

COPENHAGEN BUSINESS SCHOOL

FINANCE AND INVESTMENTS

MAY 15, 2019

HIGH FREQUENCY LEAD-LAG RELATIONSHIPS IN THE BITCOIN MARKET

*“An empirical analysis of the price movements of bitcoin on different cryptocurrency
exchanges during the year of 2018”*

MASTER’S THESIS

Author

Bendik Norheim Schei

DE93305

Supervisor

Christian Rix-Nielsen

Abstract

In a rational and efficiently functioning market, returns on financial products that represent the same underlying asset should be perfectly simultaneously correlated. However, due to market imperfections, lead-lag relationships are a commonly observed phenomenon in traditional financial markets. This thesis examines price movements in a new and emerging market. Bitcoin is the oldest and most liquid cryptocurrency and is traded on numerous exchanges. By the use of a traditional cointegration and causality approach, bidirectional relationships are confirmed between all bitcoin prices tested. A modern high frequency approach with the use of tick-by-tick data reveals strongly asymmetric cross-correlation functions. Some bitcoin prices follow the path of others with a time lag up to 15 seconds. The analysis furthermore confirms that the lead-lag relationships are affected by the rate of information arrival, whose proxy is the unexpected trading volume on the exchanges. Moreover, sophisticated investors have a more significant effect on the lead-lag relationship than non-sophisticated ones. A simple trading strategy is used to forecast mid-quote changes in lagging exchanges with directional accuracy of up to 70%. Profitable arbitrage opportunities are found by the use of an algorithm-based trading strategy, under the assumptions of trading at the lowest fee levels and mid-quote execution. Nevertheless, trading fees, price slippage and lack of liquidity are found as the most important limits to arbitrage. Several aspects could explain why lead-lag relationships are found. Exchange characteristics like infrastructure, fee structure, location and investors types are important. However, the analysis in this thesis points towards liquidity of exchanges as the most likely explanation.

TABLE OF CONTENTS

<u>1</u>	<u>INTRODUCTION</u>	<u>4</u>
1.1	RESEARCH QUESTIONS	5
<u>2</u>	<u>LITERATURE REVIEW</u>	<u>6</u>
2.1	LEAD-LAG RELATIONSHIPS	6
2.2	CRYPTOCURRENCY	7
<u>3</u>	<u>THEORETICAL FRAMEWORK</u>	<u>8</u>
3.1	BLOCKCHAIN & BITCOIN	8
3.1.1	BLOCKCHAIN TECHNOLOGY	8
3.1.2	THE HISTORY OF BITCOIN	9
3.2	CRYPTOCURRENCY EXCHANGES	10
3.3	EFFICIENT MARKETS	12
3.4	STATIONARITY	13
3.5	LEAD-LAG RELATIONSHIP	15
3.5.1	COINTEGRATION	15
3.5.2	GRANGER CAUSALITY	16
3.5.3	HIGH FREQUENCY HAYASHI-YOSHIDA CROSS-CORRELATION	17
3.5.4	DISCUSSION	18
3.6	INFORMATION ARRIVAL	19
3.6.1	TRADING VOLUME	20
<u>4</u>	<u>METHODOLOGY</u>	<u>21</u>
4.1	RESEARCH PHILOSOPHY AND APPROACH	21
4.2	RESEARCH DESIGN	21
4.3	DATA COLLECTION	21
4.4	DATA PREPARATION	22
4.5	RESEARCH QUALITY	22
4.6	LIMITATIONS AND WEAKNESSES	23
4.7	CAUSALITY APPROACH	23
4.7.1	TESTING FOR STATIONARITY	23
4.7.2	VECTOR AUTOREGRESSIVE MODEL	26
4.7.3	JOHANSEN CO-INTEGRATION TEST	26
4.7.4	WALD TEST FOR GRANGER CAUSALITY	28
4.8	THE HAYASHI-YOSHIDA ESTIMATOR	29

4.9	LINEAR REGRESSION	32
4.9.1	ASSUMPTIONS	33
5	DATA PRESENTATION	35
5.1	DATA SELECTION	35
5.2	DATA CLEANING	37
5.3	VALIDITY AND RELIABILITY	37
6	DATA ANALYSIS	38
6.1	DESCRIPTIVE STATISTICS	38
6.2	LEAD-LAG RELATIONSHIPS	41
6.2.1	CAUSALITY APPROACH	41
6.2.2	HAYASHI-YOSHIDA CROSS-CORRELATION ANALYSIS	47
6.2.3	REGRESSION ANALYSIS	56
6.3	TRADING STRATEGY	66
6.3.1	A SIMPLE FORECASTING STRATEGY	66
6.3.2	ALGORITHM-BASED STRATEGY	67
6.3.3	LIMITATION OF ARBITRAGE OPPORTUNITIES	70
7	DISCUSSION	72
7.1	REASONS FOR LEAD-LAG RELATIONSHIPS	72
7.1.1	INFRASTRUCTURE	72
7.1.2	FEE STRUCTURE	74
7.1.3	LOCATION	75
7.1.4	INVESTORS	76
7.1.5	TRADING VOLUME	76
7.2	A FUTURE PERSPECTIVE	77
8	CONCLUSION	79
9	BIBLIOGRAPHY	81
10	APPENDIX	91
10.1	TABLES	91
10.2	FIGURES	114
10.3	PYTHON CODE	121

1 INTRODUCTION

In October 2008, Satoshi Nakamoto published a white paper of eight pages concerning a new peer-to-peer payment system. Bitcoin was born, a cryptocurrency emerging from the financial crisis that put a shock through the financial markets during the same year. The object was simple; to make people financially independent from banks, governments and other third parties, through a borderless, decentralized currency (Nakamoto, 2008). It took almost ten years before this new asset got the attention of the public. 2017 was called the year of hype and bubble, with a price increase of over 2,500% from the bottom to the top (CoinMarketCap, 2019). 2018 painted a different picture. The overall cryptocurrency market saw an 80% decrease throughout the year, and discussions between critics and bitcoin maximalists flourished. Behind all the media coverage, a new asset class in rapid development emerged. Significant players like the New York Stock Exchange, Microsoft, and Starbucks entered, with huge plans to make bitcoin and cryptocurrency accessible to the public (Dale, 2018).

There is no definite solution to how bitcoin and cryptocurrencies should be valued. Some compare it to the tulip bubble, while others say it is the future of financial markets. Regulators and governments around the world both praise and ban this new market. This thesis will not discuss the technology, the regulations, nor the opinions of the Bitcoin community itself, but will solely study the price movements of bitcoin as a financial asset, with an objective and neutral approach.

Bitcoin is unique. It is the first cryptocurrency and dominates over 50% of the total market capitalization of the cryptocurrency market. It is driven by market forces only and is traded at nearly 90 different exchanges, against several traditional currencies and cryptocurrencies (“Bitcoin Exchanges”, 2019).

This thesis seeks to understand how the bitcoin market behaves, by analyzing the lead-lag relationships across different cryptocurrency exchanges. The phenomenon of lead-lag relationships has been studied for decades in traditional financial markets, with new theories, approaches, and methods evolving with the technological improvements of trading. High frequency trading and algorithm-based methods have pushed academics to approaches that study movements down to milliseconds.

Today, the largest cryptocurrency exchanges are highly efficient with minimal arbitrage opportunities of the bitcoin price itself (Bitwise Asset Management, 2019). Is this the case for the movements of the bitcoin price as well, or are some exchanges leading the price movements of bitcoin?

1.1 RESEARCH QUESTIONS

This evolvment, in addition to the discussion above results in the following research question for this thesis:

As a globally traded, borderless cryptocurrency, how are the price movements of bitcoin on different exchanges connected, and what can explain potential differences?

In order to answer this question, a series of sub-questions are formed.

- What kind of lead-lag relationships are found between the most efficient exchanges?
- How are lead-lag relationships affected by information arrival through unexpected volume changes?
- To what extent can arbitrageurs take advantage of the possible lead-lag relationships?
- Which exchange characteristics are possible explanations of lead-lag relationships?

This thesis will study the bitcoin price on different cryptocurrency exchanges and try to understand why price movements fluctuate differently across them. The thesis' structure is as follows:

Section 2 presents the findings of academic literature and set the foundation for the theoretical approach of the thesis. Section 3 presents the theoretical frameworks that are relevant to this study. In Section 4, the methodology will be presented and describes the approaches that will be used in the analysis. Furthermore, it describes the research design, data collection, preparation, and limitations. Section 5 describes the data used in the analysis. Section 6 consists of the data analysis. The analysis includes descriptive statistics, two different approaches to lead-lag relationships, a study on the effect of information arrival and trading strategies based on the findings. Section 7 consists of a profound discussion of lead-lag relationships and a future perspective. Finally, Section 8 concludes and summarizes the research question and its sub-questions.

2 LITERATURE REVIEW

2.1 LEAD-LAG RELATIONSHIPS

Several studies have been done on the lead-lag relationships between assets. However, studies exploring these relationships in the cryptocurrency market are minimal. Most studies focus on the equity market, which can be divided between studies on equities in the same country and between countries. Furthermore, studies on lead-lag relationships in the same country focus on related securities and unrelated securities. Related securities are normally spot markets and derivatives instruments, and unrelated focuses on different stocks. This thesis will study the lead-lag relationship of bitcoin prices, which represent the same asset. Hence, the presented literature will be related to lead-lag relationships on related securities. These are normally studies on the relationship between spot and futures markets. Spot and futures have been studied for a long time. Several studies show that the relationship is bidirectional. Chiang and Fong (2001), Nam et al. (2006) and Ergün (2009) confirm a bidirectional relationship, where the leadership of the futures is both stronger and over a longer time period. Kawaller et al. (1987) found that futures lead spot markets by up to 45 minutes. Results where spots lead futures have also been found, by up to 15 minutes in Chan (1992).

These relationships are commonly explained by a couple of different factors. Both Shyy et al. (1996) and Brooks et al. (1999) describe the relationship by infrequent and non-synchronous trading. However, Stoll and Whaley (1990), Grünbichler et al. (1994), Martikainen et al. (1995), and Fleming et al. (1996) argue that these implications can be corrected for. Their results showed lead-lag relationships, even after considering these troubling trading patterns. Others focus on the cost of trading as the main reason. Martikainen et al. (1995) and Fleming et al. (1996) argue that trading of an index is cheaper in derivate markets than in spot markets. Hence, new information updates faster there and show that these markets lead spot markets. Chen and Gau (2009) show that spot markets will contribute more to this price discovery when the bid/ask spread is smaller, and the minimum tick size decreases. Furthermore, Grünbichler et al. (1994) indicate that the trading mechanism is an important factor. They conclude that futures lead spot markets when they are screen-traded, since the price discovery speed increases.

Nam et al. (2008) do not recommend using low-frequency data, as this can lead to information loss and incorrect results. This recommendation will be followed in this thesis, as technology has improved and

lead-lag time has been reduced dramatically. Huth and Abergel (2012) use the model of Hayashi and Yoshida (2005), which uses the original tick data and does not require any modifications such as interpolation or resampling. They confirm that the most liquid assets tend to lead, especially in the setting of futures and stocks. Dao et al. (2018) use the same approach, extended by Hoffman et al. (2013), and include a study on the effect of information arrival on the lead-lag relationships. Their study concludes that unexpected trading volume affects the lead-lag relationships.

Brooks et al. (2001) use lead-lag relationships for accurate forecasting. However, trading on these results did not yield profits that outperformed the benchmark due to trading fees. This is furthermore confirmed in the high frequency environment, where both Huth and Abergel (2012) and later on Alsayed and McGroarty (2014) point toward trading fees and bid/ask spreads as limits to arbitrage.

2.2 CRYPTOCURRENCY

Only a few have studied the behavior of bitcoin on different cryptocurrency exchanges. Brandvold et al. (2015) use theory on information share to address the fraction of price discovery that happens on different bitcoin exchanges. They concluded that larger exchanges provide more information to the market and that smaller exchanges usually follow the market with a time lag. Bariviera (2017) and Phillip et al. (2018) found long memory in the bitcoin volatility. Catania and Sandholdt (2019) studied high frequency returns of bitcoin on two of the largest cryptocurrency exchanges. They present results of an intra-daily seasonality pattern and abnormal trade and volatility intensity close to the weekend. Furthermore, they found predictability for sample frequencies for up to six hours. Eross et al. (2017) found that volume, bid-ask spread, and volatility have n-shaped patterns throughout the day which suggests that European and North American traders are the main drivers of bitcoin trading and volatility. Urquhart (2016) concludes that bitcoin is in an inefficient market, but may be in the process of moving towards an efficient market. Nadarajah and Chu (2016) followed up on this study and showed that power-transformed bitcoin returns could be weakly efficient.

To my knowledge, this thesis contributes to the growing literature on bitcoin by being the first to study the high frequency lead-lag relationships between the most efficient cryptocurrency exchanges, and how information arrival affects these relationships.

3 THEORETICAL FRAMEWORK

This section will provide a detailed description of important topics addressed in this thesis. It is crucial for the reader to understand the concept of bitcoin and cryptocurrency exchanges, alongside theories on lead-lag relationships. Furthermore, efficient markets theory is included to touch upon the concept of arbitrage. This section will give the reader an essential fundament for understanding the analysis and discussions later in the thesis.

3.1 BLOCKCHAIN & BITCOIN

3.1.1 BLOCKCHAIN TECHNOLOGY

To understand the concept of Bitcoin, and cryptocurrency in general, some basic knowledge of the underlying technology is needed. During the financial crisis, when trust in businesses in the financial sector was at an all-time low, this new technology emerged. This solution made it possible to transact without the need for third-party intermediaries, proposed by an unknown actor with the pseudonym Satoshi Nakamoto (2008). This means that the typical approach of making transactions through a trusted middleman, like a bank, becomes theoretically unnecessary. A common definition describes blockchain as a distributed, decentralized, public ledger. Extremely simplified, one can imagine a chain of blocks. These blocks are digital information, stored in a public database that is the chain. As presented by Tasca & Tessone (2019), blockchain will be explained by looking at the fundamental principles; data decentralization, transparency, immutability, and privacy.

On a blockchain, the distributed nature of the network requires untrusted participants to reach a consensus. This kind of consensus is based on a set of rules. This can be what kind of transactions that are allowed, specifications on the block reward on mining difficulty, or details on transactions history that let participants review ownership of transactions. The distributed ledger updates when consensus is reached by the participants in the network. These participants are local nodes that independently verify transactions, making the process completely decentralized. In other words, due to this consensus mechanism, there is no need for a centralized third party. Transactions are done without a trusted authority who verify these or set the rules. The ledger of transactions is entirely open and accessible to a predefined set of participants. On public blockchains, like Bitcoin, everyone holds equal rights and ability to access. This means that anyone that is interested in the transactions on the blockchain can

go through this, whenever they want. This specification makes the blockchain totally transparent, and equally important, traceable.

Without elaborating too much on the technical details, the one-way cryptographic hash functions are essential to address. This is the base immutability. These hash functions take a variable-length input string (pre-image) and convert it to a fixed-length output string, a so-called hash value. This makes it exceptionally computationally challenging to calculate an alphanumeric text that has a given hash. Moreover, this makes the hash function collision-free: It is hard to generate two pre-images with the same hash value (Schneier, 1996). This means that the records on the blockchain are irreversible, indicating that recordings in the ledger are tamper-proof. When transactions are done, private keys possessed only by the sender are used to make signatures for the transactions. This makes the transaction tamper-proof since the signatures are being used to confirm that the transaction has come from the sender. By the specifications given in this section, blockchains are shared, tamper-proof, replicated ledgers where records are irreversible and cannot be forged (Tasca & Tessone, 2019) – making them relatively secure.

3.1.2 THE HISTORY OF BITCOIN

The Bitcoin blockchain follows the specifications of blockchain technology, and the cryptocurrency bitcoin is the unit of account, which describes the transactions on the Bitcoin blockchain. As mentioned in the introduction to this section, the Bitcoin blockchain was established in 2008. The next year, the first bitcoin block was mined and marked the birth of the cryptocurrency bitcoin. This included a text, which clearly expressed that the birth of Bitcoin was related to the financial crisis: “*The Times 03/Jan/2009 Chancellor on brink of second bailout for banks*”. This is the text from the front page of The Times, 3rd of January 2009 (Jenssen, 2019). The fact that the blockchain and the cryptocurrency have the same name can be confusing. This thesis will refer to the blockchain as “Bitcoin” and the cryptocurrency as “bitcoin”. That is, with and without a capital letter.

On the Bitcoin blockchain, new transactions are compiled into a new block every 10 minutes and sent to the Bitcoin network for confirmation. The blocks are secured based on a proof-of-work concept. This basically means that participants in the network, “miners”, utilize computational power to solve the hash related to the block. This will not be explained in detail due to the scope of this thesis, but Franco

(2015) can be used to get a deeper understanding. As this proof-of-work algorithm is both extensive and requires a lot of computational power, “miners” are rewarded with bitcoin for their effort. This reward has throughout the years systematically declined and will continue this way until the maximum amount of bitcoin has been mined. When the maximum amount of 21 million bitcoin are mined, the “miners” will only be rewarded through transaction fees (Jenssen, 2019).

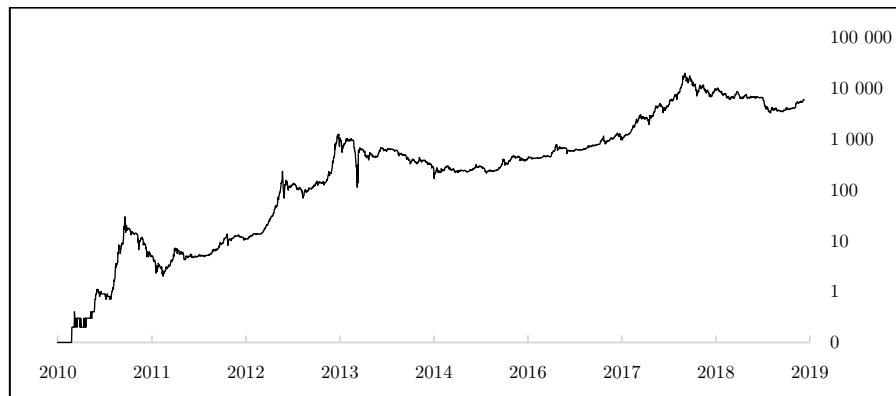


Figure 3.1 - Price chart of bitcoin (BTC) 2009-2019 on a log-scale, indicating the percentage change.

The price of one bitcoin has seen several periods with a significant increase. It increased rapidly at the end of 2010 and reached \$1 in February 2011. As seen in Figure 3.1, the price continued rising to over \$20 during the spring of 2011. The next rapid increase that came during the first months of 2013, was a surge from around \$10 to over \$150 was seen. Yet again, during the last months of 2013, the price increased 10 times from \$100 to over \$1,000. While falling down to just above \$200 in 2014 and most parts of 2015, the price started growing again in 2016. From the beginning of 2017, media really started covering cryptocurrency and the year saw the price rise from around \$1,000 to over \$19,000. As seen several times during bitcoin's lifetime, the price once again took a tumble in 2018 and ended the year just below \$4,000. As of May 2019, the price is just below \$8,000 (CoinMarketCap, 2019).

3.2 CRYPTOCURRENCY EXCHANGES

The function of cryptocurrency exchanges and how they work compared to traditional stock exchanges are essential to investigate. This section will explain some key differences and look at the development of cryptocurrency exchanges since bitcoin's birth. First of all, cryptocurrencies are normally traded on a large number of exchanges. Bitcoin, as the leading cryptocurrency, is traded on nearly 90 different exchanges ("Bitcoin Exchanges", 2019). This is quite different from stocks, which normally are traded on one exchange only. The fact that cryptocurrency markets never close is also a significant difference

from stock markets. Regular opening hours are generally during working hours and not during the weekends. This leads to news happening outside of trading hours, which is not the case in cryptocurrency markets. Furthermore, trading fees on cryptocurrency exchanges are normally high compared to stock markets. There is a common misunderstanding that trading cryptocurrency is cheap because they are easily transferable and globally accessible through blockchains. However, cryptocurrency exchanges have high fees and profit a lot on these fees (Babayan, 2019).

The process of opening stock trading accounts can be challenging, with thorough procedures. Many cryptocurrency exchanges are easy to access, without requirements. Although this could be favorable for users, it could come with a lot of risks. Regulations are on the rise in many countries, and anti-money laundering regulations with stricter KYC procedures are coming ("Regulation of Cryptocurrency Around the World", 2018). Furthermore, cryptocurrency typically has a limitation on supply. Bitcoin is capped at 21 million and will not be adjusted. Stocks do not have this kind of restriction imposed, and a company can at any time issue more stocks. Moreover, the unregulated nature of cryptocurrency makes market manipulation possible. Markets often face low liquidity, and so-called "pump and dump" schemes are a well-known problem (Xu & Livshits, 2018). Several exchanges have also experienced brutal hacks, resulting in billions of dollars stolen. One of the most used cryptocurrency exchanges in the history of bitcoin, Mt. Gox, was hacked in 2014. This will be addressed later in this section. Clearly, the risk of cryptocurrency trading is different from stock trading. However, the market of cryptocurrencies is still young, and risk, volatility, and varying liquidity is a natural consequence.

Lastly, cryptocurrency exchanges typically have a slight price mismatch. Several models for valuing both bitcoin and other cryptocurrencies have been made. There are different opinions on the fundamental values and on the fact that speculation is the only driver of the prices. This is outside the scope of this thesis and will not be addressed further. With a variety of specifications that may affect the bitcoin price on different cryptocurrency exchanges, this leads to clear price differences. The size of the exchange, trading volume and the currency pairs that the cryptocurrency is traded against, are some factors. The price differences are usually just a couple of dollars on the most liquid bitcoin exchanges, which is around 0.1%-0.2%. However, some exchange prices are several hundred dollars away from the mean market price, normally due to other currency pairings to bitcoin, low liquidity or capital restrictions ("Bitcoin Markets Arbitrage Table", 2019).

As already mentioned, Mt. Gox was the market leading cryptocurrency exchange for many years. During the years 2010, 2011 and 2012, the exchange had a market share of more than 80%. This started decreasing as the bitcoin price surged in the first quarter of 2013. Diversification of the exchange market started, and several exchanges appeared. These specialized in different traditional currencies and other cryptocurrency pairings to bitcoin. During 2013, actors like Bitfinex and Bitstamp, which are well known today, started taking market shares from Mt. Gox. Already during the last half of 2013, Mt. Gox users faced issues with withdrawals. On February 7th, 2014, Mt. Gox halted all bitcoin withdrawals. The exchange filed for bankruptcy only a couple of weeks later, after news and rumors emerged. The exchange allegedly lost about 850,000 bitcoins in a hack, corresponding to over \$450 million at that time (Dougherty & Huang, 2014). Since this, the exchange market has exploded. There are now over 200 exchanges and over 2100 cryptocurrencies, and bitcoin's dominance in market capitalization has decreased from over 90% in 2014 to around 50% as of March 2019 (CoinMarketCap, 2019).

3.3 EFFICIENT MARKETS

The Efficient Market Hypothesis was formulated by Fama (1970). This hypothesis is based on the fact that stock prices follow a random walk, which Fama (1965) was unable to reject in a study on serial correlation conducted on American stock prices. When trying to generate higher profit by the use of mechanical trading strategies, results showed that these did not outperform a standard buy-and-hold strategy. This led the way for the Efficient Market Hypothesis. The hypothesis states that if successive stock prices truly are independent, this will suggest that stock prices are efficient and absorb and reflect all available information as it reaches the markets. This is also true for financial securities markets where stocks are the underlying asset. This means that the price of the security should be an unbiased estimate of the underlying asset, thanks to the information efficiency (Fama, 1970).

The Efficient Market Hypothesis can be divided into three different levels that describe how efficient the market is. These three levels are based on assumptions on how much information that is available and reflected in prices of traded assets. The weak form of market efficiency describes a situation where all historical information is reflected in the current price. The semi-strong form means that prices reflect all historical information and all publicly available information. Lastly, the strong form of market efficiency should reflect private information in addition to public and historical information (Fama, 1970). As the Efficient Market Hypothesis implies that stocks and financial securities are traded at a fair price, there should not be any possibility for arbitrage in efficient markets. This suggests that there

is no possibility for excess return in the market, without taking additional risk. As there is no arbitrage opportunity, the law of one price occur. This states that the price of an identical security traded anywhere should have the same price. If not, an arbitrageur could purchase the security cheaper in one market and sell it in the market where the price is higher to make a profit. When the law of one price does not hold, the arbitrageur will use this opportunity until the price converges across markets (McDonald et al. 2006). However, this law does not always hold in practice, as there is often significant transaction cost, barriers to trade and other trade restrictions that apply. This is clearly seen in the cryptocurrency market. As described earlier, the price of bitcoin is not consistent, and some exchanges deviate in price with several hundred dollars. As the price of one bitcoin does not have any direct way to be calculated, except the amount the next buyer is willing to pay, it is difficult to maintain a symmetric market at all the different exchanges. Furthermore, barriers to enter exchanges, trading fees, transaction time, and other relevant factors will affect the price on each exchange. Some people say that these price differences create arbitrage opportunities across exchanges. However, time is a relevant factor, as moving bitcoins across exchanges is often quite slow. This time risk is especially relevant in this market, where the bitcoin price is very volatile.

3.4 STATIONARITY

For the analysis in this thesis, it is crucial to be familiar with the concept of stationarity. Hence, the difference between stationary and non-stationary time series, in addition to strict and weak stationarity, will be explained in this section. Brooks (2008) defines a time series as strictly stationary if the probability of its values does not change over time:

$$f(y_t, y_{t+1}, \dots, y_T) = f(y_{t+k}, y_{t+1+k}, \dots, y_{T+k})$$

This strict kind of stationarity suggests that all higher-order moments are constant, including mean and variance. Nevertheless, these kinds of time series are rarely found, and strict stationarity is therefore not common. It is more common to use the concept of weak stationarity. When a time series shows constant mean, variance and autocovariance over time, it is said to be a weakly stationary process. These process assumptions are sufficient to call a time series stationary. A time series is called non-stationary if its properties change over time. The variance of a non-stationary process will increase as the sample size moves toward infinity (Enders, 2008). A simple autoregressive process can be used to explain stationarity:

$$y_t = \phi y_{t-1} + u_t$$

This model shows that the variable y_t have no drift and depends on the lagged value y_{t-1} and the error term u_t . The value of ϕ indicates of the time series process is stationary or non-stationary. Three possible values of ϕ can occur (Brooks, 2008):

1) $\phi < 1 \Rightarrow \phi^T \rightarrow 0 \text{ as } T \rightarrow \infty$

In this case, a shock to the system is temporary and will gradually die away. This is called a **stationary** process.

2) $\phi = 1 \Rightarrow \phi^T = 1 \forall T$

In this case, shocks persist in the system and never die away. This means that the current value of y is just an infinite value sum of past shocks, in addition to the starting value y_0 . This case is also called the unit root case and is regarded **non-stationary**, since the variable y contains a unit root.

3) $\phi > 1$

In this case, the shocks will become more influential as the time series moves on, because if $\phi > 1$, then $\phi^3 > \phi^2 > \phi$, etc. This is also a **non-stationary** process and is called the explosive case and is not common. Hence, $\phi = 1$ is normally used to describe non-stationary.

Figure 3.2 shows two different processes. This illustration gives a better overview of the concept of stationarity. The time series to the left shows a non-stationary I(1) process, with a non-zero mean which indicates that $\phi = 1$. The time series to the right shows a stationary process, hence $\phi < 1$. The time series which is stationary return to its mean value throughout the time period.

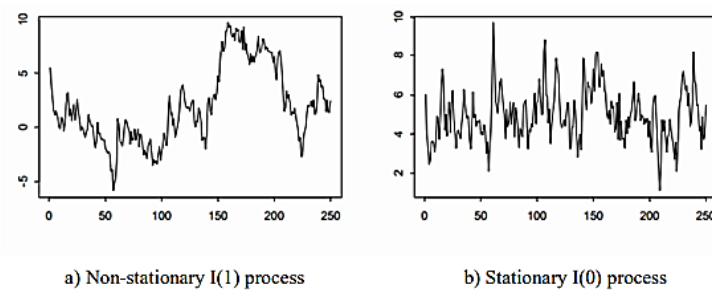


Figure 3.2 - Non-stationary / Stationary process

When working with time series, it is important to be aware of non-stationarity. False regression result can occur in a model with non-stationary variables, which can be misleading in an analysis. The R^2 , which shows how much of the variation in the dependent variable that can be explained by the independent variables, can be unusually high. This high value of R^2 indicates a relationship between two independent, random variables, when there in reality is no significant relationships between these

two (Granger & Newbold, 1974). A time series that is non-stationary, needs to be differenced d times before it becomes stationary. After that, the time series is said to be integrated of order d . The number of unit roots, i.e. the order of integration, decides the number of differences (d) to make the time series stationary. This can be written as $y_t \sim I(d)$, where $d \geq 1$. Moreover, a stationary time series can be written $y_t \sim I(0)$, since it is integrated by order 0 (Brooks, 2008).

3.5 LEAD-LAG RELATIONSHIP

As the efficiency of financial markets can be questioned, several studies have been conducted on the lead-lag relationships of different financial securities and assets. Section 3.3 explained the different levels of market efficiency. If some markets are more efficient than others and absorb and reflect available information faster, it should theoretically be possible to find leading and lagging price movements between markets. This section will describe some of the different theories that explore these relationships.

