causes a short-term increase in market quality due to misleading order book pressure. However, we find a market quality distortion after the spoofing event. More spoofing leads to faster-growing spreads just after the spoofing order cancellation, so the spread tends to increase more quickly. This effect persists for two periods subsequent to the spoofing.

Relationships in various panel regressions show a similar result. The effect identified is robust to different specifications and several modifications of the spoofing ratio. Also, our results hold after controlling for volatility, day trading volume, and periods of intensive trading during the day. We contribute to the literature by building out an analysis using intraday time dimensions according to the nature of spoofing. The change in market quality conditions leads to an unstable trading environment. We find that spoofing orders have a destabilizing effect on market quality.

Having the results that spoofing distorts market quality, the next logical step is to detect spoofing. However, detecting spoofing by itself gives low contribution to an improvement of the market environment, while forecasting the market state with a high risk of spoofing manipulation may help investors avoid entering the trade in misleading market conditions, which in the end will lead to less spoofing activity due to its inefficiency for the manipulator. In **Chapter 2**, we introduce a novel data-driven approach to the real-time prediction of market state when a spoofing event is highly probable. Our Real-Time Spoofing probability measure indicates the risk of intraday manipulative activity. We show how to identify periods when one might see suspicious activity in the order book.

Our study reveals how exchanges may improve their surveillance system by having data from the order book and suspect spoofing orders for the previous five trading days and machine learning methodology. Using our approach, regulators and vendors may forecast in real time the market state, specifically the next tick, with a high probability of spoofing order placement. The designed RTSP measure has an essential future as an adjustment possibility depending on the asset, market microstructure, and the type of manipulative activity. So, as we construct the RTSP measure using ML algorithms trained on the given dataset, the measure may forecast spoofing events and other fraudulent activities on different financial markets.

Our work does not consider the model to account for market shocks, news, dividend activity or similar macro events. We focus on intraday high-frequency data, where spoofing manipulations tend to occur. However, on top of our model, users may add layers of market performance, intraday shocks and other events. Future research may use the designed methodology for other disruptive activities and other markets, adjusting market variables and model parameters; hence, the models capture essential features such as volatility, tick size and trading activity and automatically adapt depending on the training dataset. That flexibility and the robustness of the developed model make our research an essential brick in market microstructure literature.

Our study in **Chapter 3** contributes to the debate on informed trading. We show empirically using data on the emerging futures market how different trading practices by client types affect market outcomes from intraday and daily perspectives in the futures market. Our results suggest that the buying power and order of different client groups are highly informative about daily and intraday future returns. We also find differences in customers' intraday and longer trading practices. While our main findings highlight that retail traders and institutional investors correctly anticipate intraday and short-term returns, we also identify how these types of traders differ. Retail traders', institutional investors' and non-residents' order flows are good predictors for intraday returns in commodity futures, while only institutional investors' order flow positively predicts returns in stock futures. Intraday order flows of corporate clients, retail and non-resident traders predict returns on currency futures, while the order flow of institutional investors does not.

We show empirically in three dimensions (correlation, regression and portfolio analyses) that the order flow signal differs for daily and intraday trading, and further studies need to be built separately for high-frequency and algorithmic trading and daily trading. Also, we find the major differences while splitting the data into subsamples based on the underlying asset classes. For example, we discovered that non-residents anticipate future returns in currency and commodity futures.

Our empirical study in Chapter 3 is a multi-dimensional research on informed trading in the futures market, which may be a base level for further more profound and niche research on, for example, informed trading on commodity futures. Our main contribution is showing that customer order flow is highly informative for intraday time dimensions.

As an emerging innovation in recent years, Bitcoin has received much attention due to its unique features. Design of cryptocurrency Exchange Traded Products opens to gain exposure to the underlying asset without the need to hold Bitcoin, which could be attractive for investors. Our study in **Chapter 4** investigates price discovery between Bitcoin ETPs and spot markets using four popular metrics from the literature. Our robustness analysis with 1-minute, 5-minute, 60-minute, and 1-day sampling frequencies shows a clear price leadership of the bitcoin spot market.