3.5.1 COINTEGRATION

Cointegration was first presented by Engle and Granger (1987) and is often confused with correlation. Correlation is perhaps the term people are most familiar with and measures how well two variables move in tandem with each other. That is, a measure between -1 and 1 which determine if the variable move in tandem in the same direction or opposite directions. A positive correlation means that the variables move in the same direction, and negative correlation suggests that the variables move in opposite directions. Studies show that highly correlated variables often are cointegrated as well, but this is not always the case. This is because cointegration measures whether the difference between the variables' means remains constant, and not how well they move together (Kammers, 2017). More specifically, Engle and Granger (1987) define cointegration as a shared stochastic trend in the long-run between two variables. Given two non-stationary variables $\{x, y\}$ that is integrated of order one (i.e. $\{x, y\} \sim I(1)$), and there is a linear combination of the two variables that is stationary ($I(0)$), these variables are said to be cointegrated (Brooks, 2008). A regression model of two non-stationary $I(1)$ variables y_t and x_t can be written as:

$$y_t = \mu + \beta x_t + \mu_t$$

Given the residuals $\mu_t = y_t - \beta x_t$ the variables y_t and x_t are said to be cointegrated if this error term is stationary, $\mu_t \sim I(0)$.

Theoretically, cointegration should only exist between variables that have a true relationship (Asteriou & Hall, 2007). This is the case for financial products like stocks and futures, that are based on the same underlying asset. In addition, bitcoin which is traded on several exchanges should indeed follow the theoretical concept of cointegration. Engle and Granger (1987) were the first to establish models for testing the relationship between variables that are cointegrated, including error correction models (ECM). These ECM's use both the lagged and the first-differenced levels of the variables. Hence, the models explore both the short-term relation and the long-term relation between the cointegrated variables (Brooks, 2008). Section 4.7 will give a detailed explanation of how cointegration can be tested.

3.5.2 GRANGER CAUSALITY

Already in 1969, Granger presented the concept of Granger causality. This is used to established causality between two variables and empirically testing the direction of this causality when it is assumed that two variables are related. The direction of the causality can be unidirectional or bidirectional. When two variables are unidirectional, one of the variables are said to Granger cause the other. More specifically, variable x_t is said to Granger cause y_t if it can be shown that lagged values of x_t will improve the forecast of y_t . That is, lagged values of x_t will provide statistically significant information about future values of y_t . Furthermore, the word causality should not be interpreted as how movements of one variable cause the movement of another variable. Causality refers to a correlation between the current values of one variable and the lagged values of another variable (Brooks, 2008). Figure 3.3 illustrates an example where time series x_t Granger causes time series y_t . The patterns in x_t are approximately repeated in y_t after a given time lag, which is shown by the arrows. Hence, past values of x_t can be used for the prediction of future values of y_t (Liu & Bahadori, 2012).

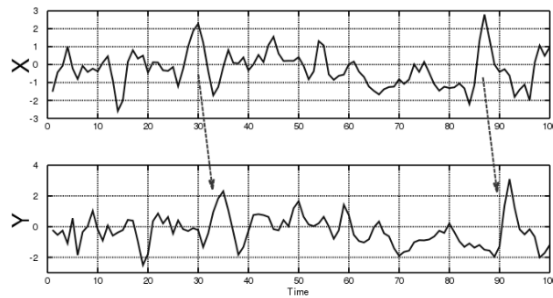


Figure 3.3 - Illustration of Granger causality, regenerated from Liu & Bahadori (2012).

As this thesis look at the same asset at different exchanges, Granger causality between these time series will be examined. The method for this test will be presented in detail in Section 4.7.4.

3.5.3 HIGH FREQUENCY HAYASHI-YOSHIDA CROSS-CORRELATION

As the technology of financial markets evolves, new approaches and theories are established. As a result of advanced computational power in recent decades, data can be collected at an extremely fine scale. When data is available down to milliseconds, new approaches and challenges occur when dealing with time series and econometrics in general (Engle, 2000). This section will dive into the details of a more modern approach to lead-lag relationships. Hayashi & Yoshida (2005) were the first to present this method, which focuses on data collected “tick-by-tick”, and not time intervals. This means that the data is not divided into time intervals of seconds, minutes or hours, but includes every single observation. To put this in perspective; high frequency data collected from one day in a liquid market can equal the amount of daily data collected for over 30 years (Dacorogna, 2001).

Hayashi & Yoshida (2005) describes the process of estimating the covariance of two diffusion processes when they are observed only at discrete times in a non-synchronous manner. This can, for example, be a stock price or foreign exchange rate, where the variable changes on a random and continuous basis. Previous studies use a popular approach of realized covariance estimation, which is based on regular spaced, synchronous data. Andersen et al. (2001) look at realized exchange rate volatility and approximate the quadratic variation and covariation directly from high frequency data. Basically, they use quadratic variations as estimators of variances and covariances of multivariate security price processes. These discrete observations of two security prices, $(P_{t_i}^1 \text{ and } P_{t_i}^2)_{i=0,1,\dots,m}$ of size $m + 1$, are Itô processes, or so-called continuous time Itô semi-martingales where $0 \leq t \leq T$. The covariation $V := \langle P^1, P^2 \rangle_T$ of the two processes, also called the realized covariance estimator, is then defined in the following way by Andersen et al. (2001):

$$V_{\pi(m)} := \sum_{i=1}^m (P_{t_i}^1 - P_{t_{i-1}}^1)(P_{t_i}^2 - P_{t_{i-1}}^2)$$

Usually, equally spacing is chosen, so that $t_i - t_{i-1} = \frac{T}{m}$ ($=: h$) for each i .

Hayashi & Yoshida (2005) point out two important implication of this realized covariance estimator. First of all, high frequency transaction data are recorded at random times. This means that two transaction prices are rarely observed at the same specific time, which the formula above assume. Secondly, some parts of the original transaction data will be missing if prespecified grid points are set, due to these random transaction times. Prespecified grid points can be time intervals of length h , i.e.,

second or minutes, which is necessary when using the formula given above. Hence, imputation or interpolation of the missing observation in these prespecified time intervals. This is problematic according to Hayashi & Yoshida (2005), because regular time intervals and data interpolation schemes may lead to unreliable estimations. In addition, this realized covariance estimator depends heavily on the choice of time interval length, h . Hence, they propose a new approach that is free from any adjustments of the original data. This leads to data that is free of bias and other problems related to synchronization processes. The Hayashi-Yoshida covariance estimator will be explained in detail in Section 4.8.

3.5.3.1 THE EPPS EFFECT

In relation to high frequency data, it is relevant to include a phenomenon called the Epps effect. Epps (1979) reports empirical evidence of how sampling frequency on high frequency stock returns will affect correlation estimators. He discovered that the empirical correlation between the returns of two different stock decreases with the length of the interval for which price changes are measured (Epps, 1979). That is, the correlation decreases when the sampling frequency increases. This effect has been studied for decades, and considerable effort has been used to explain the phenomenon found by Epps. Several factors have been found to explain the effect. However, the most important factor is related to the non-synchronicity of time series, as is present for raw transaction data. Empirical results show that data where only synchronous ticks are included, i.e., equal time intervals for two variables, clearly reduces the degree of the Epps effect (Toth & Kertesz, 2009). This effect is highly relevant for this thesis, as the method of Hayashi & Yoshida (2005) deals with high frequency data. Hence, the estimation of the true lead-lag relationships will be more accurate, but this will clearly reduce the true correlation.

3.5.4 DISCUSSION

Section 2.1 presented a range of different studies on lead-lag relationships. It would be optimal to study several approaches to these relationships. However, this thesis narrows down to two main approaches. When choosing which approaches to focus on, several aspects have been considered. Presented in this theory section, are approached that are normally used on different kind of datasets. That is, we operate with both high frequency data and data sorted in low frequency time intervals. By using a general and well-known approach to cointegration with the use of Granger causality, a good overview of the empirical facts is established. This will provide indications on how the bitcoin price behaves on the different exchanges. Not necessarily detailed results that can be used in a potential trading strategy,

but helpful results when a decision should be taken on what interesting aspects that should be explored further. If Granger causality is found between certain exchanges, further analysis can then be done with the above-mentioned theory of Hayashi and Yoshida (2005). This approach focuses on high frequency transaction data, and would hopefully be able to provide new results in the study of the lead-lag relationship between certain exchanges.

When using such different approaches, interesting results will hopefully occur. This could be related to the fact that high frequency data reveal totally different lead-lag relationships than data frequencies of for example minutes. Contrary, this could be results that strengthen the already found results. However, the point of this discussion is to clarify the use of more than one approach to lead-lag relationships. Interesting observations can be done with a classic cointegration approach, and the Hayashi-Yoshida estimator is a tool to identify relationships that a traditional approach cannot do. This way, the complete analysis will give a solid foundation to make a conclusion about the lead-lag relationships of the bitcoin price on different exchanges.

3.6 INFORMATION ARRIVAL

Lead-lag relationships have been studied for a long time, and that also includes research on what factors that affects this relationship. Earlier, studies looked at trading costs and trading mechanisms of the financial instruments that showed lead-lag relationships. Research shows that trading cost is important since information, i.e., the price, is updated faster where trading is cheaper. The fact that trading of an index is cheaper in a derivate market than in a spot market suggests that information arrives earlier in the derivate market. Hence, the price is updated in the derivate market before the spot market (Martikainen & Perttunen, 1995). Furthermore, Fleming et al. (1996) show that there is no lead-lag relationship between put and call options because they have the same type of trading cost structure.

When evaluating the trading mechanism of the financial instruments, another interesting aspect is found. An example can be a situation where two financial instruments are both floor-traded. Floor-trading is based on the physical interaction between people who buy and sell, using a so-called open outcry method. This kind of trading involves people who shout and use hand signals to transfer information about buy and sell orders (Shell, 2007). If one of the two financial instruments change trading mechanism to a more modern approach, this clearly influences the lead-lag relationship. When

a leading financial instrument improves from being floor-traded to being screen-traded, the lead-lag relationships strengthens significantly (Grünbichler et al. 1994). This is an intuitive observation since the speed of information arrival is on an entirely different level when the financial instruments are screen-traded.

The two factors that affect the lead-lag relationship mentioned above are not that relevant for the research in this thesis. When the price of bitcoin is evaluated, this only includes electronically traded prices and spot instruments. Hence, no derivatives are considered, nor different trading mechanisms. Naturally, one would assume the same level of trading costs across similarly spot prices on different exchanges. The cryptocurrency markets and exchanges operate with differences in trading costs. However, these are minimal and vary among individual traders. Hence, it will not be included when the effect of information arrival on the lead-lag relationships are analyzed in Section 6.2.3.

3.6.1 TRADING VOLUME

Dao et al. (2018) studied the information flow in trading volume to the market and the lead-lag relationship between high frequency spot instruments. They base this on the fact that the lead-lag relationship exists because some instruments reflect information faster than others, and the fact that information is important in financial markets. Trading volume has been shown to explain some aspects of information arrival to financial markets (Arago & Nieto, 2005). In addition, traders that provide volume to the market are not alike. One can differentiate so-called sophisticated investors from non-sophisticated investors. Sophisticated investors can also be called institutional investors, and they usually provide trades of a large number of shares or value (Madhavan & Sofianos, 1998).

Furthermore, this form of information arrival can be even more precisely explained when analyzing what affect the lead-lag relationship. Arago & Nieto (2005) point out that the expected and unexpected volume flow to the market should be researched. Here, the expected volume is said to capture the normal level of market activity, and the unexpected volume captures the arrival of new information to the market. Trading volume is easily observable across cryptocurrency exchanges, and the theory in this section shows its importance in relation to the lead-lag relationship. Thus, this will be elaborated on later in the thesis when the effect of information arrival is analyzed in Section 6.2.3.

4 METHODOLOGY

This section elaborates on how the research is conducted and how the data that is collected, in addition to the approaches used in the analysis. Most data handling, calculations and testing have been done with the Anaconda Spyder software, using the programming language Python. Several open-source Python packages have been downloaded, and Arcane Crypto AS has contributed with both server capacity and development of certain codes.

4.1 RESEARCH PHILOSOPHY AND APPROACH

This thesis seeks to explore the high frequency lead-lag relationships between bitcoin prices on different exchanges. Specifically, the purpose is to investigate the correlations in the increasingly liquid bitcoin market. However, this relatively new asset class does not show the same liquidity levels as traditional financial markets. This leads to the uncertainty of efficient markets and raises questions about arbitrage opportunities. Evaluating the price movements across several cryptocurrency exchanges, allow a deeper understanding of the lead-lag relationships.

4.2 RESEARCH DESIGN

Due to the aim of this thesis and inadequate existing academic literature on high frequency lead-lag relationships of the bitcoin price, an exploratory research design is applied. Although existing academic literature on lead-lag relationships is intuitively applicable for this thesis, this will to my knowledge be the first study on bitcoin. Hence, an exploratory design is suitable, since there are no earlier studies to rely upon to predict an outcome (USC Libraries, 2019). Due to the quantitative nature of this thesis, numerical data is a favorable fit.

4.3 DATA COLLECTION

The data collection for this thesis is somewhat challenging, due to the size of the datasets. The programming language Python is favorable when handling the data, as more basic programs like Excel does not provide the necessary tools and have certain limits related to data capacity. The data is collected through an Application Programming Interface, also called an API. This software intermediary let two applications talk to each other. The use of an API basically means that an

application connects to the cryptocurrency exchanges and sends trade data to a server. This happens continuously, and all trades made are therefore saved and available. This raw data includes transaction prices of every trade and the volume of the trade, specified in bitcoin. Not all cryptocurrency exchanges offer the use of API on historical data. In that scenario, the website Bitcoincharts.com is used to download data, which is the world's leading provider for financial and technical data related to the Bitcoin network ("Markets API", n.d.).

4.4 DATA PREPARATION

Usually, downloaded data is specified in time intervals or have to be transformed in these intervals. This means that price data is grouped into candles, normally 1-minute candles or larger. However, due to the specifications of the Hayashi-Yoshida estimator, it is possible to use the tick data directly. Hence, data files with all transactions will not be adjusted when this estimator is applied. For other theoretical approaches like the classic causality approach that will be explained in Section 4.7, the data is aggregated to minute-candles. This is also the case for the datasets used in the regression analysis in Section 6.2.3, which operates with daily candles. In addition, information on opening, closing, average, highest, and lowest price in the interval, can be included.

4.5 RESEARCH QUALITY

This thesis is solely based on secondary data sources since all trade data is historical information available for download. This differentiates from primary data, which is collected directly from the researcher. This could be through surveys, interviews or direct observations. It is important to consider the advantage and disadvantage of secondary data. According to Saunders et al. (2016), secondary data is more straightforward to collect than primary data. Hence, less effort in data collection is needed. This can be allocated to other parts of the process, which includes analysis and interpretation. Secondary data provide the opportunity to compare and give context, which means that the result can be placed in a more general context. Furthermore, as secondary data is more accessible to the public, it can easier be checked by others. There are also some disadvantages of secondary data. It can be collected for other purposes than what is intended use by yourself. This can make lead to data that is not applicable for answering your research questions. However, for this thesis, data collected is raw trade data, which makes this disadvantage irrelevant. It can also be challenging to judge the reliability

and credibility of, e.g. online sources. This can affect the quality of the data, which will be addressed in Section 5.3.

4.6 LIMITATIONS AND WEAKNESSES

The datasets for this thesis are subjectively chosen. That is, both the period that the datasets are collected from and the selected cryptocurrency exchanges. As the thesis only use data from 2018, this can lead to results that miss out on essential trends or findings on lead-lag relationships, both before 2018 and the first part of 2019. The cryptocurrency market is young and evolving every day with new exchanges and investors. The last couple of years have been quite different in terms of price movements, and the subjective selection of the time period in this thesis will probably impact the result. In addition, the seven cryptocurrency exchanges could reveal relationships that are not presented in the rest of the market. However, the selection of exchanges is based on a number of factors, including trustable trading volume reporting, security aspects and popularity among investors. The selection should provide a meaningful result that can be used to conclude on the overall market. Nevertheless, it is important to be aware of these kinds of limitations and weaknesses that can affect the overall conclusion.

4.7 CAUSALITY APPROACH

Throughout the last decades, several approaches to lead-lag relationships have been used to explore this area. As discussed in the theory section, this thesis will go into detail on two different approaches. One of the well-known approaches based on cointegration relationships and Granger causality will now be explained and will be addressed as the causality approach.

4.7.1 TESTING FOR STATIONARITY

As described in Section 3.4, it is important to determine if the time series in the datasets are stationary or not. Misleading result can occur if non-stationary time series are included. The most used approach when testing for stationary is the Dickey-Fuller test (Dickey & Fuller, 1979).

4.7.1.1 THE DICKEY-FULLER TEST

The objective of the Dickey-Fuller (DF) test is to test if the null hypothesis that $\phi = 1$ in the equation:

$$y_t = \phi y_{t-1} + u_t$$

holds against the one-sided alternative $\phi < 1$. This gives the following hypotheses:

H_0 : The time series contains a unit root

H_1 : The times series is stationary

A common approach is to difference the time series before applying the DF test. The equation above is then transferred into the first difference by subtracting y_{t-1} from both sides, for ease of computation and interpretation (Brooks, 2008):

$$y_t - y_{t-1} = \phi y_{t-1} - y_{t-1} + u_t$$

$$\Delta y_t = (\phi - 1)y_{t-1} + u_t$$

$$\Delta y_t = \psi y_{t-1} + u_t$$

The equation above now includes ψ , which is equal to $\phi - 1$. This means that the null hypothesis presented above is equivalent to a test of $\psi = 0$, against the alternative that $\psi < 1$. Moreover, if $\psi = 1$, this represent a time series that follows a pure random walk where the lagged value of the variable has no influence on the value at time t (Brooks, 2008). Two alternative equations for the Dickey-Fuller test can also be used. These models include an intercept and linear trend to the equation presented above, and are presented below (Dickey & Fuller, 1979):

$$\Delta y_t = \alpha + \psi y_{t-1} + u_t$$

$$\Delta y_t = \alpha + \beta_t + \psi y_{t-1} + u_t$$

The first of the two equations represent a time series variable that is a random walk with a drift. This is often seen for macroeconomic variables and includes a predictable trend as well as the stochastic unpredictable trend (Asteriou & Hall, 2007). The second equation represents a time series variable where a time trend variable is included as an independent variable. Hence, stationarity is ensured around the trend, meaning that the stochastic part of the process will disappear over time instead of adding to the deterministic trend. Before applying a DF test to a times series, it is necessary to decide which equation to use. Using the wrong equation can lead to biased results of the test due to different levels of critical values. A visual examination can be done to evaluate the need for an intercept, a linear trend, or both. The Dickey-Fuller test statistic is defined as:

$$\tau = \frac{\hat{\psi}}{SE(\hat{\psi})}.$$

Here, $\hat{\psi}$ is the estimated coefficient from OLS and $SE(\hat{\psi})$ is the standard error of the coefficient $\hat{\psi}$. This test statistic follows a non-standard distribution that is skewed to the left, and not the usual t-distribution under the null hypothesis (Dickey & Fuller, 1979). This is due to the non-stationarity of the null and leads to special DF critical values used for comparison with the computed test statistics. These values are much bigger than standard normal critical values, in absolute terms. Hence, more evidence against the null hypothesis of the DF test is required to reject it, compare to a standard t-test (Brooks, 2008).

4.7.1.2 THE AUGMENTED DICKEY-FULLER TEST

The test presented above is based on the fact that the error term u_t is a white noise process, i.e. $u_t \sim IID(0, \sigma^2)$. However, this is normally not the case and an extended equation is presented called the Augmented Dickey-Fuller test (Brooks, 2008):

$$\Delta y_t = \psi y_{t-1} + \sum_{i=1}^p \alpha_i \Delta y_{t-1} + u_t$$

This equation includes p lags of the dependent variable to account for the autocorrelation of the dependent variable, and hence in the error term u_t as well. This inclusion of lags absorbs the autocorrelation and ensure an error term that is a white noise process. This extended test follows the same hypotheses, test statistics and critical values as the standard DF test. However, this test leads to a new problem as the optimal number of lags needs to be chosen. If more lags than necessary are included, this can lead to more type II errors. That is, accepting a false null hypothesis. As more lags are included, the degrees of freedom will be affected and cause the absolute value of the test statistics to decrease. On the other hand, choosing too few lags will lead to more type I errors as some of the autocorrelation is left in the model. That is, rejecting a true null hypothesis. To determine the right number of lags, it is recommended to use the number of lags that minimizes an information criterion, normally through the Schwartz Bayesian Information Criteria (SBIC) or the Akaike Information Criterion (AIC) (Brooks, 2008). The number of lags that minimizes the AIC and SBIC values is chosen as the optimal lag length. SBIC is more consistent, but inefficient, and the AIC is not consistent, but more efficient. Hence, no criterion is definitely superior to the other (Brooks, 2008). In this thesis SBIC will be used, and can be written as:

$$SBIC = \ln|\hat{\sigma}^2| + \frac{k}{T}(\ln T)$$

In this equation, $\hat{\sigma}^2$ the residual variance, T is the number of observations, and k is the number of parameters estimated.

4.7.2 VECTOR AUTOREGRESSIVE MODEL

A vector autoregressive model (VAR) is a systems regression model, which indicates that there is more than one dependent variable. This is considered to be a “hybrid” between univariate time series models and simultaneous equations models. The simplest case in a bivariate VAR, where only two variables are included. The current values of these two variables, y_{1t} and y_{2t} , depends on different combinations of the previous k values of both variables and their error terms (Brooks, 2008). This can be written as:

$$y_{1t} = \beta_{10} + \beta_{11}y_{1t-1} + \cdots + \beta_{1k}y_{1t-k} + \alpha_{11}y_{2t-1} + \cdots + \alpha_{1k}y_{2t-k} + u_{1t}$$

$$y_{2t} = \beta_{20} + \beta_{21}y_{2t-1} + \cdots + \beta_{2k}y_{2t-k} + \alpha_{21}y_{1t-1} + \cdots + \alpha_{2k}y_{1t-k} + u_{2t}$$

Where u_{it} is the white noise disturbance term with $E(u_{it}) = 0, (i = 1, 2), E(u_{1t}, u_{2t}) = 0$.

A VAR can be extended to include several variables, where each have an equation like the ones presented above. When a model includes a set of g variables with k lags of each variable, a general notation can be used:

$$\begin{array}{ccccccc} y_t = & \beta_0 & + & \beta_1 y_{t-1} & + & \beta_2 y_{t-2} & + \cdots + \beta_k y_{t-k} & + & u_t \\ g \times 1 & g \times 1 & & g \times g & g \times 1 & & g \times g & g \times 1 & g \times 1 \end{array}$$

This model can also be extended to a model that includes first difference terms and cointegrating relationship, called a vector error correction model (VECM).

4.7.3 JOHANSEN CO-INTEGRATION TEST

There are several methods for testing cointegration between two or more variables. The most common method is presented by Engle & Granger (1987) and is called the Engle and Granger two-step method. However, this is a univariate approach, which means that it only analysis pairwise relationships between variables. In this thesis, the focus will be on a multivariate approach presented by Johansen (1988). This test explores cointegration in a system of more than two variables. That is, a more general approach to cointegration. This is fitting for this thesis as time series from several cryptocurrency exchanges will be included in the analysis, and the main focus will be on short-run dependence with

the Hayashi-Yoshida cross-correlation estimator. Hence, the cointegration test will be used as an introducing test, and another approach will be used to determine the detailed short-run relationships between the time series from different exchanges. In addition, the Johansen method improves some of the drawbacks with the Engle-Granger method. The multivariate test allows all variables to be endogenous, and the Engle-Granger method only allows one endogenous and one exogenous variable. Moreover, the Johansen method makes it possible to provide results that explain all the cointegration relationships between variables (Brooks, 2008).

The Johansen-test is based on a vector autoregressive model (VAR). Furthermore, this VAR needs to be transformed into a Vector Error Correction Model (VECM) which was briefly touched upon in the previous section. A VECM can be written as (Brooks, 2008):

$$\Delta y_t = \Pi y_{t-k} + \Gamma_1 \Delta y_{t-1} + \Gamma_2 \Delta y_{t-2} + \dots + \Gamma_{k-1} \Delta y_{t-(k-1)} + u_t$$

where

$$\Pi = (\sum_{i=1}^k \beta_i) - I_g \text{ and } \Gamma_i = (\sum_{j=1}^i \beta_j) - I_g$$

The VAR presented above is now a set of g variables in first differenced form on the left-hand side, and $k - 1$ lags of the dependent variables (differences) on the right-hand side, each with a Γ coefficient matrix attached to it. This test focuses on the Π matrix, which can be understood as a long-run coefficient matrix. This is due to the fact that when in equilibrium, all the values of Δy_{t-1} will be zero. In addition to this, when the error terms u_t are set to their expected value of zero, this will result in $\Pi y_{t-k} = 0$ (Brooks, 2008). The Johansen test uses the eigenvalues to determine a rank of the Π matrix. This rank, r , will be equal to the number of eigenvalues that are significantly different from zero. Eigenvalues, λ , are also called characteristic roots, and will indicate the number of cointegrated vectors in the system of variables. Furthermore, the test result can result in three different cases when assessing the rank of the Π matrix (Johansen & Juselius, 1990):

- $r = g$

This case is a **full rank**, as all eigenvalues are significantly different from zero. This indicates that the variables in the system are all stationary, and no cointegration is possible.

- $r = 0$

In this case the **rank is zero**, as no eigenvalues are significantly different from zero. This also indicates no cointegration, since there are no linear combinations of the variables in the system that are $I(0)$, i.e. stationary processes.

- $0 < r < g$

This case has a **reduced rank**, as some of the eigenvalues are significantly different from zero.

This indicates that there are r linear combinations of variables in the system that are $I(0)$.

Hence, cointegration exist in the system, with r cointegrated variables.

There are two sets of test statistics under the Johansen approach (Brooks, 2008). These two are presented below, and are called the Trace test and the Maximum Eigenvalue test:

$$\lambda_{trace}(r) = -T \sum_{i=r+1}^g \ln(1 - \hat{\lambda}_i)$$

$$\lambda_{max}(r, r + 1) = -T \ln(1 - \hat{\lambda}_{r+1})$$

where r is the number of cointegrated vectors under the null hypothesis and $\hat{\lambda}_i$ is the estimated value of the i th ordered eigenvalue from the Π matrix.

The Trace test's null hypothesis states that the number of cointegrated vectors is less than or equal to r . This is testes against the alternative that there are more than r cointegrated vectors. Hence, the Trace test is a joint test. The other test statistic, the Maximum eigenvalue test, take each eigenvalue individually when testing. The null hypothesis states that the number of cointegrated vectors is precisely r , and is tested against the alternative $r + 1$ cointegrated vectors (Brooks, 2008). The distributions of these tests are non-standard and the critical values to be used depend on the value of $(g - r)$ (Johansen & Juselius, 1990). Furthermore, these critical values are sensitive to the choice of lags and the number of deterministic terms in the VAR. Hence, one should carefully determine the optimal lag length, and decide if a constant or a trend should be included.

4.7.4 WALD TEST FOR GRANGER CAUSALITY

The sections above will provide results on cointegration and determine if the variables have a long-run relationship in general. However, it would be useful with information about the short-run relationship. Variables might be related in the short-run, even if there is no sign of cointegration in the long-run. Thus, this section will provide an explanation of the Wald test that explores Granger causality between variables. As explained in the theory section, variable x_t is said to Granger cause y_t if it can be shown that lagged values of x_t will improve the forecast of y_t . The Wald test is a standard F-test, and a simple VAR can be considered:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_q x_{t-q} + u_t$$

Given significant α -coefficients of the lagged values of x , x is said to Granger cause y . Furthermore, this is tested for y as well, to determine if there is causality in both directions, i.e. bidirectional causality. The estimated VAR is used in the Granger causality test when restrictions are imposed with the following hypotheses (Brooks, 2008):

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0 \quad (x \text{ does not Granger cause } y)$$

$$H_1: \text{at least one of } \alpha_p \neq 0 \quad (x \text{ does Granger cause } y)$$

That is, testing the null hypothesis that the estimated coefficients α_p on the lagged values of x are jointly zero. If the results fail to reject the null hypothesis, x does not Granger cause y .

4.8 THE HAYASHI-YOSHIDA ESTIMATOR

The presented approach of causality in the previous section will only provide results on the fact that lead-lag relationships exist between the time series on different cryptocurrency exchanges. However, one would prefer a more comprehensive approach. This section will describe a modern approach through the Hayashi-Yoshida estimator used on two series of non-synchronous tick data (Hayashi & Yoshida, 2005). This will enable information on how strong the lead-lag relationship is, in addition to information about the time aspect of the relationship, i.e. at what time the relationship is strongest. The methodological approach will follow the work of Huth & Abergel (2012) and Hoffman et al. (2013).

As described in Section 3.5.3, this Hayashi-Yoshida (HY) cross-correlation estimator does not require any kind of data synchronization, which means that raw tick data of transactions is used. Hence, no data modifications in the form of interpolation or resampling at regular intervals are done. By doing this, potential biases are avoided (Huth & Abergel, 2012). The primary purpose of this HY approach is to calculate the correlation between one series and timestamp-adjusted version of another series. This will provide results on which time adjustment that will maximize the two series' correlation. Hayashi & Yoshida (2005) introduce this new estimator of the linear correlation coefficient by looking at two diffusive processes. Given two Itô processes, X and Y , such that:

$$\begin{aligned} dX_t &= \mu_t^X dt + \sigma_t^X dW_t^X \\ dY_t &= \mu_t^Y dt + \sigma_t^Y dW_t^Y \\ d\langle W^X, W^Y \rangle_t &= \rho_t dt \end{aligned}$$

The observations times, that must be independent for X and Y, are given by:

$$\begin{aligned} 0 = t_0 \leq t_1 \leq \dots \leq t_{n-1} \leq t_n &= T & \text{for } X \\ 0 = s_0 \leq s_1 \leq \dots \leq s_{m-1} \leq s_m &= T & \text{for } Y \end{aligned}$$

Then, an unbiased and consist estimator of the covariance, $\int_0^T \sigma_t^X \sigma_t^Y \rho_t dt$, between the series is given by the following equation:

$$C = \sum_{i,j} r_i^X r_j^Y 1_{\{O_{ij} \neq \emptyset\}}$$

where

$$\begin{aligned} r_i^X &= X_{t_i} - X_{t_{i-1}} \\ r_j^Y &= Y_{t_j} - Y_{s_{j-1}} \\ O_{ij} &=]t_{i-1}, t_i] \cap]s_{j-1}, s_j] \end{aligned}$$

This estimator is an unbiased and consistent estimator as the largest mesh size goes to zero, which is different from standard previous-tick correlation estimators. It may seem challenging to interpret the equation given for C. However, it only shows that the covariance is calculated by summing every product of increments as soon as they share any overlap of time (Huth & Abergel, 2012). Figure 4.1 below illustrates non-synchronous observations from two series and the interval between observations.

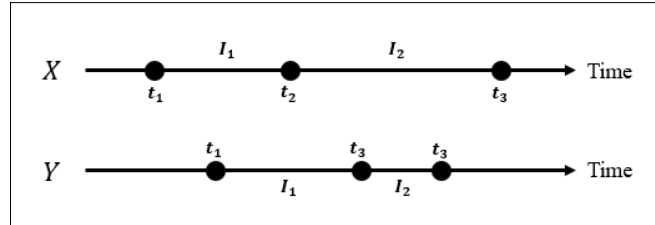


Figure 4.1 – Non-synchronous data. Each dot illustrates a data point, t is the arrival time of observations, and I is the interval between two consecutive observations

Looking at Figure 4.1, the covariance of the series X and Y is calculated by summing the products of the following pairs of return: $(r_{I_1}^X, r_{I_1}^Y)$, $(r_{I_2}^X, r_{I_1}^Y)$ and $(r_{I_2}^X, r_{I_2}^Y)$. Given constant volatilities and correlation, a consistent estimator for the cross-correlation coefficient ρ of X and Y is given by (Huth & Abergel, 2012):

$$\hat{\rho} = \frac{\sum_{i,j} r_i^X r_j^Y 1_{\{O_{ij} \neq \emptyset\}}}{\sqrt{\sum_i (r_i^X)^2 \sum_j (r_j^Y)^2}}$$

The next step is to adjust all timestamps of Y, to be able to allow for leads and lags, and re-estimate their correlation (Hoffmann et al., 2013). Figure 4.2 below shows the concept of timestamp-adjustment.

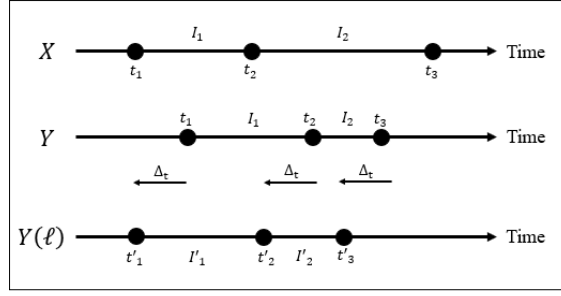


Figure 4.2 - Adjustment of timestamps.

$Y(\ell)$ is made by adjusting the observations of Y backward in time by the same amount Δt

If the series X is fixed and $Y(\ell)$ is the timestamp-adjusted series, the lagged cross-correlation coefficient $\rho(\ell)$ between the series is given by (Huth & Abergel, 2012):

$$\rho(\ell) = \frac{\sum_{i,j} r_i^X r_j^Y 1_{\{O_{ij}^\ell \neq \emptyset\}}}{\sqrt{\sum_i (r_i^X)^2 \sum_j (r_j^Y)^2}}$$

where

$$O_{ij}^\ell =]t_{i-1}, t_i] \cap]s_{j-1} - \ell, s_j - \ell]$$

The full cross-correlation function can now be computed by shifting all the timestamps of Y and then use the HY estimator given in the equation above. This will make it possible to decide if one of the series leads the other, by measuring the asymmetry of the cross-correlation between the positive and negative lags. The grid of lags will be given in seconds in this thesis, which means that Y will be moved backwards and forwards with 1-second lags. The analysis will be taken further by finding the lag where the maximum level of cross-correlation occurs. That is, at which lag the cross-correlation is highest. If this is zero, the cross-correlation of the series is highest when they are observed at the same time. Moreover, if this is positive X leads Y by this number of lags, and if this is negative Y leads X by this number of lags. The maximum correlation found is called the lead-lag correlation coefficient.

Furthermore, Huth & Abergel (2012) present an equation for calculating the relative strength of the lead-lag relationships, which is closely related to Granger Causality. If the measure is higher than one, X leads Y , and vice versa if the measure is below one. The equation, called the Lead-Lag Ratio, is presented below:

$$LLR = \frac{\sum_{i=1}^p \rho^2(\ell_i)}{\sum_{i=1}^p \rho^2(-\ell_i)} \quad (\ell_i > 0)$$

The numerator of the LLR is the sum of squared correlation coefficients at all lead of Y, and the denominator is the sum of squared correlation coefficients at all lags of Y.

To summarize, the Hayashi-Yoshida estimator will then give results on which of the two series that is leading the other, how strong this leadership is, and when this lead-lag relationship is strongest.

4.9 LINEAR REGRESSION

In order to evaluate which factors that affect the lead-lag relationship, this thesis will take us of linear regression. That is, estimating the linear relationship between the lead-lag variables and several independent variables. Both the dependent and the independent variables in the regression analysis will be presented in Section 6.2.3.

The regression analysis will be based on multiple linear regression. This regression method attempts to model the relationship between two or more independent variables and one dependent variable, by fitting a linear equation to observed data (Newbold et al., 2013). The multiple regression model is defined as:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

Where Y_i is the dependent variable (i.e. the lead-lag variable), β_0 denotes the intercept, β_k represents the slope and X_k is the chosen independent variable. The random error term ε_i indicates the variation in Y_i that is not estimated by the linear relationship (Newbold et al., 2013).

The coefficients in the regression model are typically computed with statistical software, where the Ordinary Least Square (OLS) estimator is used so that the estimated regression line is as close as possible to the observed data. Various statistical measures will be computed, including the R-squared (R^2). This measure shows the proportion of the variance in the dependent variable that is predictable from the independent variables. The level of R^2 can be in the range of 0-1, and a higher value indicates a more accurate regression model. In addition, the p-value of the model indicates the reliability of X_k to predict Y. This thesis will indicate if coefficients are significant at the 1%, 5% or 10% level when p-values are evaluated. That is, if the test statistic is greater than a given value, the null hypothesis that the chosen independent variables are equal to zero can be rejected.

4.9.1 ASSUMPTIONS

The multiple linear regression depends on five assumptions, which will be presented below.

Linearity

This assumption is based on the fact that there must be a linear relationship between the independent variable and the dependent variable. It is possible to test this assumption by the use of scatter plots with the residual values against the predicted values. One can say that the assumption of linearity is obtained when the observed points in the scatter plots are symmetrically distributed around the predicted regression line (Newbold et al., 2013).

Normality

This assumption test if the error terms are normally distributed. This is done by the use of the Jarque-Bera test (Newbold et al., 2013):

$$JB\ test = n \left(\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24} \right) \sim \chi^2$$

The test above is an adaption of the chi-square procedure and depends on the descriptive measures of skewness and kurtosis, in addition to the sample size. This normal distribution test relies on the skewness closeness to 0, and how close the kurtosis is to 3, where the test statistic is measured up against a critical value from the chi-square distribution. If the test shows that the error terms (i.e. the residuals) are not normally distributed, this is only a problem for small samples with observations less than 100 (Gujarati & Porter, 2009). Furthermore, if the normality assumption test is rejected, this can indicate that the significance tests of the coefficients in the regression model may be misleading.

Independence of residuals

This assumption is related to the fact that the residuals are statistically independent of each other. That is, there is no correlation between the error terms. One can use the Durbin-Watson test to check for this kind of auto-correlation. The test operates with an upper and a lower bound. It is rejected if the test statistic is below the given bound and accepted if it is above. The test can be classified as non-conclusive if the test statistic is between the two bounds. The result of dependent residuals can be biased estimations of the standard errors of the coefficients. Moreover, this can lead to inaccurate results of the Student test statistic and rejection of the null hypothesis when it should not be rejected, and vice versa (Newbold et al., 2013).

Constant Variance

The fourth assumption is related to the fact that the sample has to be homoscedastic. That is, the residuals have a constant variance. One can test for homoscedasticity by the use of the Breusch and Pagan's test. In this test, the null hypothesis says that the error variances are all equal. However, a rejection of the null hypothesis indicates that the residuals are heteroscedastic and are subject to non-constant variance. This will lead to a more doubtful result of the calculated p-value (Newbold et al., 2013). The problem of heteroscedasticity can be controlled for by the use of heteroskedasticity-robust standard errors (Torres-Reyna, 2007).

Multicollinearity

The fifth assumption which only applies to multiple linear regression is the last to be presented in this section. Multicollinearity is said to be present in a multiple linear regression model when a close to perfect linear relationship between some of the independent variables is observed. This can also be for all independent variables. That is, two or more independent variables are highly correlated. When multicollinearity is observed, a normal consequence is large standard errors. Wide confidence intervals are observed, which give results that are less reliable (Newbold et al., 2013). To test for multicollinearity, the Variance Inflation Factor (VIF) test will be applied. The VIF test gives indications on how much larger the standard errors are, compared to a situation where the variables have zero correlation to other independent variables in the model. Increasing value in the VIF means less reliable regression results. The limit of the VIF will be set to 10, which is in line with Bowerman et al. (2005).

4.9.1.1 THE BEST LINEAR UNBIASED ESTIMATOR

The five assumptions given above are essential. The Gauss-Markow theorem suggests that a linear regression model where the residuals are uncorrelated, have a conditional mean of zero and are homoscedastic, gives the Best Linear Unbiased Estimator (BLUE) of the coefficients. The BLUE then states that the residuals do not need to be independent and identically distributed, nor do they need to follow a normal distribution. In summary, this means that as long as these key assumptions are fulfilled, the theorem holds, and the coefficients will be BLUE. This is strengthened by the Central Limit Theorem, which states that the sampling distribution of the mean of any independent, random variable, will be approximately distributed if the sample size is large enough (Gujarati & Porter, 2009).

5 DATA PRESENTATION

This section will present the data used in the analysis. Several aspects will affect the selection of data. Luckily, transparency is one of the key features of Bitcoin. All transactions from the major cryptocurrency exchanges are found open and available, in the form of open APIs providing real-time data. One dataset from each cryptocurrency exchange is downloaded, which include raw tick data. This include timestamp in Unix time, price in BTC and volume of the trade. Unix time is a time system that counts seconds from 00:00:00 Thursday, 1 January 1970, GMT. Every day contains precisely 86,400 seconds (Matthew & Stones, 2008).

5.1 DATA SELECTION

As already mentioned, there is a lot of data available from different cryptocurrency exchanges. Firstly, a decision was made on which period to include for the analysis. As this thesis is based on high frequency data, it would make sense to choose a period with high volume. This will provide a sufficient amount of data, to enable in-depth analysis. The year 2018 is selected, including all trades on all of the chosen cryptocurrency exchanges. Although the volume peaked around the price top at the end of 2017, 2018 has shown a higher volume overall. It could be useful to include some months from 2017, but the amount of data would be overwhelming. As the largest cryptocurrency exchanges have between 25-90 million trades in 2018, this should provide sufficient information for the analysis.

This thesis includes data from seven different cryptocurrency exchanges. These are chosen carefully and are mostly based on a detailed report from the investment company Bitwise Asset Management, which was published in March 2019. The report was a part of the company's ETF filing to the U.S. Securities and Exchange Commission. This report highlighted several factors of the bitcoin market and revealed that 95% of the reported trading volume was fake. The company used the reported volume from the most popular cryptocurrency website, CoinMarketCap. The report points out that only 10 out of 81 cryptocurrency exchanges have significant real volume. Bitwise performed two data-driven tests of the volume patterns. The first test looks at trade size histograms, revealing that a lot of exchanges have completely artificial trade size histograms. The second test looks at the volume spike alignment. This tells that the exchanges with real volume have volume spikes that align perfectly since they're part of the same market. On the other hand, the exchanges with fake volume have volume spikes that do not correspond with the broader market (Bitwise Asset Management, 2019).

Based on the report from Bitwise, the following exchanges were chosen:

- **Binance**

The leading exchange worldwide by reported volume. Seen significant growth the last year. Started as a Chinese exchange, but is now based in Malta due to regulatory aspects. The exchange is well known for offering a large number of altcoins.

- **Bitfinex**

Founded already in 2012 and have been one of the leading exchanges for several years, although several hacks have occurred. The firm behind the exchange is based in Hong Kong. The exchange has not been serving U.S customers since 2017, claiming it was too expensive. Critics have raised questions about the relationship between the exchange and the stablecoin Tether, and the solvency of Tether which is supposed to be backed by the US Dollar. Tether is closely related to Bitfinex and share both common shareholders and management (Castor, 2018).

- **Kraken**

As one of the oldest cryptocurrency exchanges, Kraken was founded in 2011. It is located in San Francisco. Kraken was the first exchange to display its market data on the Bloomberg Terminal and is backed by several large investors ("Why Kraken?", n.d.).

- **Bitstamp**

This exchange is based in London and was founded to offer a Europa-based alternative to the previously dominant cryptocurrency exchange Mt. Gox. As one of the oldest exchanges, it moved to the UK in 2013, after operating in Slovenia since 2011. It is well known for offering free deposits with fiat currencies through bank transfers in Europa. Accepted as a fully regulated payment institution in the EU in 2016, it can operate in EU countries (Williams-Grut, 2018).

- **Coinbase**

Founded in 2012, this exchange has been popular for several years. It is based in San Francisco and is backed by investors like ICE, the owner of the New York Stock Exchange. The exchange has been on the forefront with partnerships, including Paypal, Overstock and Dell. The exchange is available in 42 countries ("Coinbase Inc", 2019).

- **Poloniex**

This US-based exchange launched in 2014 and got acquired by the company Circle in 2018 for \$400 million. Circle is one of the leading blockchain and cryptocurrency companies today and launched a stablecoin backed by the US Dollar in 2018 in a project together with Coinbase (Alexandre, 2018).

- **Hitbtc**

Hitbtc was not a part of the list of exchanges with real volume in the Bitwise report and is included to see if this provide any suspicious or different results. However, Hitbtc is one of the oldest exchanges, with high reported volume and advanced API services for trading.

5.2 DATA CLEANING

This thesis will include two different approaches to lead-lag relationships. In the first part of the analysis that explores causality, the data needs to be defined in specific time intervals. Hence, new datasets are produced where 1-minute time intervals are made. This is done by taking the last price in each minute and summing all the volume of the transactions for the given minute. This results in datasets for each exchange, that include details on opening, closing and average price for each minute, in addition to the volume in that period. The sampling interval should be considered carefully. It is important to have short enough intervals to capture the high frequency behavior of the data correctly. At the same time, the intervals should be long enough to contain enough observations and to avoid noise (Goodhart & O'Hara, 1997). Anderson (2000) argues that an interval of 5 minutes should cover these challenges. However, given the high frequency data obtained from the different cryptocurrency exchanges, the first part of the analysis will use 1-minute intervals. The second part of the analysis will not need any sort of data cleaning, as the HY-approach uses raw transaction data.

5.3 VALIDITY AND RELIABILITY

Validity and reliability are significant aspects to consider when assessing the quality of the research data. Validity is related to the degree that the conducted research accomplishes what it is intended to do, according to Smith (2011). Furthermore, Bryman and Bell (2011) say that the goal of a study should be to end up with findings that approximately corresponds to the real world. In this thesis, the theories and methods are all based on well-established academic papers. This approach and the data collected are based on previous empirical studies, which support the validity of this thesis. Some subjective selections of data have been made in regard of the cryptocurrency exchanges and time periods; thus the validity can be reduced.

In terms of reliability, one usually asks the question of whether the same conclusions would be reached by another researcher, following the same procedures and conducting the same study, under the same circumstances (Smith, 2011). This thesis only collects secondary data, hence rely on existing data. The cryptocurrency exchanges used in this these have the historical trade data public, and all are assumed to be reliable sources. According to Bryman and Bell (2011), the study is then considered repeatable.

6 DATA ANALYSIS

6.1 DESCRIPTIVE STATISTICS

Section 5 presents two different datasets for the analysis. They present the same data, but with different frequency. The datasets that are based on 1-minute intervals include 525,600 observations. These datasets will be used in this section to describe the characteristics of the data for the analysis. Table 6.1 below shows a summary of the return series of the bitcoin price on the different cryptocurrency exchanges.

Table 6.1 - Descriptive statistics of returns in 2018

	Mean	Median	Max	Min	Std. Dev	Skewness	Kurtosis
Binance	-1.48E-06	0	0.0650	-0.0458	0.1424%	0.639	43.869
Coinbase	-1.73E-06	0	0.0307	-0.0516	0.1255%	0.376	38.521
Bitstamp	-1.52E-06	0	0.0369	-0.0317	0.1410%	0.199	18.509
Bitfinex	-1.56E-06	0	0.0572	-0.0340	0.1324%	0.686	34.101
Kraken	-1.60E-06	0	0.0474	-0.0730	0.1400%	0.298	49.882
Hitbtc	-1.80E-06	0	0.0252	-0.0199	0.1180%	0.498	21.747
Poloniex	-1.53E-06	0	0.0321	-0.0244	0.1400%	0.268	18.455

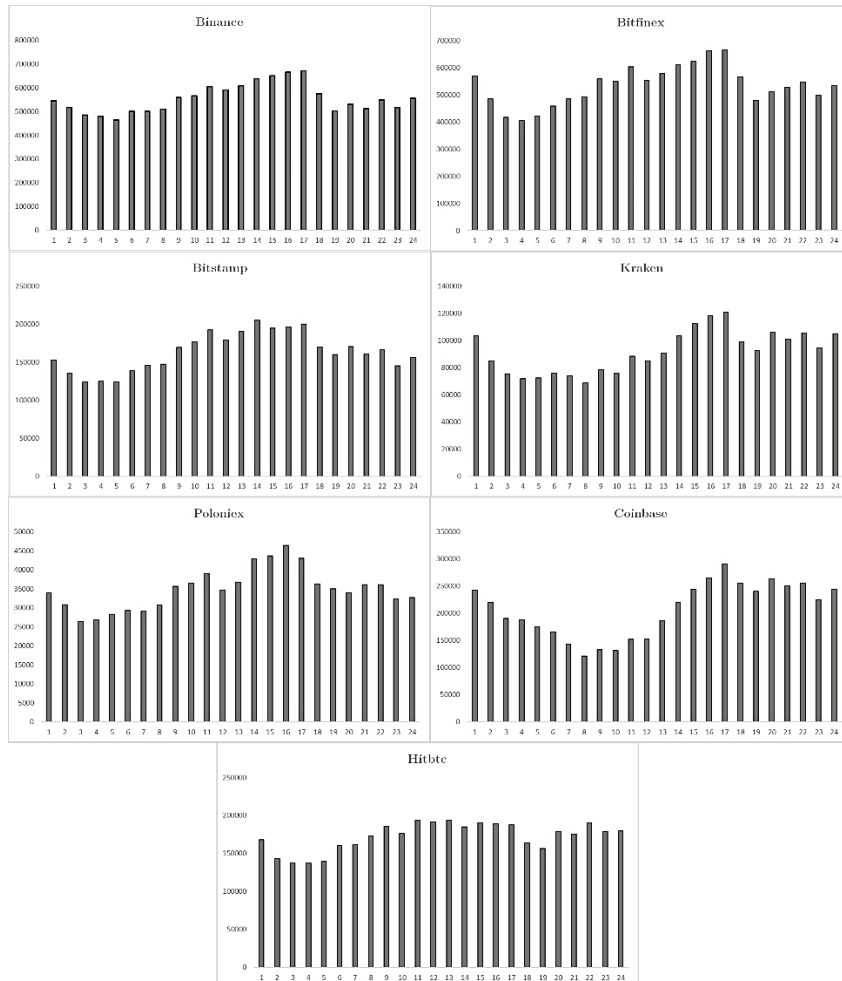
As seen from the table above, there are some interesting characteristics that should be explained. Some outliers can be seen from the Max and Min values. Binance shows a strong minute return of 6.5% as the max value. Kraken on the other hand, is the exchange with the largest negative return, with -7.3% in one minute. These extreme returns are often followed by more extreme returns, as the price usually revert back. This is because these extreme returns often are triggered by someone placing an order quite far above or below the last price, and the price will then revert back. The standard deviations do not show any apparent trends, but Binance as the largest exchange has the highest standard deviation. This indicates larger price movements on Binance, but the differences are too small to draw any conclusions.

The skewness does also show different results among the exchanges. Both Binance and Bitfinex have a skewness above 0.6, whereas Bitstamp has the lowest with 0.19. This indicates that Binance and Bitfinex are the exchanges with the longest tails to the right, and the most extreme positive returns occur on these exchanges. However, it should be noted that skewness below 1 cannot be called substantial and the distribution is not far from symmetrical, but Binance and Bitfinex have distributions that are moderately skewed (Bulmer, 1979). Bitstamp has the lowest skewness, which

indicates that Bitstamp has the shortest tail and less extreme returns than the other exchanges. The kurtosis is more divided. A higher kurtosis indicates a heavier tail. This is especially present for Kraken and Binance, which indicate that the amount of extreme returns is more substantial on these exchanges. Additional statistics on the price series and volume series of all exchanges are included in Table 1 and Table 2 in the Appendix.

As bitcoin can be traded throughout the day with no limitations in trading hours, a presentation of the trading activity can be interesting. The charts in Figure 6.1 show the amount of bitcoin traded in every hour and show some clear patterns. Coinbase has the most evident pattern, where we see a dip in trading volume during night time in the U.S. This indicates a large number of American investors on Coinbase, where the exchange also is located. Binance and Bitfinex dip most during Asian night time and suggest that a larger group of the investors could be based in Asia. Bitstamp, Kraken, in addition to Bitfinex, dip during European night hours, suggestion more European investors.

Figure 6.1 - Trade distribution over hours and weekdays on the different exchanges



An interesting observation is that Coinbase has almost 60% difference from top to bottom trading volume during a day. On the other hand, Hitbtc and Binance only have a difference of around 30%. This clearly indicates a different trading pattern on these global exchanges. Moving on to the weekday distribution seen in Figure 1 in the Appendix, most of the exchanges show the same patterns. All exchanges except Hitbtc clearly have lower volume during weekends. This could raise some questions about the legitimacy of the reported volume of Hitbtc, which is also confirmed by the report on real trading volume by Bitwise (Bitwise Asset Management, 2019).

Table 6.2 below presents the characteristics of the trade size on the different exchanges. For these statistics, the full datasets with all transactions for each exchange have been used. It is clear that small trades with a volume below 0.01 BTC, are dominant across all exchanges. The fact that Poloniex and Coinbase have almost 65% and 45% of all trades in the area below 0.01 BTC, in addition to only 1.83% and 3.66% of all trades above 1 BTC, provide valuable insight. This point towards exchanges with mostly retail investors and few professional investors. In contrast, Bitfinex seems to have a large group of professional investors, with almost 8% of trades with volume over 1 BTC, and only 17.5% of trades below 0.01 BTC. The full trade size distribution can be seen in Figure 2 in the Appendix.

Table 6.2 - Characteristics of the trade size of the individual exchanges

	< 0.01 BTC	> 1 BTC
Binance	33.48 %	2.26 %
Bitfinex	17.47 %	7.75 %
Bistamp	30.61 %	7.23 %
Kraken	24.65 %	6.71 %
Poloniex	64.74 %	1.83 %
Coinbase	44.89 %	3.66 %
Hitbtc	34.74 %	4.54 %

Lastly, it is interesting to have a look at correlations. Table 3 and Table 4 in the Appendix provide results of correlations between the price and return series. The price series are almost perfectly correlated. As they represent the same asset, different results would be troubling. Kraken and Hitbtc have the lowest correlations. This could already point towards some differences that could give rise to potential lead-lag relationships across the exchanges. This is further strengthened by looking at the correlation matrix for the return series. This also shows that especially Kraken has low correlations with other exchanges. Another interesting observation is related to Poloniex. This exchange showed the highest price correlations with other exchanges, but the lowest return correlations alongside Kraken. This can indeed point toward lagging price movements, which will be analyzed in the next section.

6.2 LEAD-LAG RELATIONSHIPS

This section of the analysis will include two different approaches to lead-lag relationships, hence be divided into two parts. To begin with, it is important to check for stationarity. This is one key assumption when working with time series analysis. Subsequently, cointegration between the series will be tested, through the Johansen cointegration test. The first part of the analysis will then end with tests for Granger causality between the series, which would provide valuable insight for further analysis. The second part of this section will make use of high frequency data, and apply the Hayashi-Yoshida cross-correlation estimator. These results will provide information on how strong the lead-lag relationships are between exchanges, in addition to details about the time lag of these relationships. Furthermore, regression analysis will be used to explore how volume affects the lead-lag relationships found with the Hayashi-Yoshida estimator.

6.2.1 CAUSALITY APPROACH

6.2.1.1 AUGMENTED DICKEY-FULLER TEST

As mentioned above, the time series need to be controlled for stationarity. As this thesis analyzes one asset on different exchanges, the time series from only one exchange will be examined in this section, assuming same results. The visual inspection below makes use of the values from Binance. Figure 6.2 shows the price of one bitcoin throughout 2018, including a plot of the price in logarithmic levels.

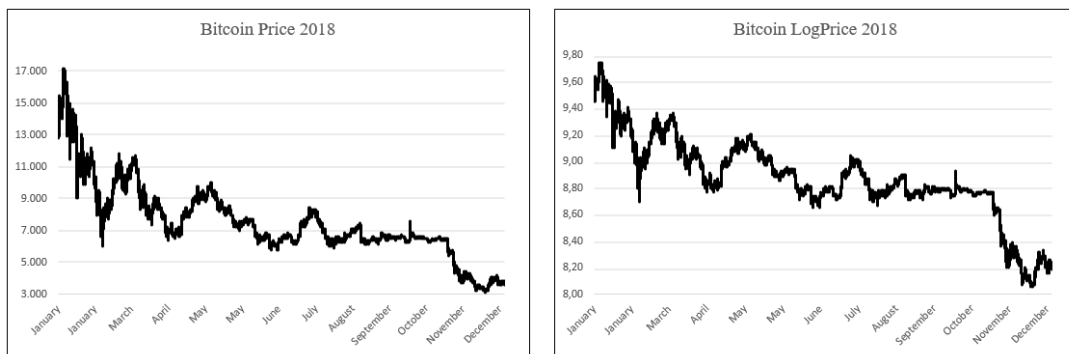


Figure 6.2 - Plot of the bitcoin price during 2018, including log-levels

A visual inspection of both plots indicates non-stationarity, with a clear tendency of a trend downwards throughout the year. This is not surprising, as stock prices and other assets typically are decent examples of non-stationary $I(1)$ series (Bollerslev et al., 1992). The figure above indicates that the

series need to be integrated to achieve stationarity. Nevertheless, the Augmented Dickey-Fuller test is applied to confirm the indications from the visual inspections. This test includes lags to account for autocorrelation in the residuals, as explained in Section 4.7.1. Table 6.3 below confirms the indications, and the null hypothesis of a unit root in the series cannot be rejected, indicating that the bitcoin price contains a unit root and is non-stationary.

Table 6.3 - Results from the Augmented Dickey-Fuller test

Augmented Dickey-Fuller (Binance)		
	BTC	LogBTC
ADF test statistic	-2.4099	-1.5978
P-value	0.1389	0.5287
H0	Not Rejected	Not Rejected

Note: Critical values, 1%: -3.430, 5%: -2.862, 10%: -2.567. Both a constant and trend are included based on the plots. 5 lags are included according to the results from the SBIC.

Stationary can possibly be achieved by differencing the log series of bitcoin. Once again, a visual inspection is needed. Figure 6.3 below indicates that taking the first differences of the log-levels to transform the series into stationarity.

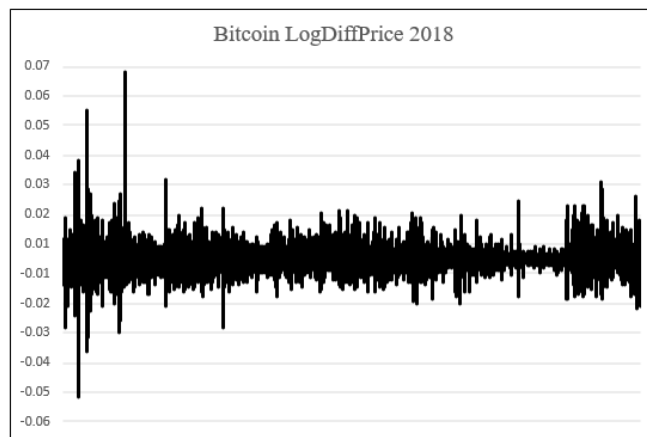


Figure 6.3 - Plot of the first-differenced log-prices of bitcoin in 2018.