Our empirical results show that spot markets dominate this process due to their deeper liquidity, continuous trading hours, and greater degree of anonymity. Nonetheless, ETPs may play a more significant role in the future as this market matures and complies with regulatory frameworks, thus gaining popularity among institutional investors. Our findings underscore the importance of trading activities on centralised crypto exchanges in determining crypto prices despite regulators broadly dismissing these markets in favour of more traditional regulated venues.

Bibliography

Aggarwal, R. K. & Wu, G. (2006), 'Stock market manipulations', *The Journal of Business* **79**(4), 1915–1953.

Ahn, K., Bi, Y. & Sohn, S. (2019), 'Price discovery among sse 50 index-based spot, futures, and options markets', *Journal of Futures Markets* **39**(2), 238–259.

Aitken, M. & Frino, A. (1996), 'The determinants of market bid ask spreads on the australian stock exchange: Cross-sectional analysis', *Accounting & Finance* **36**(1), 51–63.

Aitken, M., Harris, F. & Ji, S. (2009), Trade-based manipulation and market efficiency: A cross-market comparison, *in* '22nd Australasian Finance and Banking Conference', p. 18.

Aitken, M., Harris, F. & Ji, S. (2015), 'A worldwide examination of exchange market quality: Greater integrity increases market efficiency', *Journal of Business Ethics* **132**, 147–170.

Aitken, M. J., de B Harris, F. H. & Ji, S. (2015), 'A worldwide examination of exchange market quality: Greater integrity increases market efficiency', *Journal of Business Ethics* **132**(1), 147–170.

Akyildirim, E., Corbet, S., Katsiampa, P., Kellard, N. & Sensoy, A. (2020), 'The development of bitcoin futures: Exploring the interactions between cryptocurrency derivatives', *Finance Research Letters* **34**, 101234.

Alexander, C., Choi, J., Park, H. & Sohn, S. (2020), 'Bitmex bitcoin derivatives: Price discovery, informational efficiency, and hedging effectiveness', *Journal of Futures Markets* **40**(1), 23–43.

Allen, F. & Gale, D. (1992), 'Stock-price manipulation', *Review of financial studies* **5**(3), 503–529.

Allen, F. & Gorton, G. (1992), 'Stock price manipulation, market microstructure and asymmetric information', *European Economic Review* **36**, 624–630.

Amihud, Y. (2002), 'Illiquidity and stock returns: cross-section and time-series effects', *Journal of financial markets* **5**(1), 31–56.

Angel, J. & McCabe, D. (2010), 'The ethics of speculation', *Journal of Business Ethics* **90**, 277–286.

Balcilar, M., Bouri, E., Gupta, R. & Roubaud, D. (2017), 'Can volume predict bitcoin returns and volatility? a quantiles-based approach', *Economic Modelling* **64**, 74–81.

Barber, B. M., Odean, T. & Zhu, N. (2008), 'Do retail trades move markets?', *The Review of Financial Studies* **22**(1), 151–186.

Barrot, J.-N., Kaniel, R. & Sraer, D. (2016), 'Are retail traders compensated for providing liquidity?', *Journal of Financial Economics* **120**(1), 146–168.

Baur, D. G. & Dimpfl, T. (2019), 'Price discovery in bitcoin spot or futures?', *Journal of Futures Markets* **39**(7), 803–817.

Beneki, C., Koulis, A., Kyriazis, N. A. & Papadamou, S. (2019), 'Investigating volatility transmission and hedging properties between bitcoin and ethereum', *Research in International Business and Finance* **48**, 219–227.

Bergmeir, C., Hyndman, R. J. & Koo, B. (2018), 'A note on the validity of cross-validation for evaluating autoregressive time series prediction', *Computational Statistics & Data Analysis* **120**, 70–83.

Biais, B., Hillion, P. & Spatt, C. (1995), 'An empirical analysis of the limit order book and the order flow in the paris bourse', *the Journal of Finance* **50**(5), 1655–1689.