The Augmented Dickey-Fuller test is applied to confirm the indications. Table 6.4 below shows that all series are stationary, as the null hypothesis is rejected at a 1% significance level. Hence, the bitcoin series are stationary in their first differences, i.e. $I(1)$ non-stationary processes. This is in accordance with expectations. The visual inspections from all bitcoin prices on the different exchanges are included under Figure 3 in the Appendix, which includes plots of log-levels, in addition to ACF plots to check for autocorrelation. All results in this section have been tested with a range of different lags, as the

Augmented Dickey-Fuller test is sensitive to the chosen lag length. However, the sensitivity analysis showed that the tests still showed stationarity for the bitcoin prices.

Table 6.4 - Augmented Dickey-Fuller test of all bitcoin prices

	Binance	Bitfinex	Bitstamp	Kraken	Poloniex	Coinbase	Hitbtc
ADF test statistic	-532.915	-521.257	-539.842	-422.775	-370.931	-415.01	-702.94
P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Lags	1	1	1	2	3	2	1
H0	Rejected	Rejected	Rejected	Rejected	Rejected	Rejected	Rejected

Note: Critical values, 1%: -3.959, 5%: -3.410, 10%: -3.127. Constant and trend are not included based on the plot. Lag length is chosen by the use of SBIC.

6.2.1.2 JOHANSEN COINTEGRATION TEST

The last section showed significant results of non-stationary time series. This was an important finding, as this indicates that cointegration is theoretically possible, and leads the way for this section. As presented in Section 4.7.3, Johansen Cointegration test will be applied. This multivariate cointegration test is based on a VECM, where all seven time series from the different exchanges are included. That is, $g = 7$. As all time series are based on the same asset and should in theory be exactly the same, cointegration between the series would be expected. Nevertheless, a cointegration test is applied to confirm these expectations. Furthermore, all series are converted into logarithmic levels to smooth the data and reduce the impact of outliers (Keene, 1995).

As presented in Section 4.7.3, the Johansen approach will be affected by the selected lag length and the inclusion of deterministic terms. A VAR model needs to be formulated, to determine the characteristics before applying the cointegration test. The use of an information criterion will be applied to determine the optimal lag length. By including all seven exchanges, the VAR model will be defined with 7 variables. This model is then transformed into a VECM, where a set of the 7 variables are in first differenced form on the left-hand side, with the $k - 1$ lags of the dependent variables differenced on the right-hand side, with the coefficient matrix attached. The Johansen test will determine the number of cointegrated vectors in the system of variables, presented as the rank r . Table 6.5 below presents the results from the two test statistics of the Johansen cointegration test.

Table 6.5 - Johansen cointegration test

Trace Test Statistics					Max Eigenvalues Test Statistics				
	Variable statistic	Crit-90%	Crit-95%	Crit-99%		Variable statistic	Crit-90%	Crit-95%	Crit-99%
$r = 0$	16717.107	120.367	125.619	135.983	$r = 0$	9870.651	43.295	46.230	52.307
$r = 1$	6846.456	91.109	95.754	104.964	$r = 1$	5354.186	37.279	40.076	45.866
$r = 2$	1492.270	65.820	69.819	77.820	$r = 2$	507.429	31.238	33.878	39.369
$r = 3$	984.842	44.493	47.855	54.682	$r = 3$	429.214	25.124	27.586	32.717
$r = 4$	555.628	27.067	29.796	35.463	$r = 4$	318.357	18.893	21.131	25.865
$r = 5$	237.271	13.429	15.494	19.935	$r = 5$	235.218	12.297	14.264	18.520
$r = 6$	2.053	2.706	3.842	6.635	$r = 6$	2.053	2.706	3.842	6.635

Note: The VECM used in the Johansen test includes 2 lags given from the information criterion (SBIC). This is also suggested when dealing with a system of variables (Juselius, 2006). Both a constant and trend are included based on the log-levels time series plots of the different bitcoin prices.

The results above are in accordance with expectations. The trace test's null hypothesis states that the number of cointegrated vectors is less than or equal to r . The results above show that this null hypothesis is rejected on the 1% significance level for all values of r , except $r = 6$. At this rank, the test statistic is 2.053, which is below all critical values. Hence, the null hypothesis is not rejected. This indicates a rank of 6, and the result show a reduced rank, with 6 cointegrated equations in the VECM. The other test statistic, the maximum eigenvalues test, test each eigenvalue individually. The null hypothesis states that the number of cointegrated vectors is precisely r . As for the trace test, the null hypothesis is rejected on a 1% significance level for all values of r , except $r = 6$. The test statistics at rank 6 is exactly the same as for the trace test, and the null hypothesis of 6 cointegrated vectors cannot be rejected. Hence, both tests show a rank of 6.

The results presented above is satisfying and as expected. As earlier described, these results show a reduced rank, i.e. $r < g$. A sensitivity analysis on the lag length is performed as well, without any change of the results. Hence, the log-level of the bitcoin prices are cointegrated. Different results would be troubling, as the series are based on the same asset, and should follow the same path over time.

6.2.1.3 GRANGER CAUSALITY

The previous section shows cointegration of the bitcoin price on the different cryptocurrency exchanges in the long run. It is also interesting to look at the short-run relations that could exist. To end the first part of the analysis, the lead-lag relationships between the bitcoin prices in the short-term will be analyzed in this section. This will be done by testing for Granger causality between the bitcoin prices.

Table 6.6 - Granger Causality test of all 21 return pairs

Exchanges			Test stat	P-value	Conclusion
Coinbase	->	Bitstamp	1619.39	0	Bidirectional
Bitstamp	->	Coinbase	2220.85	0	
Coinbase	->	Kraken	1943.46	0	Bidirectional
Kraken	->	Coinbase	244.37	0	
Bitstamp	->	Kraken	2618.12	0	Bidirectional
Kraken	->	Bitstamp	114.82	0	
Coinbase	->	Bitfinex	98.15	0	Bidirectional
Bitfinex	->	Coinbase	4641.24	0	
Bitstamp	->	Bitfinex	1433.68	0	Bidirectional
Bitfinex	->	Bitstamp	6037.44	0	
Kraken	->	Bitfinex	154.57	0	Bidirectional
Bitfinex	->	Kraken	6553.70	0	
Hitbtc	->	Bitfinex	46.68	0	Bidirectional
Bitfinex	->	Hitbtc	6681.28	0	
Hitbtc	->	Kraken	204.70	0	Bidirectional
Kraken	->	Hitbtc	580.50	0	
Coinbase	->	Hitbtc	185.72	0	Bidirectional
Hitbtc	->	Coinbase	93.73	0	
Binance	->	Coinbase	5235.88	0	Bidirectional
Coinbase	->	Binance	108.14	0	
Binance	->	Bitstamp	5235.88	0	Bidirectional
Bitstamp	->	Binance	5304.34	0	
Binance	->	Bitfinex	6479.27	0	Bidirectional
Bitfinex	->	Binance	4055.86	0	
Binance	->	Kraken	4898.49	0	Bidirectional
Kraken	->	Binance	317.59	0	
Poloniex	->	Kraken	117.13	0	Bidirectional
Kraken	->	Poloniex	431.43	0	
Poloniex	->	Hitbtc	83.08	0	Bidirectional
Hitbtc	->	Poloniex	373.18	0	
Poloniex	->	Bitstamp	32.98	0	Bidirectional***
Bitstamp	->	Poloniex	1212.04	0	
Poloniex	->	Bitfinex	41.78	0	Bidirectional
Bitfinex	->	Poloniex	9694.21	0	
Poloniex	->	Coinbase	82.17	0	Bidirectional
Coinbase	->	Poloniex	914.06	0	
Poloniex	->	Binance	96.01	0	Bidirectional
Binance	->	Poloniex	6561.84	0	
Binance	->	Hitbtc	8268.02	0	Bidirectional
Hitbtc	->	Poloniex	41.30	0	
Bitstamp	->	Hitbtc	1175.10	0	Bidirectional
Hitbtc	->	Bitstamp	122.62	0	

* H_0 says that X does not Granger Causes Y , reject if the test statistic is above the critical level.

** Test at 1% significance level with critical value: 9.210.

*** 2 lags used for all tests. Results are not affected by a change in lags, except for Poloniex --> Bitstamp. This pair cannot reject the H_0 with 1 lag included, and the only unidirectional relationship is found, indicating that Bitstamp Granger causes Poloniex, but not vice versa.

The previous section showed that the bitcoin prices do indeed share a stochastic trend in the long-run. This section will now analyze if one bitcoin price can be used to predict another bitcoin price in the short-term, i.e. if one bitcoin price Granger causes another bitcoin price. According to Granger (1988), a cointegration relationship between two series indicate that at least one of the series Granger causes the other. Further analyses need to be done to identify if the lead-lag relationships are unidirectional or bidirectional.

The Granger causality tests will be based on a VAR model, where restrictions are applied. To identify specific relationships between the individual exchanges, pairwise analysis between the bitcoin prices will be applied. Furthermore, to apply the Granger causality tests the series needs to be stationary. This section will make use of the return series of all prices, which are all stationary. Details on the stationarity of the return series can be found in Table 5 in the Appendix. Table 6.6 above show Granger causality Wald's test for all 21 pairs. All exchanges show a bidirectional relationship. The null hypothesis is rejected for all tests on a 1% significance level. This basically means that all return series can be used to improve the predictions of another return series. Once again, this is not surprising, as all series represent the same asset. When lead-lag relationships are found, it is not uncommon to find bidirectional relationships (Kawaller et al., 1987). This could also be related to the fact that the series are based on 1-minute intervals, and do not provide enough details of the price movements.

Moreover, the results from the Granger causality test also could be related to something completely different. Bidirectional Granger causality can be read either as instant causality or common cause fallacy (Maziarz, 2015). This kind of misunderstanding casual relations between series can impact the conclusion. There could be another series that cause both series that are tested. This is the most common reason for spurious causality (Chu et al., 2005). It is not unlucky that common cause fallacy is presented for the tested series in this thesis. It could perhaps be that the most liquid exchange, Binance, is the true leader in price movements. Hence, Binance could cause other prices series, even when it's not a part of the test. As the pairs where Binance is included also show the bidirectional relationship, there could be even more exchanges that are true leaders of price movements. This discussion is important to be aware of and will be tested further in Section 6.2.3, where factors that affect the lead-lag relationships are explored through the regression analysis.

The first part of the analysis can conclude that there clearly exist lead-lag relationships between the

bitcoin prices on the different cryptocurrency exchanges. In addition, this part has highlighted the issues with this classic causality approach to determine lead-lag relationships. As the goal is to find results that possibly can lead to profitable trading strategies and identify true leading exchanges in price movements, these results do not provide a satisfactory level of details. The next part of the analysis will explore another approach, where high frequency trade data will be used.

6.2.2 HAYASHI-YOSHIDA CROSS-CORRELATION ANALYSIS

As an introduction to the second part of the analysis, the datasets will be presented. The HY-estimator works with high frequency trade data, and an overview of the trades on the different exchanges is useful. Table 6.7 shows all trades and total volume on the different cryptocurrency exchanges throughout 2018.

Table 6.7 – Overview over the number of trades and the volume on each cryptocurrency exchange for 2018.

Exchange	Trades	Volume (BTC)
Binance	87,343,449	13,305,165
Bitfinex	30,183,548	12,827,302
Coinbase	25,125,016	4,961,375
Bitstamp	10,949,631	3,932,073
Hitbtc	9,265,218	4,140,081
Poloniex	9,115,848	837,495
Kraken	7,029,062	2,367,024

This part of the analysis undoubtedly includes datasets with different characteristics than the first part, which only included 525.600 observations from each exchange. Binance has the most trades with over 87 million during 2018. There is a relatively large gap down to Bitfinex and Coinbase, which had just over 30 and 25 million trades, respectively. The rest of the exchanges had between 7 and 10 million trades during 2018. As explained in Section 4.8, the high frequency approach provides an unbiased estimation with non-synchronous data. The differences in trading activity should not affect the results. One intuitive thought regarding the lead-lag relationship is that the asset with the highest trading activity would lead the other. This is also confirmed by several studies on futures contracts and stock prices. However, the HY cross-correlation function has been tested through several simulation studies and has shown not to be affected by different levels of trading activity. This characteristic of the estimator is essential for this analysis as the trading activity varies between the cryptocurrency exchanges. Allowing the results to avoid being fooled by liquidity effects would yield correct results, and not automatically yield the most traded asset to be the leader (Huth & Abergel, 2012).

6.2.2.1 THE CROSS-CORRELATION FUNCTION

All 21 pairs of exchanges have been tested, with an initial lead-lag grid of -60 seconds and +60 seconds. This will let the results present a leading or lagging relationship between two exchanges up to 1 minute. Remembering from Section 4.8, this will generate a cross-correlation function. This function makes it possible to decide if one of the series leads the other, by measuring the asymmetry of the cross-correlation between the positive and negative lags. If the function tops out on lag 0, the correlation is highest when no adjustments are made in lags, indicating a weak lead-lag relationship. However, a function that indicates a stronger lead-lag relationship will have a top that is found either at a positive or a negative lag. All functions are presented in Figure 4 in Appendix, and a selection of interesting results are presented in Figure 6.4 below.

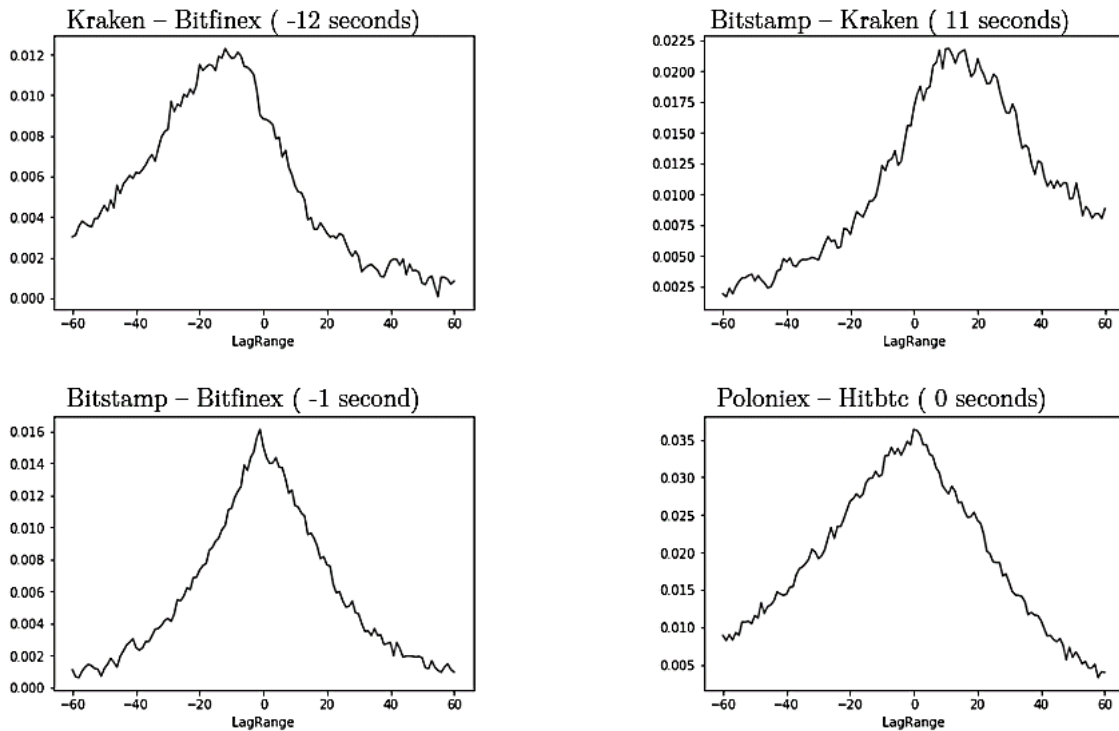


Figure 6.4 - Hayashi-Yoshida cross-correlation functions.

X-axis: 120 lags (seconds), indicating -60 seconds to +60 seconds adjustment of the Y variable in the estimator.

Y-axis: Correlation between series at a given lag adjustment.

These functions provide examples of different lead-lag relationships. The first line of functions indicates strong lead-lag relationships between the exchanges. The function for Kraken and Bitfinex is skewed to the left. This result indicates that the correlation between the bitcoin prices on these two exchanges is strongest when Bitfinex is leading Kraken with 12 seconds. The function of Bitstamp and Kraken shows similar results. This function is skewed to the right, and the correlation between the exchanges is strongest when Bitstamp is leading Kraken with 11 seconds. It is essential to understand that this

is the same as saying that Kraken is lagging behind Bitstamp with 11 seconds. The two functions on the first line are only skewed to different directions because different series are chosen as X variable and Y variable in the Hayashi-Yoshida estimator. Both functions indicate that Kraken is a lagging exchange, which will be explored further later in the analysis. The two bottom functions show a different picture. Bitstamp and Bitfinex have an almost symmetrical function, indicating a weak lead-lag relationship. The correlation on both sides dies out in the same way, and the maximum correlation is found where Bitstamp is lagging 1 second behind Bitfinex. Poloniex and Hitbtc show similar results. The pair of these two exchanges is the only one with a maximum correlation at lag 0.

The same four functions are presented in Figure 6.5 below, that zooms in on lags smaller than 10 seconds. These functions give an even better picture of the observed lead-lag relationships. The two functions on the first line are now far from symmetrical around lag 0. The two bottom functions are still symmetrical, but Bitstamp and Bitfinex could be said to have a more symmetrical function when the zoom is applied on these two pairs. Hence, Bitstamp and Bitfinex could have a less present lead-lag relationship than Poloniex and Hitbtc, even though the max correlation is at -1 second and not at lag 0.

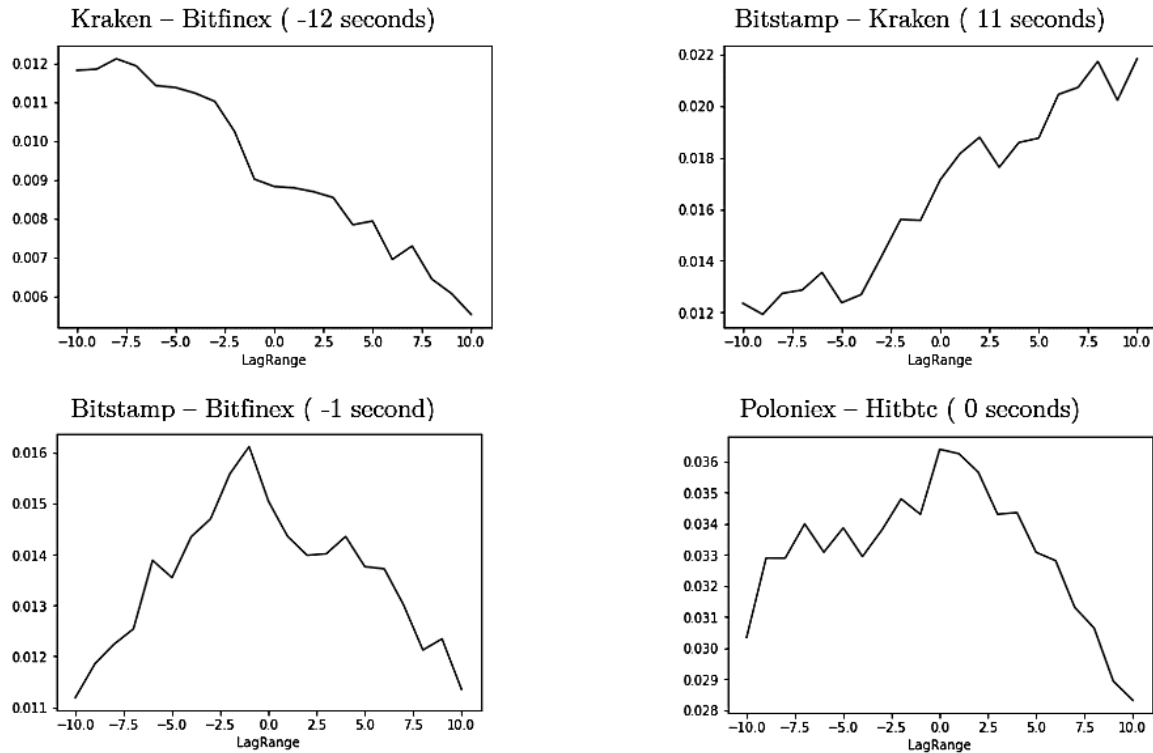


Figure 6.5 - Figure 6.4 zoomed in on only 10 seconds

6.2.2.2 LEAD-LAG RELATIONSHIPS

To investigate these relationships further, Table 6.8 presents all 21 pairs with the maximum correlation, at which lag (given in seconds) this correlation occurs, and the strength of the lead-lag relationship (Lead-Lag Ratio). That is, tests where the Y series is adjusted 60 seconds lagging and 60 seconds leading for each timestamp of the X series. The same results for 10 lags can be seen in Table 6 in the Appendix. Due to the high number of pairs, all results will not be described. However, some of the most interesting results will be elaborated on.

Starting with the pairs that are presented through their functions in Figure 6.4, noticeable differences can be found. Kraken and Bitfinex have a lead-lag ratio of 0.3684. As described in Section 4.8, this ratio represents the asymmetry of the cross-correlation function. A ratio of 1 would indicate that weights on each side of lag 0 in the cross-correlation function are precisely the same, i.e. the sum of the correlations are the same on each side. The lead-lag ratio of 0.3684 indicates a strong leading relationship for Bitfinex, and that Bitfinex correlates more with lags of Kraken, then the other way around.

Table 6.8 - Results from the Hayashi-Yoshida cross-correlation estimator.

This show the maximum correlation, the time lag where it occurs, and the lead-lag ratio.

X	Y	Seconds	Correlation	LLR
Binance	Bitfinex	-1	0.01741	0.8109
Binance	Bitstamp	-3	0.01325	0.8574
Binance	Coinbase	-1	0.01804	1.1971
Binance	Hitbtc	2	0.03707	1.6134
Binance	Kraken	15	0.01829	2.2693
Bitstamp	Bitfinex	-1	0.01612	1.0158
Bitstamp	Hitbtc	4	0.03877	1.7373
Bitstamp	Kraken	11	0.02186	2.3919
Coinbase	Bitfinex	-2	0.03292	0.6219
Coinbase	Bitstamp	-1	0.03620	0.5351
Coinbase	Hitbtc	3	0.08644	1.2911
Coinbase	Kraken	8	0.04584	1.6734
Hitbtc	Bitfinex	-3	0.02954	0.6121
Hitbtc	Kraken	8	0.03738	1.6509
Kraken	Bitfinex	-12	0.01231	0.3684
Poloniex	Binance	-7	0.00170	0.6536
Poloniex	Bitfinex	-7	0.01157	0.5163
Poloniex	Bitstamp	-7	0.01284	0.5284
Poloniex	Coinbase	-5	0.01449	0.6936
Poloniex	Hitbtc	0	0.03640	0.8333
Poloniex	Kraken	4	0.01983	1.1971

Even more remarkable is the results of Bitstamp and Kraken. As the function in Figure 6.4 presents, Bitstamp is leading Kraken, with a fat tail on the right, indicating that the correlation between Bitstamp and lags of Kraken die out slowly. There is actually a lot of correlation left even when Bitstamp is leading Kraken with 60 seconds. This is also reflected in the result from Table 6.8, with a lead-lag ratio of 2.3919. However, the maximum correlation between the exchanges is not among the highest of the 21 pairs and could indicate that the lead-lag relationship will be difficult to make a profit on when a trading strategy is applied.

Moving on to Bitstamp and Bitfinex, the results are different. The function looks almost entirely symmetrical, indicating a weak lead-lag relationship. The lead-lag ratio is indicating the same, with a value of 1.0158. This value indicates that Bitstamp is leading Bitfinex, but the maximum correlation is found when Bitfinex is leading Bitstamp with one second. The results for only 10 lags support that Bitfinex is the leading exchange, with a lead-lag ratio of 0.9782. Hence, the result is sensitive to the number of lags included. Due to the symmetrical function, small changes in lags included will affect the lead-lag ratio, as the weights on both sides are almost identical. Moreover, the correlation is relatively weak for this pair. Overall, the results for these two exchanges are satisfying. The trading activity is significantly higher for Bitfinex, having over 30 million trades in 2018 compared to Bitstamp's 10 million trades. Hence, the credibility of the HY estimator is strengthened as it doesn't automatically yield the exchange with the highest trading activity to be the leading one. The last pair of Poloniex and Hitbtc show similar results. The function with 60 lags looks symmetrical; however, the correlation dies out slower when Hitbtc is leading Poloniex. The function where only 10 lags are included shows this even more explicit, where the correlation stays higher when Hitbtc is leading Poloniex. The lead-lag ratio is below 1 for both tests, indicating that Hitbtc is leading Poloniex. However, these two exchanges are the only pair that has a maximum correlation at lag 0. This means that the exchanges correlate the most when no adjustments are made and when there theoretically is no lead-lag relationship.

Moreover, Coinbase and Hitbtc have a maximum correlation that is way above the other pairs, with 0.08644. This occurs when Coinbase is leading Hitbtc with 3 seconds. The function for this pair is almost symmetrical and shows that the correlation is nearly the same for the lags from -1 to 4. This points towards a pair that is relatively highly correlated, where it stays high throughout several seconds.

Even though the lead-lag ratio 1.2911 for 60 lags and only 1.1061 for 10 lags, this could be an interesting pair to test through a trading strategy.

6.2.2.3 DISCUSSION OF RESULTS

The 21 pairs presented in this analysis show different characteristics. This section discusses some of the highlights and trends that are found between all the pairs. This discussion is a central part of understanding how the bitcoin market works and how the bitcoin prices on the different cryptocurrency exchange behave.

6.2.2.3.1 THE EFFICIENCY OF THE LARGEST EXCHANGES

Binance, Bitfinex and Coinbase are the exchanges with the highest trading activity during 2018. Although Binance had around 3 times the number of trades compared to the two other exchanges, these three stand out from the rest of the exchanges. Binance and Bitfinex are clearly the exchanges with the highest volume as well, with 13.3 million and 12.8 million bitcoin traded in 2018, respectively. They are followed by Coinbase with almost 5 million bitcoin traded. As seen from Table 6.8, the maximum correlation occurs at 1 second for all these pair, except for one pair that has the maximum correlation at 2 seconds. Furthermore, these pairs show lead-lag ratios that are closer to 1 than most of the other pairs. Overall, these results are satisfying and also as expected. It would be strange if the largest bitcoin exchanges in the world showed significant differences in price movement, and hence provided substantial arbitrage opportunities. When the HY estimator provides results that point toward the fact that the largest cryptocurrency exchanges are more efficient than the smaller ones, with less pronounced lead-lag relationships, this strengthens the credibility of the approach.

6.2.2.3.2 KRAKEN AS THE LAGGING EXCHANGE

The behavior of one exchange needs to be discussed. Kraken is the exchange with the lowest trading activity. However, the overall volume is around 2.4 million bitcoin in 2018, which is higher than Poloniex with only 0.8 million bitcoin traded. Almost all pairs that include Kraken show that the exchange is lagging the other exchange with around 10 seconds. Poloniex and Kraken are a bit different, with a lead-lag ratio closer to 1, and only 4 seconds lead for Poloniex. 4 seconds is still more than the efficient exchanges mentioned above, and Kraken can be said to behave differently than other

exchanges. The cross-correlation functions that include Kraken are presented below in Figure 6.6. These show that when the correlation is higher when Kraken is lagging behind, i.e. the correlation dies slower.

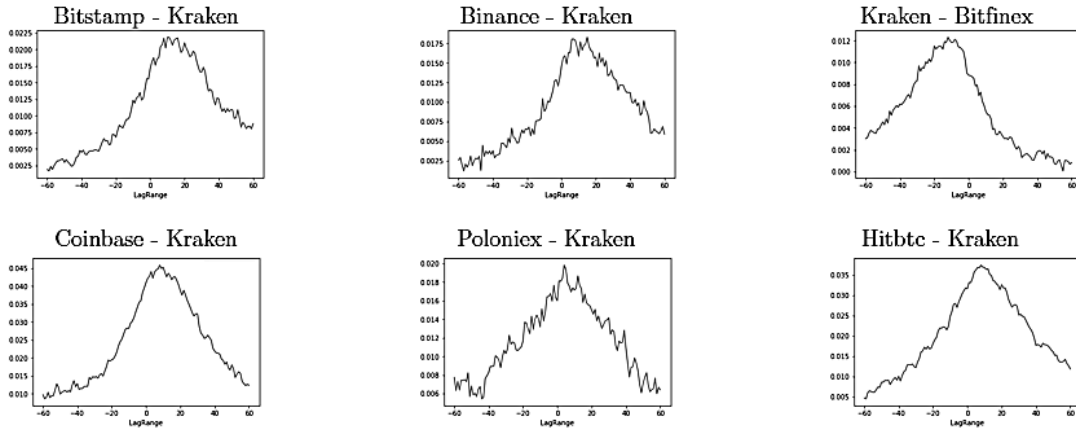


Figure 6.6 - All cross-correlation functions where Kraken is included.

This leads to the discussion of why this is the case. An initial look at the characteristics of the different exchanges points toward the low trading activity. However, the Hayashi-Yoshida cross-correlation is supposed to correct for differences in trading activity, as confirmed in Huth & Abergel (2012). To get a deeper understanding of the trading activity effect, one can look at the other exchanges with low activity. A natural starting point is Poloniex, with the lowest trading volume and just 2 million trades more than Kraken. Poloniex lags precisely 7 seconds behind three of the biggest exchanges that all vary relatively much in trading activity. Furthermore, Poloniex has a weaker lead-lag relationship to Coinbase compared to Bitstamp. As Coinbase has nearly 14 million trades more than Bitstamp, these relationships don't seem to be affected by the trading activity. An interesting observation is that Poloniex and Hitbtc show a weak lead-lag relationship, with the maximum correlation at lag 0. Moreover, they have approximately the same trading activity, but still totally different trading volume, which is noteworthy.

A look at other exchanges should also contribute to a better understanding of the effect of trading activity. Hitbtc leads Kraken with 8 seconds, which is the same relationship that Coinbase and Kraken show. However, Coinbase has 25 million trades in 2018, compared to Hitbtc's 9 million trades. Furthermore, Hitbtc only lags 2 seconds behind Binance, which has 87 million trades in 2018 and is by far the largest exchange. Hitbtc then shows a weaker relationship to Binance, than to the other large exchanges. Moreover, Bitstamp shows a leading relationship to Coinbase, that have 15 million trades more than Bitstamp in 2018 and a lower overall volume. Hence, no clear trends can be confirmed by

just looking at the trading activity. Indeed, smaller exchanges have a lagging relationship to the larger exchanges. However, based on the discussion above it is too easy to point at the trading activity as the reason for this. One can discuss if the trading volume and activity can explain some of the differences that are observed, but no clear trends are seen. Especially when Hitbtc and Poloniex are compared with the largest exchanges, trading activity cannot be blamed as the only reason for the presented relationships. An interesting discussion could be if the HY-estimator face problems when differences in trading activity become too large, or if a lower threshold should be set for trading activity. As a trade on Kraken is observed every 4-5 seconds on average in 2018 and Binance show almost 3 trades per second, this could certainly affect the results. Nevertheless, Kraken behaves differently than the other exchanges and seems to be a decent candidate for a trading strategy based on lead-lag relationships.

6.2.2.3 GRANGER CAUSALITY

Remembering the results in Section 6.2.1.2, all 21 pairs showed a bidirectional relationship. This is in accordance with the results from the Hayashi-Yoshida estimator. The Granger causality tests were done with 1-minute intervals, whereas the cross-correlations functions presented in this section only include 1 minute in total. A comparison of the Granger causality tests with only 1 lag included and the cross-correlation functions can be made, as both then represent 1-minute lead-lag relationships. When only 1 lag was applied to the Granger causality tests, Poloniex and Bitstamp could not reject the null hypothesis that Poloniex does not Granger cause Bitstamp. When studying the cross-correlation function for this pair with 99% confidence bands, one cannot reject that the correlation is zero when Poloniex is leading Bitstamp with around 55-60 seconds, as seen in Figure 6.7.

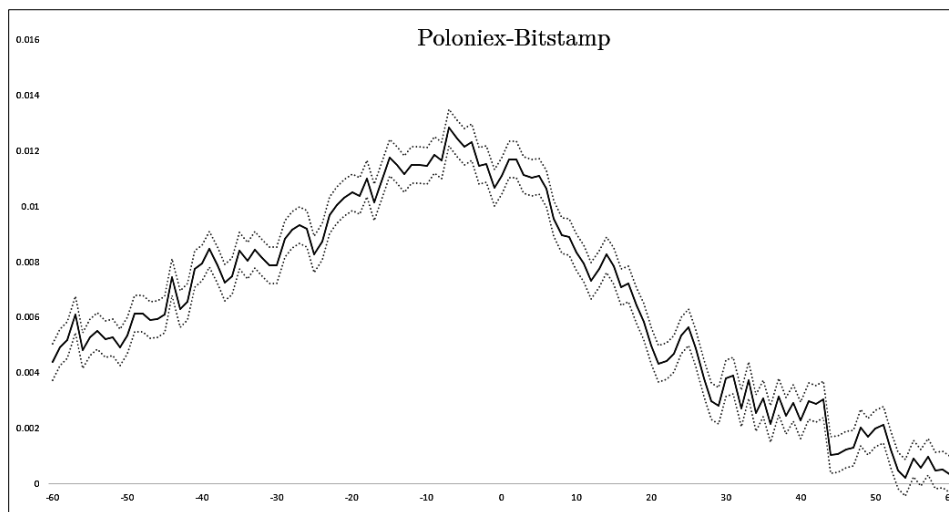


Figure 6.7 - The cross-correlation function for Poloniex and Bitstamp, with 99% confidence bands included.

For most of the pairs that show a relatively low test statistic in the Wald test for Granger causality, the same scenario as presented above occur. The correlation drops down to levels around zero when the lag move closer to 60. One reason for the fact the Poloniex and Bitstamp are the only pair that fail to reject the null hypothesis with 1 lag in the Wald test, could be related to the overall low correlation for most of the lags when Poloniex is leading Bitfinex.

To summarize, the bidirectional relationships is in accordance with the calculated cross-correlation functions. The HY-estimator gives valuable insight into the details of the lead-lag relationships, even when the relationships are bidirectional. This is given by the lead-lag ratio and will be necessary for setting up trading strategies later in the analysis. Furthermore, a unidirectional relationship would be expected if the correlation dropped sharply on one side of lag 0 only, indicating low correlation when one exchange is leading the other, but higher correlated when the other exchange leads. For setting up trading strategies, functions with clear drops on one side of lag 0 in the cross-correlation function would be the best. However, this is not the case for any of the pairs in this analysis. The pair of Bitstamp and Kraken is perhaps the best example, where the correlation stays quite high for several lags on one side of lag 0 and drops faster on the other side. Once again, this leads to the conclusion that Kraken should be tested as a lagging exchange in a trading strategy.

6.2.2.3.4 CORRELATION VALUES

The last discussion of the Hayashi-Yoshida results will be related to the correlation values. The observant reader has probably already found the correlation values to be extremely low. The correlation matrices presented in the descriptive statistics showed highly correlated price series close to 1, and correlations around 0.5-0.7 for the return series. These were based on 1-minute candles, which is quite different from the HY-estimator that uses seconds as lags. This is unfortunate for the interpretation of the cross-correlations, but there is a logical reason for these low values as mentioned in Section 3.5.3.1. The high frequency data used in the HY-estimator include trades every second, or even include several trades per second. Naturally, these movements can be extremely small and often totally insignificant for the price changes. When every non-zero price variation is included, this leads to a lot of noise in the data. These insignificant movements in price could just be noise around a given price level, and not indicate a real movement of the price. Furthermore, when analyzing two datasets of high frequency trades that are measured in returns from trade to trade, it would be surprising if these return series followed the exact same movement. Hence, correlation values become extremely small.

To avoid and minimize this microstructure noise, thresholds can be applied. That is, only movements that are above a given threshold will be included in the analysis. By doing this, small and insignificant movements will be avoided, and the high frequency datasets that are compared will most likely yield higher levels of correlation. Huth & Abergel (2012) present a thresholded version of the Hayashi-Yoshida cross-correlation estimator, to check if substantial returns are more informative than small ones. Their test does indeed yield satisfying results. Overall, the trend is that the lead-lag relationship becomes more and more pronounced as the focus moves to larger price variations. Both the lead-lag ratio and the maximum correlation increases with the given threshold. When this threshold changes from 0.5 to 3, which is determined by the tick size, the maximum correlation seems to double in value. However, the most important result from the test is related to the maximum lag. The results show that the lag where the maximum correlation occurs is more or less independent of the threshold (Huth & Abergel, 2012).

These findings are essential for this thesis. As the end goal is to establish trading strategies, the most critical indicator is the maximum lag which indicates the number of seconds an exchange leads another exchange. Hence, the results from the HY estimator and the cross-correlation functions can be used to build trading strategies, as confirmed by Huth & Abergel (2012). However, when these strategies are being built, one should filter out insignificant moves to avoid too much noise and trading fees.

6.2.3 REGRESSION ANALYSIS

As the previous section presented several lead-lag relationships between the different cryptocurrency exchanges, this part of the analysis will try to locate some of the relevant factors that impact these relationships. By looking at the effect of information arrival on lead-lag relationships, a deeper understanding of why some exchanges lead others will hopefully be achieved. Section 4.9 introduced linear regression, which will be used in this part of the analysis. This section will follow the study of Dao et al. (2018), which analyzes the effect of information arrival on high frequency lead-lag relationships.

6.2.3.1 DEPENDENT VARIABLES

In the regression analysis, the dependent variables will be related to the lead-lag relationships. That is, variables that represent the maximum correlation coefficient, at which time this occur given in seconds,

and the lead-lag ratio that measures the strength of the leadership. It is important to include both the maximum correlation coefficient and the lead-lag ratio, as they can yield differences in which exchange that is leading. The peak of the cross-correlation function can be on one side, but the lead-lag ratio can indicate that the correlation is generally higher on the other side. Hence, the two results indicate different leading exchange (Dao et al., 2018). These three lead-lag quantities are calculated for each day in 2018. That is different from the previous part of the analysis, where these were calculated for the whole period. Now, 2018 will include 365 observations with lead-lag characteristics for each day of the year. Hence, the pairs of cryptocurrency exchanges that will be tested through the regression analysis will have three daily series of lead-lag variables that will be used as the dependent variables in the regression analysis.

6.2.3.2 INDEPENDENT VARIABLES

Section 3.6 highlights the fact that several aspects can affect the lead-lag relationship. Trading mechanisms will not be included as bitcoin is electronically traded. Regarding trading cost, this varies across the different cryptocurrency exchanges and is not observable due to individual fee structures that traders achieve based on trading volume. Hence, this will not be included in the regression analysis but will be elaborated on in Section 7.2. However, trading volume will be the key factor to explain information arrival to the market that may affect the lead-lag relationships.

All independent variables will be based on the total volume for each exchange per day. The two first divide the daily trading volume into block trades and non-block trades, which will be proxies for sophisticated and non-sophisticated investors. Remembering from Section 3.6, sophisticated investors are institutional investors that provide large blocks of volume. Hence, the non-block volume will be the difference between the total daily volume and the volume that is defined as block volume. Block volume will be defined as trades of 1 bitcoin or above that size. Letting V , BV and NBV denote the total, block and non-block volume, respectively, the following equation is obtained (Dao et al., 2018):

$$NBV_t = V_t - BV_t$$

To get even more details from the trading volume, these two volume characteristics are divided into the expected and the unexpected component to capture the normal level of market activity and the arrival of new information, as found by Arago and Nieto (2005). For a given day, this expected component is equal to the volume of the previous day. Hence, the unexpected component is the new

volume that occurs, on top of the volume of the previous day. If EBV , UBV , $ENBV$ and $UNBV$ denote the expected block, unexpected block, expected non-block and unexpected non-block, the following equations can be made (Dao et al., 2018):

$$\begin{aligned} EBV_t &= BV_{t-1} \\ UBV_t &= BV_t - EBV_t \\ ENBV_t &= NBV_{t-1} \\ UNBV_t &= NBV_t - ENBV_t \end{aligned}$$

To summarize, the trading volume is now divided into four different components that will be used as independent variables in the regression analysis. These represent two dimensions of the trading volume; the block and non-block, in addition to the expected and the unexpected.

6.2.3.3 THE OVERALL MODEL

Three hypotheses are presented for the regression analysis to explore the effect of trading volume on the lead-lag relationships:

1. The information flow (i.e. unexpected volume) affects the lead-lag relationship (i.e. the dependent variables)
2. Sophisticated investors (i.e. block volume) have a more significant impact on the lead-lag relationship than non-sophisticated investors (i.e. non-block volume).
3. Market leading exchange by volume (i.e. Binance volume) affects the lead-lag relationship of other exchanges.

The regression model includes four independent variables that indicate different characteristics of the daily volume. The exchange pairs that will be used in the regression analysis are selected based on how strong their lead-lag relationships are. When Binance is not included in the regression analysis, the model will also include the total volume of Binance. This is to test if the volume of the largest exchange is significant for lead-lag relationships, even when this exchange is not a part of the analysis. This could provide valuable information on the fact that a dominant exchange in the bitcoin market actually impacts the behaviour of smaller exchanges. Furthermore, this provides valuable information on the discussion of common cause fallacy that was discussed in Section 6.2.1.3. A separate regression will be

estimated for each lead-lag variable, i.e. lead-lag correlation coefficient, lead-lag time and the lead-lag ratio. If Y denotes this variable, the regression model can be written as:

$$Y_t = \alpha + \beta_1 EBV_{t,X} + \beta_2 UBV_{t,X} + \beta_3 ENBV_{t,X} + \beta_4 UNBV_{t,X} + \gamma_1 EBV_{t,Y} + \gamma_2 UBV_{t,Y} + \gamma_3 ENBV_{t,Y} + \gamma_4 UNBV_{t,Y} + \delta_1 V_{t,BINANCE}$$

The regression analysis will include the four pairs that are presented in the beginning of Section 6.2.2.1, which result in 12 different regression models. These pairs have different characteristics in both volume and the lead-lag results from the HY-estimator.

6.2.3.4 ASSUMPTIONS

To provide reliable results in the regression analysis, a look at all assumptions presented in Section 4.9.1 is important. The goal is to achieve the Best Linear Unbiased Estimator (BLUE) of the coefficients. All assumptions are tested for the 12 models in the regression analysis. The results are troubling. Nine out of twelve models violate all assumptions except linearity. The other three models violate at least two of the assumptions. Table 7 in the Appendix show all tests of assumptions for these twelve models. This indicates that the models need to be adjusted. A natural starting point is related to the dependent variables. A closer look at these shows troubling results, as they are not stationary and seem to include a trend. The plots for the dependent variables are included in Figure 5 in the Appendix. One example is included in Figure 6.8 below, showing the values for the maximum correlation for the pair of Poloniex and Hitbtc. The logarithmic values of the lead-lag time are spurious for some plots and not relevant, as these are undefined for negative values. As seen, the dependent variables become stationary when they are transformed to first differenced values. Present for all plots is the fact the variance increases over time.

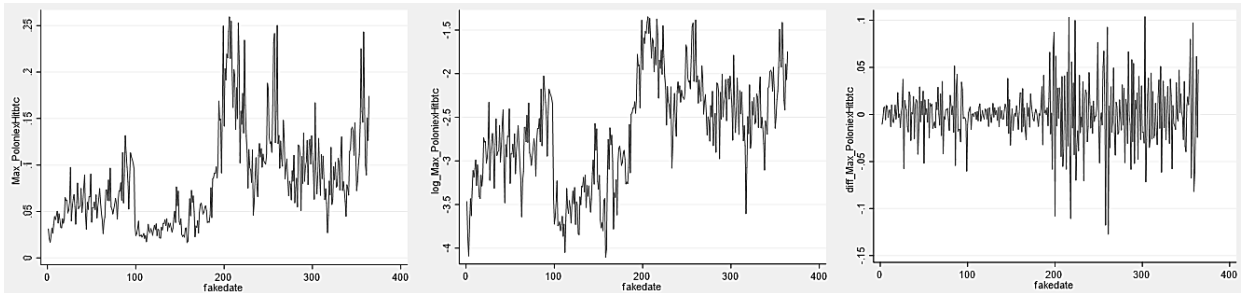


Figure 6.8 - Plots of lead-lag results from Poloniex-Hitbtc. From left: Original values, log values, first-differenced values.

Based on the discussion above, all dependent variables are transformed into first-differenced values. Hopefully, this leads to no violation of the regression assumptions and avoidance of spurious regression.

An evaluation of all assumptions will now be presented for the pair of Poloniex and Hitbtc, with lead-lag max correlation as the dependent variable. The rest of the assumption tests are found in Table 8 in the Appendix, as all regression models are based on the same type of variables that represent exchange volume. However, critical differences in assumption violations between the twelve regression models will be addressed.

To check for linearity, all independent variables are plotted against the dependent variable. Figure 6.9 below shows all scatterplots. The first column to the left shows the dependent variable plotted against independent variables and the rest of the matrix show independent variables against each other. If any plots look nonlinear, additional augmented partial residual plots are investigated, where the residuals are plotted against the independent variable. An example of this is shown in Figure 6.10, with the unexpected non-block volume of Hitbtc. The smoothed line is close to the ordinary regression line, and there is no evident nonlinearity. This is the case for all independent variables, and the assumption of linearity is not violated for the regression models.

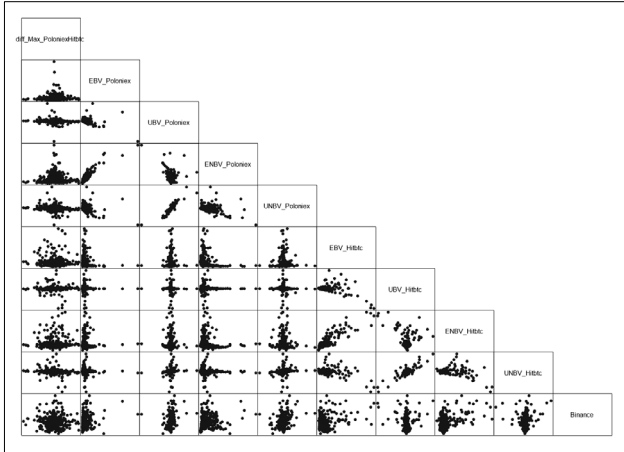


Figure 6.9 - Graph matrix of the pair Poloniex-Hitbtc

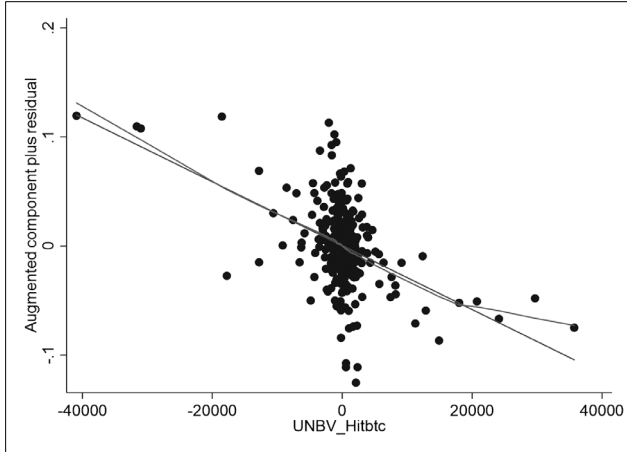


Figure 6.10 - Augmented partial residuals plot of UNBV_Hitbtc

Normality is checked with the use of the Jarque-Bera test. The null hypothesis for the test is that the data is normally distributed, and the alternate hypothesis is that the data does not come from a normal distribution. As can be seen from Table 8 in the Appendix, the test statistics are above the critical values for all regression models, and the null hypothesis is rejected at a 1% significance level. This leads to a violation of the assumptions of normality. However, normality is not required in order to obtain unbiased estimates of the regression coefficients, as described in Section 4.9.1.1. The third assumption of independent residuals is, however, a requirement for OLS regression. This is tested with the use of the Durbin-Watson test, where the critical value of the lower band is 1.78182 and the upper band

1.87261 for the regression models with 363 observations at a 5% significance level. The test statistic for the regression model of Poloniex and Hitbtc with max correlation as dependent variable shows a value of 2.753. This is above the critical band value, and the assumption of independent residuals are not violated. As seen in Table 8 in the Appendix, this is the case for all regression models.

The fourth assumption is related to the constant variance of the residuals. This is tested with the Breusch and Pagan's test, in addition to plots with the residuals against fitted (predicted) values. As seen in Table 8 in the Appendix, the null hypothesis that the variance of the residuals is homogeneous is not rejected at a 5% significance level for all regression models, except the pair of Bitstamp and Bitfinex with max correlation as the dependent variable. For this regression model, heteroscedasticity-robust standard errors are applied, as explained in Section 4.9. The last assumption of multicollinearity is tested with the VIF-test. As the independent variables are the same for each of the three regression models for each pair, all test results are presented below in Table 6.9. A VIF value over 10 is worrying and could indicate that the variable is considered as a linear combination of other independent variables. The pair of Poloniex and Hitbtc is not violating the assumption of no multicollinearity. However, the three other pairs show different results. This is, however, a natural result, as the independent variables in this regression analysis are based on the same total volume, and strongly correlated exchanges and their volume. Luckily, even extreme multicollinearity (so long as it is not perfect) does not violate OLS assumptions. OLS estimates are still unbiased and BLUE (Best Linear Unbiased Estimators). Greater multicollinearity leads to greater standard errors. This will affect the confidence intervals of the coefficients, which will be wide with small t-statistics (Williams, 2015). Hence, it will be more challenging to achieve significant coefficients, and this will be a critical remark when evaluating the regression results in the next section.

Table 6.9 - Test of multicollinearity between all independent variables of the pairs (from left in the table):
Poloniex-Hitbtc, Bitstamp-Bitfinex, Bitstamp-Kraken and Kraken-Bitfinex

Variable	VIF	1/VIF	Variable	VIF	1/VIF	Variable	VIF	1/VIF	Variable	VIF	1/VIF
ENBV_Hitbtc	8.83	0.113262	UNBV_Bitfi-x	24.38	0.041009	ENBV_Kraken	14.73	0.067876	UNBV_Bitfi-x	22.84	0.043788
EBV_Hitbtc	7.26	0.137820	ENBV_Bitfi-x	23.76	0.042084	UNBV_Kraken	14.34	0.069755	UBV_Bitfinex	21.61	0.046279
UNBV_Polon-x	6.98	0.143171	UBV_Bitfinex	20.13	0.049666	EBV_Kraken	11.22	0.089137	ENBV_Kraken	18.92	0.052865
ENBV_Polon-x	6.35	0.157379	ENBV_Bitst-p	15.95	0.062702	UBV_Kraken	8.79	0.113815	EBV_Bitfinex	17.45	0.057299
UBV_Poloniex	6.14	0.162775	EBV_Bitfinex	15.67	0.063817	UNBV_Bitst-p	8.40	0.119092	UNBV_Kraken	17.19	0.058184
EBV_Poloniex	6.00	0.166739	UNBV_Bitst-p	12.74	0.078520	ENBV_Bitst-p	7.32	0.136648	ENBV_Bitfi-x	17.17	0.058241
UBV_Hitbtc	5.75	0.173782	UBV_Bitstamp	5.62	0.177922	UBV_Bitstamp	5.78	0.172991	EBV_Kraken	11.26	0.088833
UNBV_Hitbtc	5.69	0.175656	EBV_Bitstamp	4.90	0.203903	EBV_Bitstamp	5.77	0.173308	UBV_Kraken	9.59	0.104270
Binance	2.34	0.427681	Binance	3.22	0.310539	Binance	3.24	0.308768	Binance	3.35	0.298479
Mean VIF	6.15		Mean VIF	14.04		Mean VIF	8.84		Mean VIF	15.49	

Overall, the assumptions for Best Linear Unbiased Estimators are fulfilled, and the transformation of the dependent variables to first differenced values yielded satisfying results.

6.2.3.5 REGRESSION RESULTS

Table 6.10 represents the regression models for the four pairs, where max correlation and lead-lag ratio are the dependent variables. That is, testing the effect of information arrival on the daily lead-lag relationships among the four pairs. The results of the regression models for the lead-lag time are included in Table 9 in the Appendix since none of the coefficients in the models showed to be significant.

Table 6.10 - Regression results – Effect of information arrival on lead-lag relationships

	Poloniex - Hitbtc		Bitstamp - Bitfinex		Bitstamp - Kraken		Kraken - Bitfinex	
Panel A: lead-lag correlation coefficient								
Intercept	-0.005848	0.1637	-0.0026953	0.2584	-0.006087	0.2032	-0.002354	0.5374
Excepted Poloniex block volume	-0.000017	0.6151						
Unexpected Poloniex block volume	0.000015	0.7486						
Excepted Poloniex non-block volume	0.000001	0.6760						
Unexpected Poloniex non-block volume	-0.000002	0.4109						
Excepted Hitbtc block volume	-0.000003	0.5280						
Unexpected Hitbtc block volume	0.000014***	0.0033						
Excepted Hitbtc non-block volume	0.000000	0.7239						
Unexpected Hitbtc non-block volume	-0.000003***	0.0001						
Excepted Bitstamp block volume			-0.000002	0.6006	0.000002	0.6307		
Unexpected Bitstamp block volume			0.000002	0.7381	-0.000000	0.9344		
Excepted Bitstamp non-block volume			0.000000	0.3058	0.000000	0.4668		
Unexpected Bitstamp non-block volume			0.000002**	0.0120	0.000001	0.1500		
Excepted Bitfinex block volume			0.000001	0.5667			0.000001	0.7839
Unexpected Bitfinex block volume			0.000001	0.8027			0.000001	0.8507
Excepted Bitfinex non-block volume			-0.000000	0.3653			0.000000	0.8605
Unexpected Bitfinex non-block volume			-0.000001***	0.0004			-0.000001	0.1476
Excepted Kraken block volume					0.000000	0.9926	0.000001	0.9346
Unexpected Kraken block volume					0.000022*	0.0947	0.000003	0.7682
Excepted Kraken non-block volume					-0.000001	0.6797	-0.000001	0.6135
Unexpected Kraken non-block volume					-0.000008***	0.0001	-0.000001	0.5122
Total Binance volume	0.000000	0.2656	0.000000	0.4893	0.000000	0.6073	0.000000	0.6507
R-squared	0.0877		0.2144		0.1668		0.1462	
P-value	0.0002***		0.0000***		0.0000***		0.0000***	

	Poloniex - Hitbtc		Bitstamp - Bitfinex		Bitstamp - Kraken		Kraken - Bitfinex	
Panel C: lead-lag ratio								
Intercept	0.007212	0.9020	0.023336	0.7121	-0.159859	0.3000	0.050009	0.3465
Excepted Poloniex block volume	0.000059	0.9008						
Unexpected Poloniex block volume	0.000423	0.5235						
Excepted Poloniex non-block volume	-0.000004	0.8605						
Unexpected Poloniex non-block volume	-0.000010	0.7842						
Excepted Hitbtc block volume	-0.000072	0.2309						
Unexpected Hitbtc block volume	0.000161**	0.0138						
Excepted Hitbtc non-block volume	0.000008	0.3004						
Unexpected Hitbtc non-block volume	-0.000021**	0.0315						
Excepted Bitstamp block volume			0.000045	0.3846	-0.000004	0.9792		
Unexpected Bitstamp block volume			0.000014	0.8438	0.000088	0.6065		
Excepted Bitstamp non-block volume			-0.000014	0.2890	0.000015	0.4893		
Unexpected Bitstamp non-block volume			0.000003	0.8111	0.000035	0.2142		
Excepted Bitfinex block volume			-0.000033	0.4356			0.000001	0.9767
Unexpected Bitfinex block volume			-0.000048	0.3670			0.000065*	0.0753
Excepted Bitfinex non-block volume			0.000006	0.2453			-0.000002	0.4455
Unexpected Bitfinex non-block volume			0.000001	0.8530			-0.000008**	0.0321
Excepted Kraken block volume					0.000051	0.8875	-0.000033	0.7185
Unexpected Kraken block volume					0.000148	0.7316	-0.000044	0.7705
Excepted Kraken non-block volume					-0.000027	0.5831	0.000013	0.2947
Unexpected Kraken non-block volume					-0.000114*	0.0828	0.000004	0.8439
Total Binance volume	-0.000001	0.6148	-0.000001	0.8084	0.000004	0.4765	-0.000002	0.2531
R-squared	0.0426		0.0199		0.0261		0.0315	
P-value	0.0774*		0.6209		0.3390		0.2488	

Note: Significance on *10%, **5% and ***1% level

As expected, the regression models do not provide a high explanatory level. Naturally, the volume cannot explain the lead-lag relationships alone, and the R-squared for the regression models are relatively low. However, satisfying results are found related to the effect of information arrival. According to Panel A and C in Table 6.10, there is evidence that the lead-lag relationships among the cryptocurrency exchanges are influenced by the rate of information arrival. This effect is captured by the unexpected trading volume on these exchanges and confirms the hypothesis presented earlier. Noteworthy is the fact that the only significant variables in every regression model are the ones related to the arrival of information, i.e. unexpected volume. A complete overview of all detailed regression results can be found in Table 10 in the Appendix.

6.2.3.5.1 DISCUSSION

Panel A presents the models for the lead-lag correlation coefficient, where all regression models are significant on a 1% level. Few coefficients are significant, but some trends can be seen. As already explained, only information arrival seems to affect the lead-lag correlation. Furthermore, it seems that when this information arrives from sophisticated investors, this will lead to an increase in the lead-lag correlation. On the other hand, when unexpected trading activity from non-sophisticated investor occurs, the lead-lag correlation seems to decrease. Moreover, it is difficult to conclude anything on the effect of information arrival from the leading or the lagging exchanges from Panel A. Poloniex and Hitbtc indicate that the leading exchange is the one affecting the lead-lag correlation. However, Bitstamp and Kraken show the opposite, where the lagging exchange seems to affect the lead-lag correlation more.

It could be tempting to divide the four pairs based on their characteristics. Remembering from Section 6.2.2.1, both pairs that include Kraken show a strong lead-lag relationship, where Kraken is lagging. If the focus in Panel A shifts to only these two pairs, these indicate that the lagging exchange is affecting the lead-lag correlation. This is clear for Bitstamp and Kraken, where unexpected volume from Kraken is significant for the lead-lag correlation. For Kraken and Bitfinex, the coefficient closest to the significance level is the unexpected volume from non-sophisticated investors at Kraken. Although this coefficient is not statistically significant, it has a p-value that is much lower than the other coefficients in the model, and according to Wasserstein and Lazar (2016), this relative comparison is still valid. Even if the coefficients do not match in the direction of the sign, this may overall indicate that

information arrival at lagging exchanges affects the lead-lag correlation the most. However, one can only conclude that this is purely indications.