Blume, M. E. & Goldstein, M. A. (1992), 'Displayed and effective spreads by market', *Rodney L. White Center for Financial Research Working Paper* (27-92).

Boehmer, E., Jones, C. M., Zhang, X. & Zhang, X. (2021), 'Tracking retail investor activity', *The Journal of Finance* **76**(5), 2249–2305.

Bohl, M. T., Salm, C. A. & Schuppli, M. (2011), 'Price discovery and investor structure in stock index futures', *Journal of Futures Markets* **31**(3), 282–306.

Booth, G. G., So, R. W. & Tse, Y. (1999), 'Price discovery in the german equity index derivatives markets', *Journal of Futures Markets: Futures, Options, and Other Derivative Products* **19**(6), 619–643.

Brandt, M. W. & Kavajecz, K. A. (2004), 'Price discovery in the us treasury market: The impact of orderflow and liquidity on the yield curve', *The Journal of Finance* **59**(6), 2623–2654.

Brandvold, M., Molnár, P., Vagstad, K. & Valstad, O. C. A. (2015), 'Price discovery on bitcoin exchanges', *Journal of International Financial Markets, Institutions and Money* **36**, 18–35.

Brauneis, A. & Mestel, R. (2018), 'Price discovery of cryptocurrencies: Bitcoin and beyond', *Economics Letters* **165**, 58–61.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Brogaard, J., Carrion, A., Moyaert, T., Riordan, R., Shkilko, A. & Sokolov, K. (2018), 'High frequency trading and extreme price movements', *Journal of Financial Economics* **128**(2), 253–265.

Brogaard, J., Li, D. & Yang, J. (2022), 'Does high frequency market manipulation harm market quality?', *Available at SSRN* .

Broman, M. S. (2016), 'Liquidity, style investing and excess comovement of exchange-traded fund returns', *Journal of Financial Markets* **30**, 27–53.

Brown, D. C., Davies, S. W. & Ringgenberg, M. C. (2021), 'Etf arbitrage, non-fundamental demand, and return predictability', *Review of Finance* **25**(4), 937–972.

Cao, Y., Li, Y., Coleman, S., Belatreche, A. & McGinnity, T. M. (2014), Detecting price manipulation in the financial market, *in* '2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)', IEEE, pp. 77–84.

Cao, Y., Li, Y., Coleman, S., Belatreche, A. & McGinnity, T. M. (2015), 'Adaptive hidden markov model with anomaly states for price manipulation detection', *IEEE transactions on neural networks and learning systems* **26**(2), 318–330.

Cartea, A., Jaimungal, S. & Wang, Y. (2020), 'Spoofing and price manipulation in order-driven markets', *Applied Mathematical Finance* **27**, 1–2, 67–98.

Cartea, Á., Payne, R., Penalva, J. & Tapia, M. (2019), 'Ultra-fast activity and intraday market quality', *Journal of Banking & Finance* **99**, 157–181.

Cerrato, M., Sarantis, N. & Saunders, A. (2011), 'An investigation of customer order flow in the foreign exchange market', *Journal of Banking & Finance* **35**(8), 1892–1906.

Cespa, G., Gargano, A., Riddiough, S. J. & Sarno, L. (2022), 'Foreign exchange volume', *The Review of Financial Studies* **35**(5), 2386–2427.

Chamberlain, T., Cheung, C. & Kwan, C. (1989), 'Expiration-day effects of index futures and options: Some canadian evidence', *Financial Analysts Journal* **45**(5), 67–71.

Chan, K. (1992), 'A further analysis of the lead–lag relationship between the cash market and stock index futures market', *The Review of Financial Studies* **5**(1), 123–152.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'Smote: synthetic minority over-sampling technique', *Journal of artificial intelligence research* **16**, 321–357.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T. et al. (2015), 'Xgboost: extreme gradient boosting', *R package version 0.4-2* **1**(4), 1–4.