Moving on to Panel C, the explanatory level of the regression models decrease a lot, with an R-squared ranging between approximately 0.02 and 0.04. None of the regression models is significant on the 5% level. However, a discussion based on the significant coefficients is still relevant. Panel C shows the regression models where the lead-lag ratio is the dependent variable. This is perhaps the most interesting variable, as it states the strength of the lead-lag relationship. Starting with Poloniex and Hitbtc, the coefficients for informational arrival of the leading exchange, Hitbtc, are significant at the 5% level. Once again, the four pairs can be divided based on their characteristics. This indicates that the information arrival on the leading exchanges affects the lead-lag relationship when exchanges with a weak lead-lag relationship are evaluated. The group of the other two pairs where the lead-lag relationship is stronger, show similar results. The unexpected volume on the lagging exchange, Kraken, seems to affect the lead-lag ratio for the pair of Bitstamp and Kraken. However, this is only at the 10% significance level. The pair of Kraken and Bitfinex is in accordance with Poloniex and Hitbtc, indicating that the leader affects the lead-lag relationship. For the other coefficients, it is not easy to conclude anything by the signs of the significant coefficients. Overall, information flow from leading exchanges seems to impact the lead-lag ratio more than information flow from the lagging exchanges.

To summarize this regression analysis, one can look at the hypotheses presented at the beginning of the analysis. The hypothesis regarding the information flow was already confirmed early in the analysis. All significant variables were related to unexpected volume. The second hypothesis stated that sophisticated investors have a more significant impact on the lead-lag relationship. It seems that sophisticated investors impact the lead-lag relationship more than non-sophisticated investors. That is, in regression models where both investor types showed significant impact, the absolute value of the coefficients for block volume is higher. Results of both these hypotheses are satisfying. One would assume that new information in the form of unexpected volume would make a more substantial impact on these relationships, in addition to the volume from sophisticated investors. Since sophisticated investors trade with larger amounts, it is natural that these types of volume changes affect the relationships more. This is also in accordance with the findings of Dao et al. (2018,), and could indicate that the lead-lag relationships on traditional financial markets and cryptocurrency markets are affected by the same type of information arrival. It is noteworthy that several coefficients related to the non-

block volume are significant. These non-sophisticated investors provide noise to the market due to a large number of small trades and are often related to investors whose decisions to buy, sell, or hold are irrational and erratic (De Long et al., 1990). These results are satisfying as noise in the sense of a large number of trades can be at least as powerful as a small number of large trades (Black, 1986).

The last hypothesis states that the volume on the market-leading exchange will affect the lead-lag relationship of other exchanges. The total volume of Binance was included as an independent variable in all regression models. This volume did not seem to have a significant impact on the lead-lag relationships of any pairs, and the third hypothesis is rejected. Nevertheless, this variable can be used to form an interesting discussion. Insignificant results are also results, and both Wasserstein and Lazar (2016) and Lopez de Prado (2019) highlight the misuse of p-values and how these significance levels can misrepresent the ground truth. P-values furthermore requires strong assumptions, which are not always realistic for financial data (Lopez de Prado, 2019). When looking at Panel A, there is not much that could indicate how the volume of Binance affects the lead-lag correlation of the pairs. However, it can be seen that all Binance coefficients are positive. These coefficients all have a value of zero in the figure above, due to the limitation in decimals included. Details of regression results can be seen in Table 10 in the Appendix, which shows that all Binance coefficients have a slightly positive value. As the volume of Binance can be assumed to be a proxy of the overall bitcoin volume, these results indicate that the overall correlations of the different bitcoin prices increase with an overall volume increase. Intuitively this makes sense, as higher volume indicates more participants in the market, which lead to decreasing spreads and more correctly priced assets (Siegel et al., 2000).

As for Panel C, dividing the pairs into groups will again give new results. When looking at the two pairs with strong lead-lag relationships, it is clear that an increase in Binance volume leads to a stronger lead-lag relationship. Bitstamp and Kraken have a positive coefficient for Binance volume, which increases the lead-lag ratio. As Bitstamp is leading, an increasing ratio strengthens the relationship (since it should be above 1). The same effect is seen for the pair of Kraken and Bitfinex, where the signs are the opposite. Since Bitfinex is the leader (and the Y variable in the original test in the HY estimator), the ratio should be below 1, and a negative coefficient indicates a stronger lead-lag relationship. It is vital to mention that the discussion above is solely included to indicate possible characteristics of the market. It is not meant to conclude anything about the cryptocurrency markets based on statistical significance, which can be a dangerous practice (Wasserstein et al., 2019).

6.3 TRADING STRATEGY

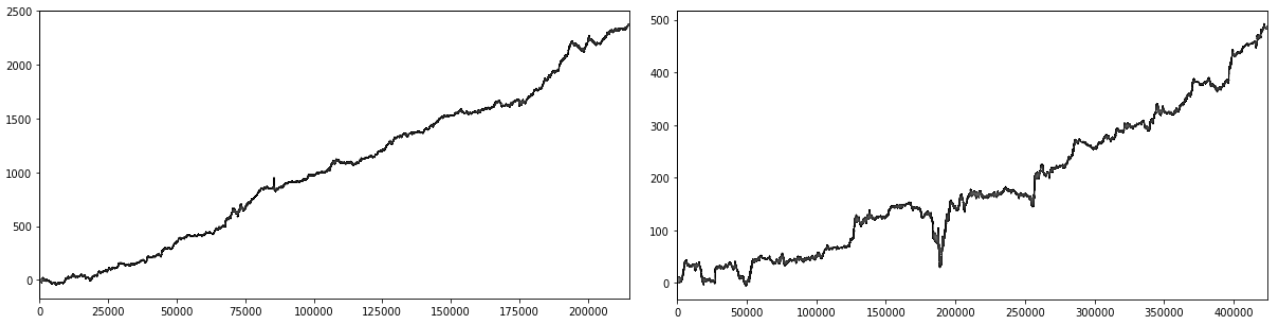
The results from this thesis can be used to build statistical arbitrage strategies that seek to take advantage of the temporary imbalance between the bitcoin prices of two cryptocurrency exchanges. The Hayashi-Yoshida cross-correlation estimator yielded several strong lead-lag relationships. Some of the pairs of exchanges will be used to backtest the possibility of arbitrage between these exchanges. Given the assumption that one exchange leads another, information from the leading exchange can be used to execute trades on the lagging exchange. That is, given price movements of the leading exchange that occur at time $t - 1$, these movement will occur at the lagging exchange at time t . The datasets used for backtesting will be for the first two months of 2019. The data includes all trades done in this period, including price, volume and the timestamp of the trades. Unfortunately, full order books are not accessible. Hence, the trading strategy will be based on mid-quote values. The implications of this assumption will be discussed later, as this clearly deviate from a real trading situation.

6.3.1 A SIMPLE FORECASTING STRATEGY

First, a simple strategy based on the movements of the leading exchange is presented. This “next tick” strategy initiates trades on the lagging exchange based on either an upwards movement or a downwards movement on the leading exchange. If the leading exchange moves upwards, one bitcoin on the lagging exchanges is bought and sold again when the next tick on the lagging exchange occurs. If the leading exchange moves downwards, the trade on the lagging will be the opposite, with a short position instead of a long position. As the trading activity varies on the two exchanges, the lagging exchange uses the last tick movement on the leading exchange that is closest in time to the tick on the lagging exchange. This simple strategy is just included to establish a forecasting device on the price movements of the two exchanges, to test the accuracy. This will tell if the markets actually behave as the lead-lag relationship results indicate. However, this is an extremely unrealistic strategy, as it is based on mid-quotes and do not include trading fees. A more realistic strategy will be presented in the next section.

Figure 6.11 below shows a test of the first 20 days of February 2019, with Binance as the leading exchange and Kraken as the lagging exchange, in addition to a test of Coinbase and Hitbtc. The selection of exchanges is simply based on strong and weak lead-lag relationships. Several pairs are tested initially, but the two pairs are included to exemplify the differences in forecasting accuracy when the lead-lag relationship varies in strength. As seen from Figure 6.11, the simple strategy yields a profit

of 2380 USD for the pair of Binance and Kraken. The pair of Coinbase and Hitbtc yields a profit of 487 USD. However, when trading fees are included these results change to extreme losses, as expected. More interesting is the accuracy of the forecast on the lagging exchanges, as the profit is influenced by the number of ticks that are traded on the lagging exchange. This varies between the exchanges, as confirmed by the two plots in Figure 6.11. Kraken follows Binance with an accuracy of 69.15%, but Hitbtc only shows an accuracy of 60.95%. Most of the tested pairs show a forecasting accuracy between 60% and 70%, depending on the strength of the lead-lag relationships. This indicates that the results of the HY approach are trustworthy.



*Figure 6.11 – Next tick strategy. Plots of profit in USD, given mid-quote execution and no trading fees.
Left: Binance and Kraken. Right: Coinbase and Hitbtc.*

6.3.2 ALGORITHM-BASED STRATEGY

This section presents an attempt of implementing a realistic trading strategy based on lead-lag relationships. The main goal of this strategy is to identify when the price difference between the two exchanges is high, and open positions based on this. If the difference is above a given threshold, a position is opened on the lagging exchange and hold until a profit target is reached. The algorithm for the backtesting strategy includes several rules, which operate as requirements for executing trades. These are necessary to avoid challenges that occur when trying to implement profitable arbitrage strategies. First of all, the algorithm will evaluate the price movements on the leading exchange in a given window of time. As the results on the lead-lag relationships yielded lead-lag times of up to 15 seconds, the timeframe will be adjusted after the individual results of the pair. As an example; Coinbase and Kraken have a lead-lag time of 8 seconds. In that situation, the interval is changed to an interval of 10 seconds, to capture the price movements around the time where the lead-lag correlation is highest. The interval needs to be reasonably tight, to avoid bias and movements that are not related to the lead-lag relationships.

The next rule defines the threshold for the price difference between the exchanges. A common mistake

when trading strategies are made is related to the frequency of trades. If the algorithm executes a large number of trades, this will, in the end, lead to trading fees surpassing profits. Hence, a threshold needs to be implemented to avoid entering non-profitable trades. For this algorithm, the threshold will be the same as the profit target, i.e. a predetermined percentage. This percentage threshold is a minimum for executing a trade. When the difference between the two exchanges reaches this threshold, a position is opened. If the price difference continues to move higher above the threshold, the profit target will be updated. This is to avoid opening new positions and instead, stay in the market. When the profit target is reached, the algorithm closes the position. Moreover, if the price on the leading exchange change direction, the position close immediately before the profit target is reached. An example of a successful trade is illustrated below in Figure 6.12.

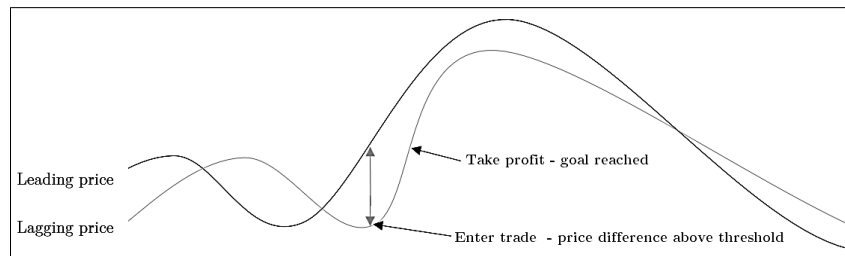


Figure 6.12 - Example of a successful trade by the algorithm.

If true lead-lag relationships are found, the prices of the two exchanges will move similar to the example above. At a certain point, the price difference of the two exchanges reach the threshold, and a long/short position is opened. The lagging exchange will then follow the price movement of the leading exchange, and a profit is achieved.

The trading strategy is tested on two pairs with different lead-lag relationships. For both examples, the initial investment is 1 bitcoin per trade. As of 01.01.2019, that equaled approximately 3,740 USD (CoinMarketCap, 2019). The algorithm will take both long and short positions, to take advantage of price movements in both directions. The threshold and profit target are set to 0.6%. This level was determined after some testing of different thresholds. As describes above, it is important to have a threshold that does not open many unprofitable trades. It cannot be too high either, as it will result in no trades opened. Hence, the threshold of 0.6% will try to take advantage of large price drops or rises, and not profit on small movements. Furthermore, all trading results will be compared with a passive buy-and-hold strategy. That is, buying 1 bitcoin at the beginning of the backtesting period and hold it until the end of the period. This is important, as the goal is to provide a trading strategy that creates a profit superior to the benchmark profit.

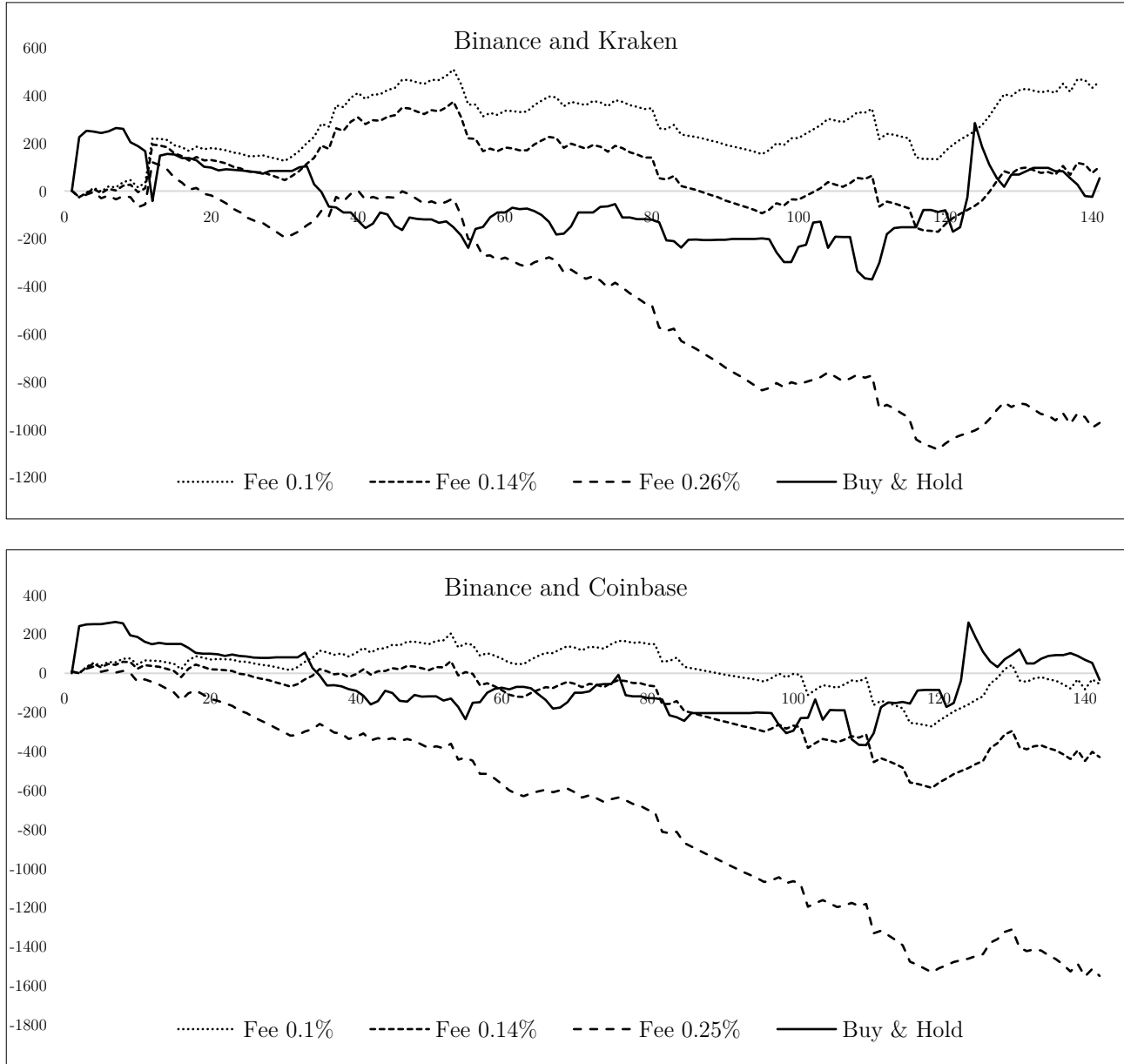


Figure 6.13 – Plots of profit from algorithm-based trading strategy, where trades are executed at the lagging exchanges. 3 tests are done for each pair, with different trading fees. The passive buy and hold strategy is also included.

Figure 6.13 presents the two pairs traded over the period of January through February 2019. For both pairs, the profit of the passive buy-and-hold strategy is included. The results are as expected. Different levels of trading fees have been included to show how important this is for the profit. The first pair of Binance and Kraken have a strong lead-lag relationship. This is reflected in the profit, as the trading strategy beats the benchmark when the trading fees are around 0.1%. Binance and Coinbase, which is a pair with a weak lead-lag relationship, does not yield a profit even with a trading fee of 0.1%. These tests show that when a strong lead-lag relationship is present, the algorithm has more successful trades. The algorithm executes 141 trades at Kraken and only 142 at Coinbase. This is a bit surprising, as one

would expect that the price differences, and hence the trading opportunities, are more present for Binance and Kraken with a strong lead-lag relationship. Moreover, the percentage of the trades that yield a profit on Kraken and Coinbase is 77.3% and 73.2%, respectively. The average profit per successful trade is 20.4 USD at Kraken and 16.4 USD for Coinbase. Hence, the algorithm stays in the positions for a longer period and update the profit target more often at Kraken. The results show that the algorithm works better when the lead-lag relationships are strong, which is a satisfying result.

Nevertheless, it is essential to remark that this is just backtesting and can be troubling. The tests presented are of a subjectively chosen period, where the bitcoin price did not move much. Moreover, the chosen exchange pairs are also subjective. The presented results are reasonable indications of how the trading strategy works, but no conclusion can be done based on the results. Backtesting can be a successful way to test trading strategies without risking any capital. However, a different sample period could reveal other results. These presented pitfalls of backtesting are important to be aware of (Beauty, 2015). Overall, this section shows that arbitrage opportunities could be possible with relatively low trading fees and under the assumption of mid-quote execution. The next section will explore why arbitrage can be challenging to achieve, even with theoretically profitable strategies.

6.3.3 LIMITATION OF ARBITRAGE OPPORTUNITIES

Due to certain constraints and challenges, there is a popular theory in financial markets related to the limits of arbitrage. Pricing inefficiencies that arbitrageurs typically trade on may stay in the market for an extended period of time (Shleifer and Vishny, 1997). This can explain why prices are different across exchanges and are highly relevant in relation to cryptocurrency exchanges. Perhaps the best example of this can be seen from South Korean exchanges. The price of one bitcoin is not in accordance with the other exchanges across the world, with a significant spread (CoinMarketCap, 2019). However, the bitcoin market is highly efficient between the largest exchanges that report real trading volume. Today, average spreads on the top 10 spot exchanges by volume range from 0.01% and 0.10%, and are constrained from falling lower primarily due to exchange fees and tick sizes (Bitwise Asset Management, 2019). Overall, this makes arbitrage cross-exchange trading of bitcoin prices nearly impossible on these exchanges. However, the algorithm-based trading strategy in this thesis seeks to find situations where the price spread of the exchanges deviates from these average levels. These deviations do happen according to the results of the HY-estimator. Since some exchanges lead others, significant movements in price results in a wider spread for a limited period of time. The HY-estimator indicate that price

movements happen 10-15 seconds later at lagging exchanges. Relying on real lead-lag relationships, the lagging exchanges will follow the same movement as the leading exchanges, and arbitrage opportunities arise. The trading results above show how these opportunities yield profits. However, there are other implications that limit these opportunities for arbitrage. These are mainly related to risk and costs. The effect of trading fees is somewhat avoided by the threshold of the algorithm-based strategy, but the results showed that this is far from enough to deal with this problem. Trading fees are high at cryptocurrency exchanges and are the main reason for the limit of arbitrage opportunities. Moreover, there typically are fixed costs by setting up the right infrastructure for this kind of high frequency trading. That is, cloud subscriptions to servers and other services, to enable the fastest execution time possible.

There are several elements of risk with the presented trading strategy. High liquidity is necessary when implementing the arbitrage strategies, as this improves the bid/ask spreads and result in efficient pricing. Moreover, high liquidity makes the order book ticker. That is, more buyers and sellers available at, or close to, the current price. This is where the theory of price slippage and execution risk need to be discussed. Price slippage means that a trade is not executed at the observed (best) price. This can happen both when entering and exiting a position. As the algorithm trade with market orders, which execute at the current price, the possible profit may disappear. In the fraction of a second it takes for an order to reach the exchange, something may change, or the quote could be slightly delayed (Mitchell, 2019). Furthermore, the trade size can cause price slippage. It could be that the volume at the required price is not enough to maintain the current bid/ask spread. Hence, trades can be executed at the second-best price, and the potential arbitrage profit vanishes. The strategies are based on a trade size of 1 bitcoin. The historical volume data, especially from Kraken, is troubling for the trade size. The risk of price slippage seems to be quite high in Kraken, as several minute intervals in the backtesting period did not show a volume of 1 bitcoin. Hence, this liquidity will most likely cause price slippage, and trouble with executing at the desired price. Implementing this strategy with Kraken as the lagging exchange will most likely see a much lower roof for trade size and generate a low profit in absolute values. However, as the historical 1-minute volume at Kraken shows spikes when the price is moving rapidly, it could be a possibility for trading with 1 bitcoin as trade size. To summarize, individual investors may possibly make profits, both in percentage and absolute terms. However, for a large investment firm, the low volume at some of the lagging exchanges will minimize the possibility for scaling, and the arbitrage strategy could be of no use for institutional investors.

7 DISCUSSION

This section of the thesis is important to understand the reasons behind the results in the analysis. Several thought-provoking results are presented, with strong lead-lag relationships in the bitcoin market among the largest cryptocurrency exchanges. As described, these exchanges are highly efficient. What aspects of the cryptocurrency market could be the cause of these relationships?

7.1 REASONS FOR LEAD-LAG RELATIONSHIPS

7.1.1 INFRASTRUCTURE

Probably the most critical factor for a trader is a well-functioning platform. All cryptocurrency exchanges have built up their technology and infrastructure and try to offer the best possible environment for traders. This is dependent on both how the interface works, but most importantly how efficient the technology is. Time is vital in a high frequency trading environment, and customers seek exchanges that have reliable and fast platforms. APIs for trading are widely used in today's financial ecosystem and are emerging in the cryptocurrency market. Traders that set up and integrate trading strategies through algorithms needs APIs to make this work. In simple terms, an API works as a messenger that takes requests and tells a system to do what you want, and finally returns the system's response back to you (Lielacher, 2018). For a trader, APIs allow for direct execution at the exchanges, generally based on pre-set algorithmic models.

All cryptocurrency exchanges in this analysis provide APIs to their customers. This is of interest for both the exchanges and their customers. Sophisticated traders use these APIs to exploit arbitrage opportunities and will over time help the cryptocurrency market to become more efficient and liquid. However, cryptocurrency markets are far from established financial markets, and the number of institutional investors is relatively small. Especially as regulation of cryptocurrency is still in the beginning, institutional money is holding back. Over time, this will probably improve, and the development of better and more secure trading APIs will take place.

There is no doubt that the API services that cryptocurrency exchange provides can be part of the reason for the lead-lag relationships that are seen. Historically, Kraken has been criticized a lot for its infrastructure and service. After months of troubling performance, Kraken updated their infrastructure

in 2018. However, this didn't seem to impress the users, which in fact argued that the service became slower and prone to more errors (Buntinx, 2018). Cryptocurrency is a relatively young financial market, and building trust is essential. As Kraken's reputation for sophisticated traders has been challenged, this could indicate that they move to other exchanges. As fewer investors make use of their trading API to exploit arbitrage opportunities, this could explain some of the lead-lag relationships that are seen with Kraken and other exchanges. However, Kraken is one of the oldest cryptocurrency exchanges and has never experienced hacks, nor lost any customer assets. The exchange is rank as number one on security out of 100 cryptocurrency exchanges, according to the CER Cyber Security Score. Kraken outscores both Coinbase and Binance, which are number two and three on the list, on especially server security (CER, 2019). The full list can be seen in Table 11 in the Appendix. This shows that simple blame of Kraken's reputation as a single reason for a lead-lag relationship may be slightly unrealistic. However, emphasizing exchanges APIs alongside the reputation effect of these as reasons for lead-lag relationships, are important.

As this part of the discussion looks at infrastructure, order speed is also necessary to include. Traders that are engaged in high frequency trading rely on the speed of execution. Only milliseconds can be vital for exploiting arbitrage opportunities and can be the difference between profit and loss. A study on order execution, conducted by the cryptocurrency derivate exchange Deribit, included three of the exchanges that are analyzed in this thesis (Sedgwick, 2018). The results for Binance, Bitfinex and Coinbase are presented in Table 7.1.

Table 7.1 - Execution delay on Binance, Bitfinex and Coinbase.

Show the average execution delay (in milliseconds) per trade and the percentage of trades that have a delay above 1 second.

	Average	> 1 second
Binance	37.2 ms	1.10%
Bitfinex	156 ms	1.50%
Coinbase	33 ms	0.10%

With only 33 milliseconds on average and 0.1% of orders taking longer than 1 second, Coinbase had the fastest speed of order execution. If a trader is dependent on execution within the frame of a second, the trading strategy will fail more often on Binance and Bitfinex, in 1.1% and 1.5% of the cases, respectively. This is perhaps not critical for the profit opportunities for the trading strategy in this thesis but gives a picture of what traders evaluate when choosing platforms to trade on. Hence, one

suggestion can be that sophisticated traders choose exchanges with low execution delay, leading to faster price adjustments at specific exchanges. This can possibly explain some of the lead-lag relationships that are found between cryptocurrency exchanges.

7.1.2 FEE STRUCTURE

Another aspect of the different cryptocurrency exchanges is the fee structures. That is, the cost of executing a trade. A general approach on cryptocurrency exchanges is to reward traders based on trading volume. All exchanges included in this thesis have trading fees structures based on the average monthly trading volume. As an example, Kraken is the most expensive at the first level, with a fee of 0.26% per trade if a person has traded for less than \$50,000 in the last 30 days. The cheapest is Binance with only 0.1% as a starting fee, and only 0.075% if their own cryptocurrency Binance Coin is used for trading fees. The fee level drops to 0.03% at Binance, but that requires a monthly trading volume of 150,000 BTC (approximately \$750 million), which is somewhat unrealistic for most traders. A complete overview of taker and maker fees can be seen in Table 12 in the Appendix, and the two first levels of all exchanges are presented in Table 7.2 below. Maker fees are not included in the discussion, as the maker fee structures follow the changes in taker fee levels for most exchanges. However, remark that maker fees are lower, as traders are rewarded for providing liquidity on the exchanges.

Table 7.2 - Trading fee structures (Taker fee)

	Level 1		Level 2	
	Fee	Volume	Fee	Volume
Binance	0.10%	< 100 BTC	0.08%	< 500 BTC
Bitfinex	0.20%	< \$500,000	0.15%	< \$10 million
Coinbase	0.25%	< \$100,000	0.20%	< \$1 million
Bitstamp	0.25%	< \$20,000	0.24%	< \$100,000
Poloniex	0.20%	< \$1 million	0.15%	< \$20 million
Hitbtc	0.20%	< 100 BTC	0.18%	< 2000 BTC
Kraken	0.26%	< \$50,000	0.24%	< \$100,000

As presented in Section 3.6, previous studies indicate that cheaper markets, i.e. futures and options, tend to lead spot markets and explain some of the lead-lag relationships. However, this analysis cannot make the same conclusion. Binance has a fee structure much lower than any other exchange in this

analysis. However, Binance only shows a leading relationship to 4 out of 6 exchanges. It is, in fact, showing a lagging relationship to both Bitstamp and Bitfinex, which both have much higher fees. Bitstamp showed a leading relationship to all other exchanges but does not have a particularly low fee structure compared to the others. This furthermore weakens the indication that exchanges with low trading fees are leading. Overall, it seems that trading fees as the primary reason for lead-lag relationship are not likely based on this discussion.

7.1.3 LOCATION

The bitcoin market is unique. Not only is the same asset traded on different platforms, but they operate from totally different locations around the world. Some exchanges do also have restrictions on where the customers can be based in the world. This can affect price movements. Moreover, the locations of the servers that an exchange use for order matching are also important. The advantage of being close to the servers of the exchange can be massive in high frequency trading. Where the matching engines of an exchange are located is normally not publicly know. Both Coinbase and Bitfinex offer colocation services, where traders pay a premium to get data center space close to the matching engines of the exchange. This will lead to considerable advantage in executing as fast as possible (Miller, 2018). During 2018 more institutional investor came to the market. These kinds of services started growing, and the price variations across exchange fell significantly. Before 2018 these variations could be as high as 4.5% and is now down to 0.1% (Godshall, 2018).

There is no doubt that location is important, but how can this be a cause for lead-lag relationships? As Binance origin from Hong-Kong, one can perhaps guess that the machine engines are placed in Asia. On the other hand, Coinbase is U.S based, and Bitfinex has servers in Switzerland (Banister, 2019). Widely divided around the globe, this will lead to differences in execution time and hence price movements. The lead-lag relationships in this thesis are on a scale far different from the millisecond adjustments that are handled with colocations. Yet, these kinds of different locations of the exchanges can be part of the reasons why lead-lag relationships are observed.

There is also reason to discuss the location of the customer group at an exchange. The descriptive statistics in Section 6.1 revealed patterns in trading volume. Coinbase showed indications of mostly American traders, as the volume clearly dropped down during night hours in the US. Moreover,

Coinbase only allows customers from certain countries, which excludes most Asian countries. This can also be a part of why lead-lag relationships are seen.

7.1.4 INVESTORS

The location of an investor was discussed above, but more investor characteristics can also be included. The described in Section 6.1 Bitfinex had most trades with a trade size above 1 bitcoin. An exchange with a larger group of sophisticated investors can impact the lead-lag relationship. The regression analysis in Section 6.2.3.5 revealed that new information arrival from sophisticated investors had the most substantial impact. If large blocks of volume come from sophisticated investors on a given exchange, the regression analysis indicated that the lead-lag relationship increases. This could be due to already discussed aspects, for example, the location of servers of sophisticated traders, i.e. fast execution time.

7.1.5 TRADING VOLUME

All the above-mentioned exchange characteristic can lead to differences in price movements and lead-lag relationships. This discussion has been highly hypothetical, but can hopefully help understand the differences that are seen in today's bitcoin market. If all the reasons behind lead-lag relationships were known, the relationships would probably not exist.

Moreover, could the main reason for most of the lead-lag relationships be due to the trading volume at the exchanges? As previously discussed, the HY-estimator should not be biased by the differences in trading volume. Yet, the results point toward exchanges with high volume being leading. It could be that the results are not affected by these differences, but that the real reason behind the relationships is related to volume. Naturally, higher volume indicates more participants in the market and faster-updated prices, which lead to decreasing spreads and more correctly priced assets (Siegel et al., 2000). The trading strategy indicated that the low volume of lagging exchanges caused scaling problems. A possible theory of the present lead-lag relationships could be that arbitrageurs do not have the possibility to apply scaled arbitrage trading due to the liquidity level on the lagging exchanges. Hence, the differences stay in the market, as real limits to arbitrage make it nearly impossible to remove. On

a long-term perspective, as cryptocurrency see an increase in trading volume, these arbitrage opportunities could possibly be exploited.

The above discussion is merely indications. Nonetheless, it gives an interesting perspective to why lead-lag relationships are present on efficient exchanges. It is not unlikely that trading volume and activity could be a big part of the explanations behind the results in this thesis. However, it is more unlikely that this is the only reason.

7.2 A FUTURE PERSPECTIVE

Most studies on lead-lag relationships focus on futures and spot markets. It is widely agreed that futures are the leading component, as noted in the literature review of this thesis. The analysis in Section 6 is only based on the relationships that are found between bitcoin spot exchanges exclusively. An interesting subject for further research could be the emerging futures market for bitcoin.

During 2018 several well-known companies showed interest in bitcoin. The owner of the New York Stock Exchange, Intercontinental Exchange (ICE), will now turn to this new market with a new company called Bakkt. Their plan is to launch a custody solution, in addition to an own exchange for digital assets, and bitcoin futures that are physically settled. These are three major steps in a market that has not seen a lot of large institutional players investing so far. Trusted custody solutions are crucial in a new financial market. New institutional investors need to feel that their assets are safe. As cryptocurrency is held with just a private key that contains several letters and numbers, new investors are skeptical. This is not without reason, with several large cryptocurrency exchanges being hacked throughout the last years. A typical procedure is to send cryptocurrency from an exchange to a wallet, that usually is a more secure place for storage. Hence, trusted services for custody of cryptocurrency is essential for the development of the cryptocurrency market (Arcane Crypto, 2019).

Moreover, the launch of a cryptocurrency exchange will most likely lead to more institutions entering the market. When the owner of the largest stock exchange in the world launches a platform for cryptocurrency, this will be a signal to institutional investors. Perhaps the most exciting part of the plans of ICE is the plan to launch bitcoin futures. As mentioned, these are physically settled. That means physical delivery of bitcoin at the expiration day of the contracts. The bitcoin futures that are

on the market today are cash settled, meaning no actual transactions of bitcoin. Physically settled bitcoin futures will be the first of its kind in the cryptocurrency space, and the trading volume of bitcoin is expected to rise (Aki, 2008).

CBOE is the primary provider of bitcoin futures today. The daily volume is approximately the same as the volume of the largest cryptocurrency exchange, Binance (Bitwise Asset Management, 2019). When Bakkt launches its new bitcoin product, it is likely that trading of bitcoin futures will increase due to the inflow of money from institutional investors. This gives rise for further research from the thesis. After the launch, the bitcoin market will probably have three main components; spot exchanges, cash-settled futures and physically settled futures. Hence, a study on the lead-lag relationships between these three markets can be investigated.

Remembering the aspects of the discussion in the previous sections, some of these are even more interesting in future research. The financial markets today are highly efficient, and high frequency trading is on a totally different level than what is seen in the cryptocurrency markets. A future situation with new bitcoin markets that are mostly traded by institutional investors will most likely increase the number of arbitrageurs and investors seeking opportunities through high frequency trading. This will not only lead to new and improved infrastructures on the new platforms but also challenge the cryptocurrency exchanges that are present today. Investors will seek the platforms that have the highest order speed, the best and most reliable APIs, and the best colocation services.

The Hayashi-Yoshida cross-correlation estimator has shown reliable results in previous studies where lead-lag relationships are found down to milliseconds in traditional financial markets. Hence, continuing this study at a future time would be highly interesting and recommended. The bitcoin environment could look totally different, with new trading platforms in place, higher volumes, and more arbitrageurs trying to exploit profitable trading strategies. The trading strategy presented in this thesis has certain scaling issues due to the low volume of the lagging exchanges. How this strategy would work in a future situation where the overall market volume has increased, could also be interesting to look at. This future institutionalization of the bitcoin market will moreover move the seemingly untouched research area of lead-lag relationships in bitcoin markets closer to previous research results of high frequency lead-lag relationships in traditional financial markets, opening up new opportunities for comparison and improvements.

8 CONCLUSION

This thesis investigated the behavior of the bitcoin price in the year of 2018. By analyzing the price series of seven cryptocurrency exchanges, several interesting results were found.

The analysis presented two different approaches to lead-lag relationships. First, the seven bitcoin price series were analyzed with the use of 1-minute candles. According to expectations, all bitcoin prices showed cointegration in the long run. As all price series represent the same asset, different results would be troubling. The first part of the analysis was ended with a test of the lead-lag relationships in the short-term. All return series were tested in pairs, and all seven exchanges showed bidirectional relationships. That is, all observed return series can be used to predict the other series. This confirms the existence of lead-lag relationships among the most efficient cryptocurrency exchanges.

The second part of the analysis sought to explain the lead-lag relationships on a deeper level, as the first part of the analysis did not provide a satisfactory level of details. This was done by the use of high frequency trade data, including all trades done on the given exchanges during 2018. Strong lead-lag relationships were found, with a time lag time up to 15 seconds. Smaller, less liquid exchanges like Kraken showed a lagging relationship to all the other exchanges. The largest and most liquid exchanges showed weak lead-lag relationships with each other.

The analysis furthermore confirmed that the presented lead-lag relationships were affected by the arrival of new information. Unexpected trading volume, both from sophisticated and non-sophisticated investors showed a significant impact on the lead-lag relationships. Volume from sophisticated investors had the most substantial impact on the lead-lag relationship. However, non-sophisticated investors also had a significant impact on the lead-lag relationships, by providing noise through a large number of small trades.

A simple trading strategy was implemented to show the forecasting accuracy of different lead-lag relationships. As expected, exchanges with a strong leading relationship were better to forecast the movements of the lagging exchanges than pairs with weak lead-lag relationships. This resulted in directional accuracy of up to 70%. Furthermore, a more realistic algorithm-based trading strategy was implemented. This trading strategy showed profitable trading results under the assumptions of low

trading fees and mid-quote executions. However, taking advantage of lead-lag relationships in the cryptocurrency market in real life is challenging due to the limits of arbitrage. High trading fees, significant risk of price slippage and scaling problems due to low liquidity are the most critical challenges.

Several aspects in the cryptocurrency market could affect the lead-lag relationships. The infrastructure of an exchange with reliable technology and effective API services could explain why exchanges like Kraken are lagging behind. Arbitrageurs prefer well-functioning exchanges with low trading fees. The location of the exchanges is also important, and colocation services could play a role. Furthermore, the location of investors and the type of investors that have access to a given exchange could also explain some of the lead-lag relationships. Finally, the big question is; are these relationships found solely because of different trading activity and liquidity at the exchanges? The results of the analysis could point towards this explanation. A complete answer to this is most likely related to several aspects, both those mentioned in this thesis and other aspects. However, further investigating would be valuable.

As an overall conclusion, the most efficient exchanges indicate that high frequency lead-lag relationships are present in the overall bitcoin market. Only the future can tell if this growing financial market will become even more efficient. With higher trading activity, more investors and technological improvements, the observed lead-lag relationships in this thesis could be eliminated.

9 BIBLIOGRAPHY

- Aki, J. (2018). ICE Backed Bakkt Reveals Physical Bitcoin Futures Will be First Crypto Product. Retrieved from <https://blockonomi.com/bakkt-bitcoin-futures/>
- Alsayed, H., & McGroarty, F. (2014). Ultra-High-Frequency Algorithmic Arbitrage Across International Index Futures. *Journal Of Forecasting*, 33(6). doi: 10.1002/for.2298
- Andersen, T., Bollerslev, T., Diebold, F. and Labys, P. (2001). Modeling and Forecasting Realized Volatility. *SSRN Electronic Journal*.
- Arago, V., & Nieto, L. (2005). Heteroskedasticity in the returns of the main world stock exchange indices: volume versus GARCH effects. *Journal Of International Financial Markets, Institutions And Money*, 15(3). doi: 10.1016/j.intfin.2004.06.001
- Arcane Crypto. (2019). The Institutionalisation of Bitcoin. Oslo: Arcane Crypto. Retrieved from <https://drive.google.com/file/d/1v-aPPQwrXfvL8t2kMXOIsmVT3tw4xsxz/view>
- Asteriou, D., & Hall, S. (2007). *Applied Econometrics: a modern approach using eviews and microfit*. New York: Palgrave Macmillan.
- Babayan, D. (2019). Analyst: Trading Crypto on Coinbase is 48x More Expensive than Stock Exchange. Retrieved from <https://www.newsbtc.com/2019/03/28/analyst-trading-crypto-on-coinbase-is-48x-more-expensive-than-stock-exchange/>
- Banister, J. (2019). Institutions Poised to take Crypto Seriously in 2019. Retrieved from <https://cryptonewsreview.com/institutions-poised-to-take-crypto-seriously-in-2019/>
- Bariviera, A. (2017). The inefficiency of Bitcoin revisited: A dynamic approach. *Economics Letters*, 161. doi: 10.1016/j.econlet.2017.09.013
- Beaty, A. (2015). 5 Common Mistakes You Make Backtesting Trading Strategies – The Option Prophet. Retrieved from <https://theoptionprophet.com/blog/5-pitfalls-of>

backtesting-option-strategies

Bitcoin Exchanges. (2019). Retrieved from <https://coinpaprika.com/coin/btcbitcoin/#!exchanges>

Bitcoin markets arbitrage table. (2019). Retrieved from
<https://data.bitcoinity.org/markets/arbitrage/GBP>

Bitwise Asset Management. (2019). Presentation to the U.S. Securities and Exchange Commission. Bitwise Asset Management. Retrieved from
<https://www.sec.gov/comments/sr-nysearca-2019-01/srnysearca201901-5164833183434.pdf>

Black, F. (1986). Noise. *The Journal Of Finance*, 41(3). doi: 10.2307/2328481

Bollerslev, T., Chou, R., & Kroner, K. (1992). ARCH modeling in finance: a review of the theory and empirical evidence. *Journal Of Econometrics*, 51(1), 5-59. doi: 10.1016/0304-4076(92)90064-X

Bowerman, B., O'Connell, R., & Koehler, A. (2005). *Forecasting, time series, and regression* (4th ed.). Brooks/Cole.

Brandvold, M., Molnár, P., Vagstad, K., & Andreas Valstad, O. (2015). Price discovery on Bitcoin exchanges. *Journal Of International Financial Markets, Institutions And Money*, 36. doi: 10.1016/j.intfin.2015.02.010

Brooks, C. (2008). *Introductory Econometrics for Finance*. New York: Cambridge University Press.

Brooks, C., Garrett, I., & Hinich, M. (1999). An alternative approach to investigating lead-lag relationships between stock and stock index futures markets. *Applied Financial Economics*, 9(6). doi: 10.1080/096031099332050

Brooks, C., Rew, A., & Ritson, S. (2001). A trading strategy based on the lead-lag relationship

between the spot index and futures contract for the FTSE 100. *International Journal Of Forecasting*, 17(1). doi: 10.1016/s0169-2070(00)00062-5

Bryman, A., Bell, E. (2011). *Business Research methods*. Oxford: Oxford University Press.

Bulmer, M. (1979). *Principles of statistics*. Cambridge, Mass.: M.I.T. Press.

Buntinx, J. (2018). Cryptocurrency Exchange Review: Kraken - The Merkle Hash. Retrieved from <https://themerke.com/cryptocurrency-exchange-review-kraken/>

Castor, A. (2018). Warning Signs? A Timeline of Tether and Bitfinex Events. Retrieved from <https://bitcoinmagazine.com/articles/warning-signs-timeline-tether-and-bitfinex-events/>

Catania, L., & Sandholdt, M. (2019). Bitcoin at High Frequency. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3309565

CER. (2019). TOP 100 CRYPTO EXCHANGES ACCORDING TO THE CER CYBER SECURITY SCORE (CSS). Retrieved from <https://blog.cer.live/analytical-assessments/top-100-exchanges/>

Chan, K. (1992). A Further Analysis of the Lead-Lag Relationship Between the Cash Market and Stock Index Futures Market. *Review Of Financial Studies*, 5(1). doi: 10.1093/rfs/5.1.123

Chen, Y., & Gau, Y. (2009). Tick sizes and relative rates of price discovery in stock, futures, and options markets: Evidence from the Taiwan stock exchange. *Journal Of Futures Markets*, 29(1). doi: 10.1002/fut.20319

Chiang, R., & Fong, W. (2001). Relative informational efficiency of cash, futures, and options markets: The case of an emerging market. *Journal Of Banking & Finance*, 25(2). doi: 10.1016/s0378-4266(99)00127-2

Chu, T., Danks, D., & Glymour, C. (2005). Data Driven Methods for Granger Causality and

- Contemporaneous Causality with Non-Linear Corrections: Climate Teleconnection Mechanisms [Ebook]. Retrieved from https://www.researchgate.net/publication/249984827_Data_Driven_Methods_for_Granger_Causality_and_Contemporaneous_Causality_with_Non_Linear_Corrections_Climate_Teleconnection_Mechanisms
- Coinbase Inc. (2019). Retrieved from <https://www.bloomberg.com/profiles/companies/0776164D:US-coinbase-inc>
- CoinMarketCap. (2019). Bitcoin. Retrieved from <https://coinmarketcap.com/currencies/bitcoin/>
- Dacorogna, M. (2001). An introduction to high-frequency finance. San Diego: Acad. Press.
- Dale, O. (2018). What Is Bakkt & How Will it Change the Cryptocurrency World?. Retrieved from <https://blockonomi.com/what-is-bakkt/>
- Dao, T., McGroarty, F., & Urquhart, A. (2018). Ultra-high-frequency lead-lag relationship and information arrival. *Quantitative Finance*, 18(5). doi: 10.1080/14697688.2017.1414484
- De Long, J., Shleifer, A., Summers, L., & Waldmann, R. (1990). Noise Trader Risk in Financial Markets. *Journal Of Political Economy*, 98(4). doi: 10.1086/261703
- Dickey, D., & Fuller, W. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal Of The American Statistical Association*, 74(366). doi: 10.2307/2286348
- Dougherty, C., & Huang, G. (2014). Mt. Gox Seeks Bankruptcy After \$480 Million Bitcoin Loss. Retrieved from <https://www.bloomberg.com/news/articles/2014-02-28/mt-gox-exchange-files-for-bankruptcy>
- Enders, W. (2008). *Applied econometric time series* (3rd ed.). New York: John Wiley & Sons, Inc.

- Engle, R. (2000). The Econometrics of Ultra-high-frequency Data. *Econometrica*, 68(1). doi: 10.1111/1468-0262.00091
- Engle, R., & Granger, C. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55(2). doi: 10.2307/1913236
- Epps, T. (1979). Comovements in Stock Prices in the Very Short Run. *Journal of the American Statistical Association*, 74(366).
- Ergün, A. (2009). NYSE Rule 80A restrictions on index arbitrage and market linkage. *Applied Financial Economics*, 19(20). doi: 10.1080/09603100802599613
- Eross, A., McGroarty, F., Urquhart, A., & Wolfe, S. (2017). The Intraday Dynamics of Bitcoin. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3013699
- Fama, E. (1965). Random Walks in Stock Market Prices. *Financial Analysts Journal*, 21(5).
- Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal Of Finance*, 25(2). doi: 10.2307/2325486
- Fleming, J., Ostdiek, B., & Whaley, R. (1996). Trading costs and the relative rates of price discovery in stock, futures, and option markets. *Journal Of Futures Markets*, 16(4). doi: 10.1002/(sici)1096-9934(199606)16:4<353::aid-fut1>3.0.co;2-h
- Franco, Pedro (2015). Understanding Bitcoin. *Cryptography, Engineering and Economics*. John Wiley & Sons, pp. 95–122.
- Godshall, J. (2018). Report Suggests Wall Street's Influence has Drastically Improved Price Variations Between Cryptocurrency Exchanges - UNHASHED. Retrieved from <https://unhashed.com/cryptocurrency-news/wall-street-price-variations-cryptocurrency-exchanges/>

- Goodhart, C., & O'Hara, M. (1997). High frequency data in financial markets: Issues and applications. *Journal Of Empirical Finance*, 4(2-3). doi: 10.1016/s0927-5398(97)000030
- Granger, C. (1988). Some recent development in a concept of causality. *Journal Of Econometrics*, 39(1-2), 199-211. doi: 10.1016/0304-4076(88)90045-0
- Granger, C., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal Of Econometrics*, 2(2). doi: 10.1016/0304-4076(74)90034-7
- Grünbichler, A., Longstaff, F., & Schwartz, E. (1994). Electronic Screen Trading and the Transmission of Information: An Empirical Examination. *Journal Of Financial Intermediation*, 3(2). doi: 10.1006/jfin.1994.1002
- Gujarati, D., & Porter, D. (2009). *Basic econometrics*. London: McGraw-Hill Irwin.
- Hayashi, T., & Yoshida, N. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 11(2). doi: 10.3150/bj/1116340299
- Hoffmann, M., Rosenbaum, M., & Yoshida, N. (2013). Estimation of the lead-lag parameter from non-synchronous data. *Bernoulli*, 19(2). doi: 10.3150/11-bej407
- Huth, N., & Abergel, F. (2012). *High Frequency Lead/lag Relationships Empirical facts*[Ebook]. Retrieved from <https://arxiv.org/ftp/arxiv/papers/1111/1111.7103.pdf>
- Jenssen, T. (2019). Bitcoin – Store norske leksikon. [online] Store norske leksikon. Available at: <https://snl.no/Bitcoin> [Accessed 6 Mar. 2019].
- Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal Of Economic Dynamics And Control*, 12(2-3). doi: 10.1016/0165-1889(88)90041-3
- Johansen, S., & Juselius, K. (1990). MAXIMUM LIKELIHOOD ESTIMATION AND INFERENCE

ON COINTEGRATION - WITH APPLICATIONS TO THE DEMAND FOR MONEY. Oxford Bulletin Of Economics And Statistics, 52(2). doi: 10.1111/j.1468-0084.1990.mp52002003.x

Kammers, E. (2017). Cointegration, Correlation, and Log Returns. Retrieved from <https://www.rbloggers.com/cointegration-correlation-and-log-returns/>

Kawaller, I., Koch, P., & Koch, T. (1987). The Temporal Price Relationship Between S&P 500 Futures and the S&P 500 Index. The Journal Of Finance, 42(5). doi: 10.2307/2328529

Keene, O. (1995). The log transformation is special. Statistics In Medicine, 14(8), 811-819. doi: 10.1002/sim.4780140810

Lielacher, A. (2018). What is API trading and how is it applied to crypto?. Retrieved from <https://bravenewcoin.com/insights/what-is-api-trading-and-how-is-it-applied-to-crypto>

Liu, Y., & Bahadori, M. (2012). A Survey on Granger Causality: A Computational View. University of Southern California.

Lopez de Prado, M. (2019). The 7 Reasons Most Econometric Investments Fail. Presentation, Cornell University.

Madhavan, A., & Sofianos, G. (1998). An empirical analysis of NYSE specialist trading. Journal Of Financial Economics, 48(2). doi: 10.1016/S0304-405X(98)00008-7

Markets API. Retrieved from <https://bitcoincharts.com/about/markets-api/>

Martikainen, T., Perttunen, J., & Puttonen, V. (1995). On the dynamics of stock index futures and individual stock returns. Journal Of Business Finance & Accounting, 22(1). doi: 10.1111/j.1468-5957.1995.tb00673.x

Matthew, N. and Stones, R. (2008). Beginning Linux Programming. 4th ed. Indianapolis, Indiana: Wiley Publishing, Inc.

- Maziarz, M. (2015). A review of the Granger-causality fallacy. *The Journal Of Philosophical Economics : Reflections On Economic And Social Issues*, 8(2), 85-105. Retrieved from <https://hrcak.srce.hr/155919>
- McDonald, R., Cassano, M., & Fahlenbrach, R. (2006). *Derivatives markets*. Boston: Pearson Education Inc.
- Miller, R. (2018). Coinbase Plans Low-Latency Colocation for Bitcoin Trading. Retrieved from <https://datacenterfrontier.com/coinbase-plans-low-latency-colocation-for-cryptocurrency-trading/>
- Mitchell, C. (2019). Reducing Order Slippage While Trading. Retrieved from <https://www.thebalance.com/day-trading-slippage-defined-1030866>
- Nadarajah, S., & Chu, J. (2017). On the inefficiency of Bitcoin. *Economics Letters*, 150. doi: 10.1016/j.econlet.2016.10.033
- Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*. Retrieved from <https://bitcoin.org/bitcoin.pdf>
- Nam, S., Oh, S., & Kim, H. (2008). The time difference effect of a measurement unit in the lead-lag relationship analysis of Korean financial market. *International Review Of Financial Analysis*, 17(2). doi: 10.1016/j.irfa.2006.09.004
- Nam, S., Oh, S., Kim, H., & Kim, B. (2006). An empirical analysis of the price discovery and the pricing bias in the KOSPI 200 stock index derivatives markets. *International Review Of Financial Analysis*, 15(4-5). doi: 10.1016/j.irfa.2006.02.003
- Newbold, P., Carlson, W., & Thorne, B. (2013). *Statistics for business and economics*. Harlow, Essex: Pearson Education.
- Phillip, A., Chan, J., & Peiris, S. (2018). A new look at Cryptocurrencies. *Economics Letters*, 163. doi: 10.1016/j.econlet.2017.11.020

Regulation of Cryptocurrency Around the World. (2018). Retrieved from

<https://www.loc.gov/law/help/cryptocurrency/cryptocurrency-world-survey.pdf>

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research Methods for Business Students* (7th ed.).

Harlow: Pearson Education.

Schneier, B. (1996). *Applied Cryptography, Second Edition: Protocols, Algorithms, and Source Code in C* (2nd ed.). John Wiley & Sons, Inc.

Sedgwick, K. (2018). Order Speed Analysis Reveals the Fastest Cryptocurrency Exchanges – Bitcoin News. Retrieved from <https://news.bitcoin.com/order-speed-analysis-reveals-the-fastest-cryptocurrency-exchanges/>

Shell, A. (2007). Technology squeezes out real, live traders - USATODAY.com. Retrieved from

https://usatoday30.usatoday.com/money/markets/2007-07-11-nyse-traders_N.htm

Shleifer, A., & Vishny, R. (1997). The Limits of Arbitrage. 52, 1, 35-55. doi: 10.1111/j.1540-6261.1997.tb03807.x

Shyy, G., Vijayraghavan, V., & Scott-Quinn, B. (1996). A further investigation of the lead-lag relationship between the cash market and stock index futures market with the use of bid/ask quotes: The case of France. *Journal Of Futures Markets*, 16(4). doi: 10.1002/(sici)1096-9934(199606)16:4<405::aid-fut3>3.0.co;2-m

Siegel, J., Shim, J., Qureshi, A., & Brauchler, J. (2000). *International Encyclopedia of Technical Analysis*. Taylor and Francis.

Smith, M., 2011. *Research Methods in Accounting*. 2 ed. London: SAGE Publications.

Stoll, H., & Whaley, R. (1990). The Dynamics of Stock Index and Stock Index Futures Returns. *The Journal Of Financial And Quantitative Analysis*, 25(4). doi: 10.2307/2331010

Tasca, P., & Tessone, C. (2019). *A Taxonomy of Blockchain Technologies: Principles of*

- Identification and Classification. Retrieved from
<http://ledger.pitt.edu/ojs/index.php/ledger/article/view/140/118>
- Torres-Reyna, O. (2007). Linear Regression using Stata [Ebook]. Princeton University. Retrieved
from <https://www.princeton.edu/~otorres/Regression101.pdf>
- Toth, B. and Kertesz, J. (2009). The Epps effect revisited. *Quantitative Finance*, 9(7).
- Urquhart, A. (2016). The Inefficiency of Bitcoin. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2828745
- USC Libraries. (2019). Research Guides: Organizing Your Social Sciences Research Paper: Types of
Research Designs. [online] Available at: <http://libguides.usc.edu/writingguide/researchdesigns>
- Wasserstein, R, Schirm, A. & Lazar, N. (2019) Moving to a World Beyond “ $p < 0.05$ ”, *The American Statistician*, 73(1), 1-19, DOI: 10.1080/00031305.2019.1583913
- Wasserstein, R., & Lazar, N. (2016). The ASA's Statement on p-Values: Context, Process, and
Purpose. *The American Statistician*, 70(2), 129-133. doi: 10.1080/00031305.2016.1154108
- Why Kraken?. Retrieved from <https://www.kraken.com/en-us/why-kraken>
- Williams, R. (2015). Multicollinearity. Retrieved from <https://www3.nd.edu/~rwilliam/>
- Williams-Grut, O. (2018). A South Korean gaming company is said to be in talks to buy Bitstamp,
the world's oldest bitcoin exchange. Retrieved from
<https://www.businessinsider.com/gaming-group-nexon-seeking-to-buy-cryptocurrency-exchange-bitstamp-2018-4?r=US&IR=T&IR=T>
- Xu, J., & Livshits, B. (2018). The Anatomy of a Cryptocurrency Pump-and-Dump Scheme.
Retrieved from
https://arxiv.org/abs/1811.10109?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+arxiv%2FQ5Xk+%28ExcitingAds%21+cs+updates+on+arXiv.org%29

10 APPENDIX

10.1 TABLES

TABLE 1 – DESCRIPTIVE STATISTICS PRICES

	Mean	Median	Max	Min	Std.Dev	Skew	Kurtosis
Binance	7,539	6,910	17,175	3,158	2,382	0.999	1.886
Coinbase	7,522	6,900	17,178	3,130	2,405	1.005	1.865
Bitstamp	7,525	6,902	17,235	3,124	2,408	1.001	1.835
Bitfinex	7,552	6,910	17,248	3,215	2,374	1.040	1.937
Kraken	7,533	6,899	17,212	3,124	2,424	1.016	1.872
Hitbtc	7,606	7,018	17,204	3,247	2,394	0.975	1.723
Poloniex	7,543	6,910	17,250	3,158	2,388	1.005	1.897

TABLE 2 – DESCRIPTIVE STATISTICS VOLUME

	Mean	Median	Max	Min	Std.Dev	Skew	Kurtosis
Binance	25.31	15.83	1235.77	0	34.2587	6.468	83.132
Coinbase	9.44	3.05	964.14	0	22.6800	9.039	147.195
Bitstamp	7.48	2.09	747.43	0	17.9456	8.199	124.039
Bitfinex	24.41	7.06	6717.52	0	65.3021	13.915	509.697
Kraken	4.20	0.76	822.83	0	12.3078	12.382	332.795
Hitbtc	7.88	2.17	Kraken	0	20.3841	7.187	76.291
Poloniex	1.59	0.16	402.99	0	4.9302	13.921	506.834

TABLE 3 – CORRELATION MATRIX PRICES

	Binance	Bitfinex	Bitstamp	Kraken	Poloniex	Coinbase	Hitbtc
Binance	1	0.99978	0.99947	0.99910	0.99987	0.99940	0.99915
Bitfinex	0.99978	1	0.99968	0.99932	0.99991	0.99962	0.99918
Bitstamp	0.99947	0.99968	1	0.99950	0.99964	0.99996	0.99910
Kraken	0.99910	0.99932	0.99950	1	0.99929	0.99950	0.99865
Poloniex	0.99987	0.99991	0.99964	0.99929	1	0.99958	0.99923
Coinbase	0.99940	0.99962	0.99996	0.99950	0.99958	1	0.99905
Hitbtc	0.99915	0.99918	0.99910	0.99865	0.99923	0.99905	1

TABLE 4 – CORRELATION MATRIX RETURNS

	Binance	Bitfinex	Bitstamp	Kraken	Poloniex	Coinbase	Hitbtc
Binance	1	0.7716	0.6247	0.4744	0.5057	0.7020	0.6537
Bitfinex	0.7716	1	0.6937	0.5438	0.5667	0.7670	0.7328
Bitstamp	0.6247	0.6937	1	0.5015	0.4911	0.6850	0.6059
Kraken	0.4744	0.5438	0.5015	1	0.4771	0.5663	0.5323
Poloniex	0.5057	0.5667	0.4911	0.4771	1	0.5556	0.5428
Coinbase	0.7020	0.7670	0.6850	0.5663	0.5556	1	0.6743
Hitbtc	0.6537	0.7328	0.6059	0.5323	0.5428	0.6743	1

TABLE 5 – STATIONARITY OF RETURNS SERIES

Note: Critical values, 1%: -3.959, 5%: -3.410, 10%: -3.127. Constant and trend are not included based on plots of the return series. Lag length is chosen by the use of SBIC.