Chen, Y.-L. & Gau, Y.-F. (2010), 'News announcements and price discovery in foreign exchange spot and futures markets', *Journal of Banking & Finance* **34**(7), 1628–1636.

Chiou, J., Wu, P., Wang, A. & Huang, B. (2007), 'The asymmetric information and price manipulation in stock market', *Applied Economics* **39**(7), 883–891.

Chu, Q. C., Hsieh, W.-l. G. & Tse, Y. (1999), 'Price discovery on the s&p 500 index markets: An analysis of spot index, index futures, and spdrs', *International Review of Financial Analysis* **8**(1), 21–34.

Claeskens, G., Magnus, J. R., Vasnev, A. L. & Wang, W. (2016), 'The forecast combination puzzle: A simple theoretical explanation', *International Journal of Forecasting* **32**(3), 754–762.

Corbet, S., Lucey, B., Peat, M. & Vigne, S. (2018), 'Bitcoin futures—what use are they?', *Economics Letters* **172**, 23–27.

Covrig, V., Ding, D. K. & Low, B. S. (2004), 'The contribution of a satellite market to price discovery: Evidence from the singapore exchange', *Journal of Futures Markets: Futures, Options, and Other Derivative Products* **24**(10), 981–1004.

Cumming, D., Johan, S. & Li, D. (2011), 'Exchange trading rules and stock market liquidity', *Journal of Financial Economics* **99**(3), 651–671.

Czech, R., Huang, S., Lou, D. & Wang, T. (2021), 'Informed trading in government bond markets', *Journal of Financial Economics* **142**(3), 1253–1274.

Da, Z. & Shive, S. (2018), 'Exchange traded funds and asset return correlations', *European Financial Management* **24**(1), 136–168.

Diaz, D., Theodoulidis, B. & Sampaio, P. (2011), 'Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices', *Expert Systems with Applications* **38**(10), 12757–12771.

Dimpfl, T., Flad, M. & Jung, R. C. (2017), 'Price discovery in agricultural commodity markets in the presence of futures speculation', *Journal of Commodity Markets* **5**, 50–62.

Dodd-Frank (2010), 'Dodd-frank wall street reform and consumer protection act'. **URL:** *https://www.govinfo.gov/content/pkg/PLAW-111publ203/pdf/PLAW-111publ203.pdf*

Dolatabadi, S., Nielsen, M. Ø. & Xu, K. (2015), 'A fractionally cointegrated var analysis of price discovery in commodity futures markets', *Journal of Futures Markets* **35**(4), 339–356.

Egginton, J. F., Van Ness, B. F. & Van Ness, R. A. (2016), 'Quote stuffing', *Financial Management* **45**(3), 583–608.

Egginton, J., Ness, B. & Ness, R. (2016), 'Quote stuffing', *Financial Management* **45**, 583–608.

Evans, M. D. & Lyons, R. K. (2002), 'Order flow and exchange rate dynamics', *Journal of political economy* **110**(1), 170–180.

Evans, M. D. & Lyons, R. K. (2008), 'How is macro news transmitted to exchange rates?', *Journal of Financial Economics* **88**(1), 26–50.

Fassas, A. P., Papadamou, S. & Koulis, A. (2020), 'Price discovery in bitcoin futures', *Research in International Business and Finance* **52**, 101116.

Fassas, A. P. & Siriopoulos, C. (2019), 'Intraday price discovery and volatility spillovers in an emerging market', *International Review of Economics & Finance* **59**, 333–346.

Fishe, R. P. & Smith, A. D. (2012), 'Identifying informed traders in futures markets', *Journal of Financial Markets* **15**(3), 329–359.

Glosten, L., Nallareddy, S. & Zou, Y. (2021), 'Etf activity and informational efficiency of underlying securities', *Management Science* **67**(1), 22–47.

Golmohammadi, K. & Zaiane, O. R. (2015), Time series contextual anomaly detection for detecting market manipulation in stock market, *in* '2015 IEEE international conference on data science and advanced analytics (DSAA)', IEEE, pp. 1–10.

Gong, J. & Kim, H. (2017), 'Rhsboost: Improving classification performance in imbalance data', *Computational Statistics & Data Analysis* **111**, 1–13.

Gonzalo, J. & Granger, C. (1995), 'Estimation of common long-memory components in cointegrated systems', *Journal of Business & Economic Statistics* **13**(1), 27–35.

Guiso, L., Sapienza, P. & Zingales, L. (2008), 'Trusting the stock market', *the Journal of Finance* **63**(6), 2557–2600.

Harris, F. H. d., McInish, T. H. & Wood, R. A. (2002), 'Security price adjustment across exchanges: an investigation of common factor components for dow stocks', *Journal of financial markets* **5**(3), 277–308.

Hasbrouck, J. (1991*a*), 'Measuring the information content of stock trades', *The Journal of Finance* **46**(1), 179–207.

Hasbrouck, J. (1991*b*), 'The summary informativeness of stock trades: An econometric analysis', *The Review of Financial Studies* **4**(3), 571–595.

Hasbrouck, J. (1995), 'One security, many markets: Determining the contributions to price discovery', *The journal of Finance* **50**(4), 1175–1199.

Hasbrouck, J. (2003), 'Intraday price formation in us equity index markets', *The Journal of Finance* **58**(6), 2375–2400.

Hasbrouck, J. & Saar, G. (2013), 'Low-latency trading', *Journal of Financial Markets* **16**(4), 646–679.

Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer.

Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2017), *The elements of statistical learning: data mining, inference, and prediction*, Springer.

Hauptfleisch, M., Putniņš, T. J. & Lucey, B. (2016), 'Who sets the price of gold? london or new york', *Journal of Futures Markets* **36**(6), 564–586.

Ho, T. K. (1995), Random decision forests, *in* 'Proceedings of 3rd international conference on document analysis and recognition', Vol. 1, IEEE, pp. 278–282.

Hu, Y., Hou, Y. G. & Oxley, L. (2020), 'What role do futures markets play in bitcoin pricing? causality, cointegration and price discovery from a time-varying perspective?', *International Review of Financial Analysis* **72**, 101569.

Huang, Y. & Cheng, Y. (2015), 'Stock manipulation and its effects: pump and dump versus stabilization', *Review of Quantitative Finance and Accounting* **44**(4), 791–815.

Imisiker, S. & Tas, B. (2013), 'Which firms are more prone to stock market manipulation?', *Emerging Markets Review* **16**, 119–130.

Israeli, D., Lee, C. M. & Sridharan, S. A. (2017), 'Is there a dark side to exchange traded funds? an information perspective', *Review of Accounting Studies* **22**, 1048–1083.

Jang, H. & Lee, J. (2017), 'An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information', *Ieee Access* **6**, 5427–5437.

Jarrow, R. (1992), 'Market manipulation, bubbles, corners, and short squeezes', *Journal of financial and Quantitative Analysis* **27**, 311–336.

Jin, M., Kearney, F., Li, Y. & Yang, Y. C. (2020), 'Intraday time-series momentum: Evidence from china', *Journal of Futures Markets* **40**(4), 632–650.

Kaniel, R., Liu, S., Saar, G. & Titman, S. (2012), 'Individual investor trading and return patterns around earnings announcements', *The Journal of Finance* **67**(2), 639–680.

Kaniel, R., Saar, G. & Titman, S. (2008), 'Individual investor trading and stock returns', *The Journal of finance* **63**(1), 273–310.

Kapar, B. & Olmo, J. (2019), 'An analysis of price discovery between bitcoin futures and spot markets', *Economics Letters* **174**, 62–64.

Kawaller, I. G., Koch, P. D. & Koch, T. W. (1987), 'The temporal price relationship between s&p 500 futures and the s&p 500 index', *The Journal of Finance* **42**(5), 1309–1329.

Kelley, E. K. & Tetlock, P. C. (2013), 'How wise are crowds? insights from retail orders and stock returns', *The Journal of Finance* **68**(3), 1229–1265.

Kelley, E. K. & Tetlock, P. C. (2017), 'Retail short selling and stock prices', *The Review of Financial Studies* **30**(3), 801–834.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P. et al. (2006), 'Handling imbalanced datasets: A review', *GESTS international transactions on computer science and engineering* **30**(1), 25–36.

Koutmos, G. & Tucker, M. (1996), 'Temporal relationships and dynamic interactions between spot and futures stock markets', *The Journal of Futures Markets (1986-1998)* **16**(1), 55.

Kuiper, W. E., Pennings, J. M. & Meulenberg, M. T. (2002), 'Identification by full adjustment: evidence from the relationship between futures and spot prices', *European Review of Agricultural Economics* **29**(1), 67–84.

Kyle, A. S. (1985), 'Continuous auctions and insider trading', *Econometrica: Journal of the Econometric Society* pp. 1315–1335.

Leangarun, T., Tangamchit, P. & Thajchayapong, S. (2016), Stock price manipulation detection using a computational neural network model, *in* '2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)', IEEE, pp. 337–341.

Lee, C. M. (1993), 'Market integration and price execution for nyse-listed securities', *The Journal of Finance* **48**(3), 1009–1038.

Lee, E., Eom, K. & Park, K. (2013), 'Microstructure-based manipulation: Strategic behavior and performance of spoofing traders', *Journal of Financial Markets* **16**(2), 227–252.

Lettau, M. & Madhavan, A. (2018), 'Exchange-traded funds 101 for economists', *Journal of Economic Perspectives* **32**(1), 135–154.

Love, R. & Payne, R. (2008), 'Macroeconomic news, order flows, and exchange rates', *Journal of Financial and Quantitative Analysis* **43**(2), 467–488.

Lyons, R. K. (1995), 'Tests of microstructural hypotheses in the foreign exchange market', *Journal of Financial Economics* **39**(2-3), 321–351.

Madhavan, A. N. (2016), *Exchange-traded funds and the new dynamics of investing*, Oxford University Press.

Madhavan, A. & Sobczyk, A. (2016), 'Price dynamics and liquidity of exchange-traded funds', *Journal of Investment Management* **14**(2), 1–17.

Marshall, B. R., Nguyen, N. H. & Visaltanachoti, N. (2013), 'Etf arbitrage: Intraday evidence', *Journal of Banking & Finance* **37**(9), 3486–3498.

Martínez-Miranda, E., McBurney, P. & Howard, M. J. (2016), Learning unfair trading: A market manipulation analysis from the reinforcement learning perspective, *in* '2016 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)', IEEE, pp. 103–109.

Menkhoff, L., Sarno, L., Schmeling, M. & Schrimpf, A. (2016), 'Information flows in foreign exchange markets: Dissecting customer currency trades', *The Journal of Finance* **71**(2), 601–634.

Menkveld, A. J. & Saru, I. L. (2022), 'Who knows? information differences between trader types'.

Montgomery, J. (2016), Spoofing, market manipulation, and the limit-order book, Technical report.

More, A. & Rana, D. P. (2017), Review of random forest classification techniques to resolve data imbalance, *in* '2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)', IEEE, pp. 72–78.

Öğüt, H., Doğanay, M. M. & Aktaş, R. (2009), 'Detecting stock-price manipulation in an emerging market: The case of turkey', *Expert Systems with Applications* **36**(9), 11944–11949.

Park, J. Y. & Hahn, S. B. (1999), 'Cointegrating regressions with time varying coefficients', *Econometric Theory* **15**(5), 664–703.

Peri, M., Baldi, L. & Vandone, D. (2013), 'Price discovery in commodity markets', *Applied Economics Letters* **20**(4), 397–403.

Perraudin, W. & Vitale, P. (1996), Interdealer trade and information flows in a decentralized foreign exchange market, *in* 'The microstructure of foreign exchange markets', University of Chicago Press, pp. 73–106.

Punniyamoorthy, M. & Thoppan, J. (2012), 'Ann-ga based model for stock market surveillance', *Journal of Financial Crime* **20**(1), 52–66.

Punniyamoorthy, M. & Thoppan, J. J. (2013), 'Ann-ga based model for stock market surveillance', *Journal of Financial Crime* .

Putniņš, T. (2020), *An overview of market manipulation*, 1st edn, John Wiley & Sons Inc., United States, pp. 13–44.

Putniņš, T. J. (2013), 'What do price discovery metrics really measure?', *Journal of Empirical Finance* **23**, 68–83.

Ranaldo, A. (2004), 'Order aggressiveness in limit order book markets', *Journal of Financial Markets* **7**(1), 53–74.

Richie, N., Daigler, R. T. & Gleason, K. C. (2008), 'The limits to stock index arbitrage: Examining s&p 500 futures and spdrs', *Journal of Futures Markets: Futures, Options, and Other Derivative Products* **28**(12), 1182–1205.

Schlag, C. & Stoll, H. (2005), 'Price impacts of options volume', *Journal of Financial Markets* **8**(1), 69–87.

Shi, H., Worden, K. & Cross, E. J. (2018), 'A regime-switching cointegration approach for removing environmental and operational variations in structural health monitoring', *Mechanical Systems and Signal Processing* **103**, 381–397.

So, R. W. & Tse, Y. (2004), 'Price discovery in the hang seng index markets: index, futures, and the tracker fund', *Journal of Futures Markets: Futures, Options, and Other Derivative Products* **24**(9), 887–907.

Stephens, A., Fardon, Z., Kirkpatrick, K., Dedman, R., Watling, M., Wissa, M. & Nettesheim, M. (2019), '"spoofing" the market: a comparison of us and uk law and enforcement'.

Stoll, H. & Whaley, R. (1987), 'Program trading and expiration-day effects', *Financial Analysts Journal* **43**, 16–28.

Tao, X., Day, A., Ling, L. & Drapeau, S. (2022), 'On detecting spoofing strategies in high-frequency trading', *Quantitative Finance* **22**(8), 1405–1425.

Theissen, E. (2012), 'Price discovery in spot and futures markets: A reconsideration', *The European Journal of Finance* **18**(10), 969–987.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Tse, Y. (1999), 'Price discovery and volatility spillovers in the djia index and futures markets', *Journal of Futures markets* **19**(8), 911–930.

Tuccella, J.-N., Nadler, P. & Şerban, O. (2021), 'Protecting retail investors from order book spoofing using a gru-based detection model', *arXiv:2110.03687* .

Van Kervel, V. & Menkveld, A. J. (2019), 'High-frequency trading around large institutional orders', *The Journal of Finance* **74**(3), 1091–1137.

Wahab, M. & Lashgari, M. (1993), 'Price dynamics and error correction in stock index and stock index futures markets: A cointegration approach', *Journal of Futures Markets* **13**(7), 711–742.

Wang, X. & Wellman, M. (2017), Spoofing the limit order book: An agent-based model, *in* 'Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems', International Foundation for Autonomous Agents and Multiagent Systems, pp. 651–659.

Wang, Y. (2015), 'Strategic spoofing order trading by different types of investors in the futures markets', *Wall Street Journal* .

Williams, B. & Skrzypacz, A. (2020), 'Spoofing in equilibrium', *Available at SSRN* .

Yan, B. & Zivot, E. (2010), 'A structural analysis of price discovery measures', *Journal of Financial Markets* **13**(1), 1–19.

Yang, J., Yang, Z. & Zhou, Y. (2012), 'Intraday price discovery and volatility transmission in stock index and stock index futures markets: Evidence from china', *Journal of Futures Markets* **32**(2), 99–121.

Zhou, Z.-H. (2012), *Ensemble methods: foundations and algorithms*, CRC press.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the royal statistical society: series B (statistical methodology)* **67**(2), 301–320.