	Binance	Bitfinex	Bitstamp	Kraken	Poloniex	Coinbase	Hitbtc
ADF test statistic	-532.76	-521.19	-539.766	-422.72	-370.907	-414.934	-702.8035
P-value	0	0	0	0	0	0	0
Lags	1	1	1	2	3	2	0
H0	Rejected	Rejected	Rejected	Rejected	Rejected	Rejected	Rejected

TABLE 6 – LEAD-LAG RELATIONSHIPS (10 LAGS)

X	Y	Seconds	Correlation	LLR
Binance	Bitfinex	-1	0.01741	0.8194
Binance	Bitstamp	-3	0.01325	0.9677
Binance	Coinbase	-1	0.01804	1.0009
Binance	Hitbtc	2	0.03707	1.1734
Binance	Kraken	7	0.01806	1.4961
Bitstamp	Bitfinex	-1	0.01612	0.9782
Bitstamp	Hitbtc	4	0.03877	1.1548
Bitstamp	Kraken	10	0.02184	1.4718
Coinbase	Bitfinex	-2	0.03292	0.8111
Coinbase	Bitstamp	-1	0.03620	0.7976
Coinbase	Hitbtc	3	0.08644	1.1061
Coinbase	Kraken	8	0.04584	1.3135
Hitbtc	Bitfinex	-3	0.02954	0.7696
Hitbtc	Kraken	8	0.03738	1.2554
Kraken	Bitfinex	-8	0.01212	0.6613
Poloniex	Binance	-7	0.00269	0.9683
Poloniex	Bitfinex	-7	0.01157	0.8252
Poloniex	Bitstamp	-7	0.01284	0.8703
Poloniex	Coinbase	-5	0.01449	0.9082
Poloniex	Hitbtc	0	0.03640	0.9782
Poloniex	Kraken	4	0.01983	1.1558

TABLE 7 – ASSUMPTIONS: ORIGINAL REGRESSION MODEL**Homoscedasticity – Poloniex and Hitbtc**

Reg 1 (Y= Correlation coefficient)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of Max_PoloniexHitbtc

chi2(1)      =    41.09
Prob > chi2   =    0.0000
```

Reg 2 (Y= Lag time)

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of Time_PoloniexHitbtc

chi2(1)      =    13.69
Prob > chi2   =    0.0002
```

Reg 3 (Y= Lead-lag ratio)

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of LLR_PoloniexHitbtc

chi2(1)      =     1.32
Prob > chi2   =    0.2506

.
```

Homoscedasticity – Bitstamp and Bitfinex

Reg 1 (Y= Correlation coefficient)

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of Max_BitstampBitfinex

chi2(1)      =    57.48
Prob > chi2   =    0.0000
```

Reg 2 (Y= Lag time)

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of Time_BitstampBitfinex

chi2(1)      =     0.23
Prob > chi2   =    0.6331

.
```

Reg 3 (Y= Lead-lag ratio)

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of LLR_BitstampBitfinex

chi2(1)      =    36.08
Prob > chi2   =    0.0000
```

Homoscedasticity – Kraken and Bitfinex

Reg 1 (Y= Correlation coefficient)

```
. estat hettest  
  
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of Max_KrakenBitfinex  
  
chi2(1)      =    7.89  
Prob > chi2   =   0.0050  
  
.
```

Reg 2 (Y= Lag time)

```
. estat hettest  
  
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of Time_KrakenBitfinex  
  
chi2(1)      =    0.05  
Prob > chi2   =   0.8302  
  
.
```

Reg 3 (Y= Lead-lag ratio)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of LLR_KrakenBitfinex  
  
chi2(1)      =   22.29  
Prob > chi2   =   0.0000  
  
.
```

Homoscedasticity – Bitstamp and Kraken

Reg 1 (Y= Correlation coefficient)

```
. estat hettest  
  
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of Max_BitstampKraken  
  
chi2(1)      =    7.49  
Prob > chi2   =   0.0062  
  
.
```

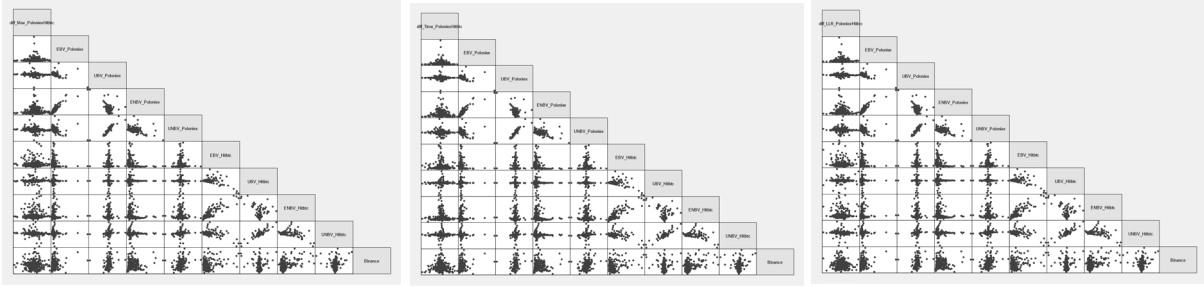
Reg 2 (Y= Lag time)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of Time_BitstampKraken  
  
chi2(1)      =   51.71  
Prob > chi2   =   0.0000  
  
.
```

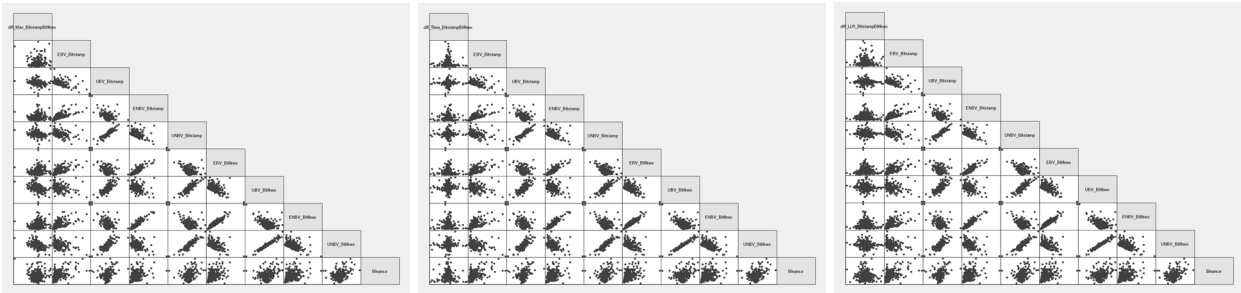
Reg 3 (Y= Lead-lag ratio)

```
. estat hettest  
  
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of LLR_BitstampBitfinex  
  
chi2(1)      =   36.08  
Prob > chi2   =   0.0000  
  
.
```

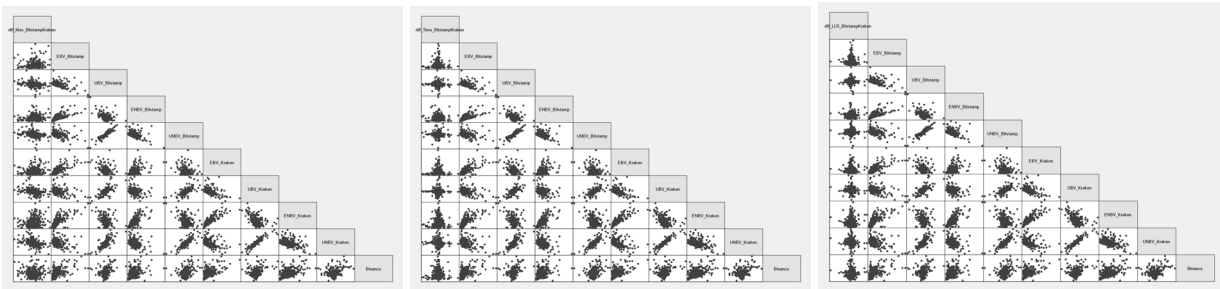
Linearity – Poloniex and Hitbtc (Reg 1,2,3 from left)



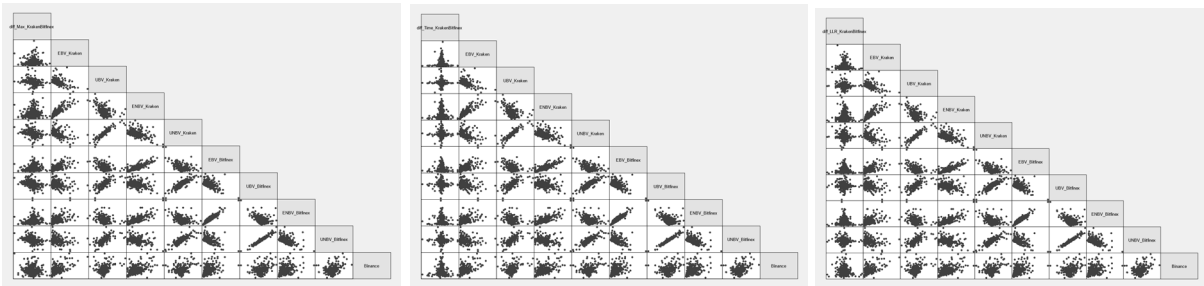
Linearity – Bitstamp and Bitfinex (Reg 1,2,3 from left)



Linearity – Bitstamp and Kraken (Reg 1,2,3 from left)



Linearity – Kraken and Bitfinex (Reg 1,2,3 from left)



Normality – Poloniex and Hitbtc

Reg 1 (Y= Correlation coefficient)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.0000	0.0002	48.67	0.0000

Reg 2 (Y= Lag time)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.5422	0.0000	34.90	0.0000

Reg 3 (Y= Lead-lag ratio)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.0000	0.0001	44.24	0.0000

```
.
```

Normality – Bitstamp and Bitfinex

Reg 1 (Y= Correlation coefficient)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.0000	0.0001	31.67	0.0000

Reg 2 (Y= Lag time)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.0001	0.0000	.	0.0000

Reg 3 (Y= Lead-lag ratio)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.0000	0.0007	31.45	0.0000

```
.
```

Normality – Kraken and Bitfinex

Reg 1 (Y= Correlation coefficient)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.0000	0.0000	58.82	0.0000

Reg 2 (Y= Lag time)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.0000	0.0000	64.04	0.0000

```
.
```

Reg 3 (Y= Lead-lag ratio)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.0000	0.0000	64.04	0.0000

```
.
```

Normality – Bitstamp and Kraken

Reg 1 (Y= Correlation coefficient)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.0000	0.2449	19.94	0.0000

```
.
```

Reg 2 (Y= Lag time)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	364	0.2244	0.0000	25.12	0.0000

```
.
```

Reg 3 (Y= Lead-lag ratio)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	363	0.0156	0.0000	59.81	0.0000

Independent residuals – Poloniex and Hitbtc (Reg 1 from top)

```
. estat dwatson
Durbin-Watson d-statistic( 10, 364) = .5761355

. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.886446

. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.379591
```

Independent residuals – Bitstamp and Bitfinex (Reg 1 from top)

```
. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.157975

. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.899976
.

. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.321461
.
end of do-file
```

Independent residuals – Bitstamp and Kraken (Reg 1 from top)

```
. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.230137

. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.552446

. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.321461
```

Independent residuals – Kraken and Bitfinex (Reg 1 from top)

```
. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.648528
.

. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.656382

. dwstat
Durbin-Watson d-statistic( 10, 364) = 1.136818
```

Multicollinearity – Poloniex and Hitbtc

. vif

Variable	VIF	1/VIF
ENBV_Hitbtc	8.79	0.113713
EBV_Hitbtc	7.25	0.137909
UNBV_Polon~x	6.68	0.149672
ENBV_Polon~x	6.20	0.161411
UBV_Poloniex	5.94	0.168438
EBV_Poloniex	5.91	0.169167
UBV_Hitbtc	5.75	0.173833
UNBV_Hitbtc	5.69	0.175799
Binance	2.30	0.434546
Mean VIF	6.06	

Multicollinearity –Bitstamp and Bitfinex

. vif

Variable	VIF	1/VIF
UNBV_Bitfi~x	24.27	0.041201
ENBV_Bitfi~x	23.61	0.042351
UBV_Bitfinex	19.99	0.050026
ENBV_Bitst~p	15.94	0.062744
EBV_Bitfinex	15.63	0.063996
UNBV_Bitst~p	12.78	0.078234
UBV_Bitstamp	5.62	0.178033
EBV_Bitstamp	4.90	0.204189
Binance	3.21	0.311654
Mean VIF	13.99	

Multicollinearity – Kraken and Bitfinex

. vif

Variable	VIF	1/VIF
UNBV_Bitfi~x	22.32	0.044805
UBV_Bitfinex	21.41	0.046708
ENBV_Kraken	18.80	0.053188
EBV_Bitfinex	17.39	0.057510
UNBV_Kraken	17.03	0.058734
ENBV_Bitfi~x	16.90	0.059181
EBV_Kraken	11.18	0.089412
UBV_Kraken	9.52	0.105032
Binance	3.34	0.299143
Mean VIF	15.32	

Multicollinearity –Bitstamp and Kraken

. vif

Variable	VIF	1/VIF
ENBV_Kraken	14.72	0.067948
UNBV_Kraken	14.26	0.070121
EBV_Kraken	11.22	0.089093
UBV_Kraken	8.76	0.114123
UNBV_Bitst~p	8.25	0.121275
ENBV_Bitst~p	7.28	0.137283
EBV_Bitstamp	5.78	0.173156
UBV_Bitstamp	5.75	0.174022
Binance	3.24	0.309109
Mean VIF	8.81	

TABLE 8 - ASSUMPTIONS: ADJUSTED REGRESSION MODELS**Homoscedasticity – Poloniex and Hitbtc**

Reg 1 (Y= Correlation coefficient)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_Max_PoloniexHitbtc

chi2(1)      =      0.50
Prob > chi2   =      0.4809
```

Reg 2 (Y= Lag time)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_Time_PoloniexHitbtc

chi2(1)      =      0.10
Prob > chi2   =      0.7561
```

Reg 3 (Y= Lead-lag ratio)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_LLRL_PoloniexHitbtc

chi2(1)      =      2.22
Prob > chi2   =      0.1365
```

Homoscedasticity – Bitstamp and Bitfinex

Reg 1 (Y= Correlation coefficient)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_Max_BitstampBitfinex

chi2(1)      =      6.03
Prob > chi2   =      0.0140
```

Reg 2 (Y= Lag time)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_Time_BitstampBitfinex

chi2(1)      =      2.68
Prob > chi2   =      0.1018
```

Reg 3 (Y= Lead-lag ratio)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_LLRL_BitstampBitfinex

chi2(1)      =      0.78
Prob > chi2   =      0.3784
```

Homoscedasticity – Kraken and Bitfinex

Reg 1 (Y= Correlation coefficient)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_Max_KrakenBitfinex

chi2(1)      =      2.57
Prob > chi2   =    0.1090
```

Reg 2 (Y= Lag time)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_Time_KrakenBitfinex

chi2(1)      =      0.32
Prob > chi2   =    0.5706
```

Reg 3 (Y= Lead-lag ratio)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_LLR_KrakenBitfinex

chi2(1)      =      3.09
Prob > chi2   =    0.0789
```

Homoscedasticity – Bitstamp and Kraken

Reg 1 (Y= Correlation coefficient)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_Max_BitstampKraken

chi2(1)      =      0.02
Prob > chi2   =    0.8982
```

Reg 2 (Y= Lag time)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_Time_BitstampKraken

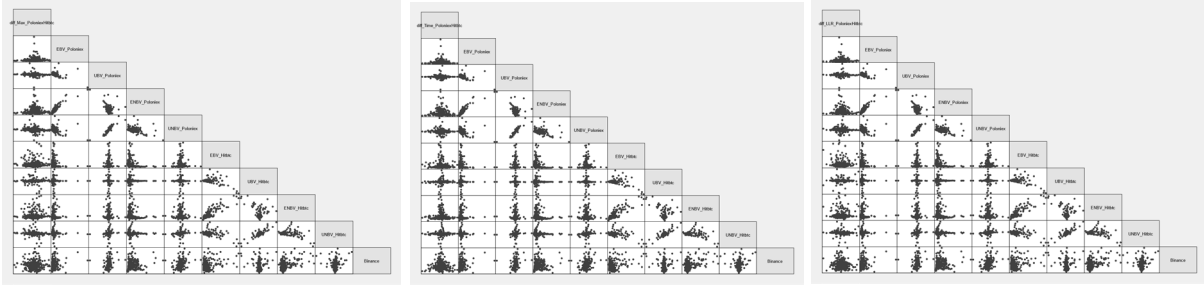
chi2(1)      =      0.30
Prob > chi2   =    0.5838
```

Reg 3 (Y= Lead-lag ratio)

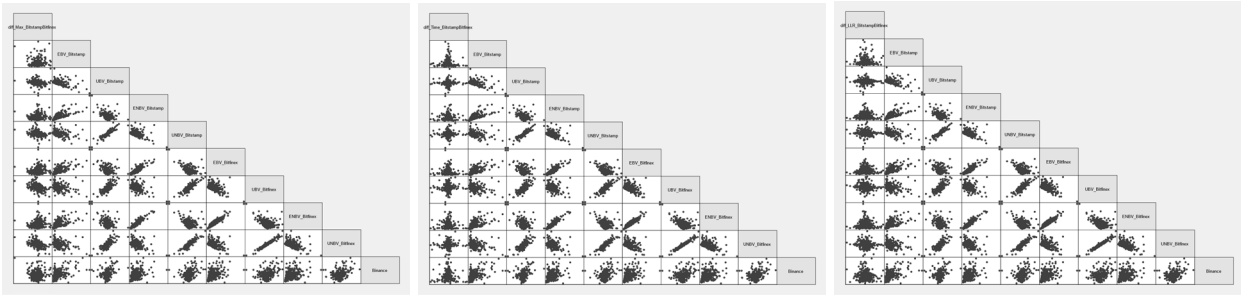
```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of diff_LLR_BitstampKraken

chi2(1)      =      0.03
Prob > chi2   =    0.8563
```

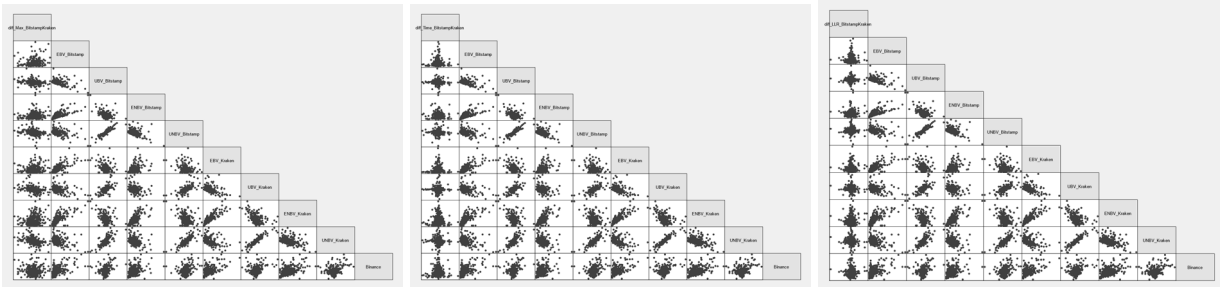
Linearity – Poloniex and Hitbtc (Reg 1,2,3 from left)



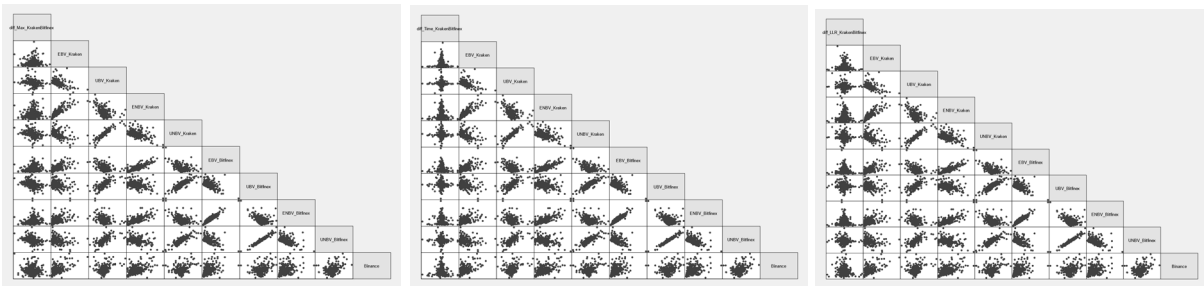
Linearity – Bitstamp and Bitfinex (Reg 1,2,3 from left)



Linearity – Bitstamp and Kraken (Reg 1,2,3 from left)



Linearity – Kraken and Bitfinex (Reg 1,2,3 from left)



Normality – Poloniex and Hitbtc

Reg 1 (Y= Correlation coefficient)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.1489	0.0000	22.07	0.0000

Reg 2 (Y= Lag time)

```
. sktest e
```

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.2686	0.0000	27.38	0.0000

Reg 3 (Y= Lead-lag ratio)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.0397	0.0000	33.20	0.0000

Normality – Bitstamp and Bitfinex

Reg 1 (Y= Correlation coefficient)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.5989	0.0000	22.75	0.0000

Reg 2 (Y= Lag time)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.4168	0.0000	55.01	0.0000

Reg 3 (Y= Lead-lag ratio)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.1776	0.0000	21.62	0.0000

Normality – Kraken and Bitfinex

Reg 1 (Y= Correlation coefficient)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.8039	0.0000	16.70	0.0002

Reg 2 (Y= Lag time)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.0002	0.0000	72.09	0.0000

Reg 3 (Y= Lead-lag ratio)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.1129	0.0000	42.93	0.0000

Normality – Bitstamp and Kraken

Reg 1 (Y= Correlation coefficient)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.0034	0.0000	26.29	0.0000

Reg 2 (Y= Lag time)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.1684	0.0000	43.72	0.0000

Reg 3 (Y= Lead-lag ratio)

Skewness/Kurtosis tests for Normality					
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	joint Prob>chi2
e	363	0.0156	0.0000	59.81	0.0000

Independent residuals – Poloniex and Hitbtc (Reg 1 from top)

```
. dwstat
Durbin-Watson d-statistic( 10, 363) = 2.753008

. dwstat
Durbin-Watson d-statistic( 10, 363) = 2.995245

. dwstat
Durbin-Watson d-statistic( 10, 363) = 3.006312
```

Independent residuals – Bitstamp and Bitfinex (Reg 1 from top)

```
Durbin-Watson d-statistic( 10, 363) = 2.716652

Durbin-Watson d-statistic( 10, 363) = 2.981582

Durbin-Watson d-statistic( 10, 363) = 3.049315
```

Independent residuals – Bitstamp and Kraken (Reg 1 from top)

```
Durbin-Watson d-statistic( 10, 363) = 2.889774

Durbin-Watson d-statistic( 10, 363) = 3.053212

Durbin-Watson d-statistic( 10, 363) = 2.744229
```

Independent residuals – Kraken and Bitfinex (Reg 1 from top)

```
Durbin-Watson d-statistic( 10, 363) = 2.985505

Durbin-Watson d-statistic( 10, 363) = 3.07444

Durbin-Watson d-statistic( 10, 363) = 2.80264
```

Multicollinearity – Poloniex and Hitbtc

. vif

Variable	VIF	1/VIF
ENBV_Hitbtc	8.79	0.113713
EBV_Hitbtc	7.25	0.137909
UNBV_Polon~x	6.68	0.149672
ENBV_Polon~x	6.20	0.161411
UBV_Poloniex	5.94	0.168438
EBV_Poloniex	5.91	0.169167
UBV_Hitbtc	5.75	0.173833
UNBV_Hitbtc	5.69	0.175799
Binance	2.30	0.434546
Mean VIF	6.06	

Multicollinearity –Bitstamp and Bitfinex

. vif

Variable	VIF	1/VIF
UNBV_Bitfi~x	24.27	0.041201
ENBV_Bitfi~x	23.61	0.042351
UBV_Bitfinex	19.99	0.050026
ENBV_Bitst~p	15.94	0.062744
EBV_Bitfinex	15.63	0.063996
UNBV_Bitst~p	12.78	0.078234
UBV_Bitstamp	5.62	0.178033
EBV_Bitstamp	4.90	0.204189
Binance	3.21	0.311654
Mean VIF	13.99	

Multicollinearity – Kraken and Bitfinex

. vif

Variable	VIF	1/VIF
UNBV_Bitfi~x	22.32	0.044805
UBV_Bitfinex	21.41	0.046708
ENBV_Kraken	18.80	0.053188
EBV_Bitfinex	17.39	0.057510
UNBV_Kraken	17.03	0.058734
ENBV_Bitfi~x	16.90	0.059181
EBV_Kraken	11.18	0.089412
UBV_Kraken	9.52	0.105032
Binance	3.34	0.299143
Mean VIF	15.32	

Multicollinearity –Bitstamp and Kraken

. vif

Variable	VIF	1/VIF
ENBV_Kraken	14.72	0.067948
UNBV_Kraken	14.26	0.070121
EBV_Kraken	11.22	0.089093
UBV_Kraken	8.76	0.114123
UNBV_Bitst~p	8.25	0.121275
ENBV_Bitst~p	7.28	0.137283
EBV_Bitstamp	5.78	0.173156
UBV_Bitstamp	5.75	0.174022
Binance	3.24	0.309109
Mean VIF	8.81	

TABLE 9 – REGRESSION RESULTS: LEAD-LAG TIME AS DEPENDENT VARIABLE

	Poloniex - Hitbtc		Bitstamp - Bitfinex		Bitstamp - Kraken		Kraken - Bitfinex	
<i>Panel B: lead-lag time (seconds)</i>								
Intercept	0.570980	0.8034	0.601683	0.7061	-0.384145	0.8608	0.858756	0.7122
Excepted Poloniex block volume	0.003299	0.8587						
Unexpected Poloniex block volume	0.038759	0.1359						
Excepted Poloniex non-block volume	-0.000183	0.8348						
Unexpected Poloniex non-block volume	-0.002330	0.1056						
Excepted Hitbtc block volume	0.000269	0.9087						
Unexpected Hitbtc block volume	0.000352	0.8903						
Excepted Hitbtc non-block volume	-0.000037	0.9041						
Unexpected Hitbtc non-block volume	0.000097	0.8040						
Excepted Bitstamp block volume			0.000385	0.7667	-0.000542	0.7803		
Unexpected Bitstamp block volume			-0.001767	0.3113	0.001203	0.6222		
Excepted Bitstamp non-block volume			0.000059	0.8546	0.000054	0.8578		
Unexpected Bitstamp non-block volume			0.000338	0.3392	-0.000213	0.5901		
Excepted Bitfinex block volume			0.000471	0.6569			-0.000268	0.8742
Unexpected Bitfinex block volume			-0.000373	0.7792			0.002559	0.2206
Excepted Bitfinex non-block volume			-0.000042	0.7321			-0.000016	0.9192
Unexpected Bitfinex non-block volume			0.000053	0.7322			-0.000387*	0.0887
Excepted Kraken block volume					0.001131	0.8264	-0.000558	0.9217
Unexpected Kraken block volume					0.001291	0.8329	-0.000642	0.3896
Excepted Kraken non-block volume					-0.000171	0.8032	0.000255	0.7653
Unexpected Kraken non-block volume					-0.000222	0.8117	0.001263	0.2594
Total Binance volume	-0.000007	0.9272	-0.000042	0.4654	0.000021	0.7946	-0.000021	0.8090
R-squared	0.0118		0.0103		0.0029		0.0110	
P-value	0.8958		0.9315		0.9994		0.9146	

TABLE 10– REGRESSION RESULTS: OUTPUT FROM STATA

Poloniex and Hitbtc

Source	SS	df	MS	Number of obs =	F(9, 353)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	.031737418	9	.003526248	363	0.0002	0.0002	0.0002	0.0002	0.0002
Residual	.330243291	353	.000935533						
Total	.361980709	362	.000939947						

diff_Max_Po-o	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Poloniex	-.000017	.0000339	-0.50	0.615	-.0000836 .0000495
UBV_Poloniex	.0000152	.0000474	0.32	0.749	-.000078 .0001084
ENBV_Poloniex	6.71e-07	1.40e-06	0.48	0.636	-2.48e-06 9.89e-06
UNBV_Poloniex	-2.16e-06	2.63e-06	-0.82	0.411	-7.33e-06 3.00e-06
EBV_Hitbtc	-2.71e-06	4.28e-06	-0.63	0.528	-.0000111 5.72e-06
UBV_Hitbtc	.0000138	4.67e-06	2.96	0.003	4.44e-06 .000023
ENBV_Hitbtc	1.99e-07	5.63e-07	0.35	0.724	-9.09e-07 1.31e-06
UNBV_Hitbtc	-2.84e-06	7.11e-07	-4.00	0.000	-4.24e-06 -1.44e-06
Binance	1.53e-07	1.37e-07	1.11	0.266	-1.17e-07 4.23e-07
_cons	-.0058463	.0041907	-1.40	0.164	-.0140963 .0023936

Source	SS	df	MS	Number of obs =	F(9, 353)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	1178.67936	9	130.964374	363	0.8958	0.0118	0.0118	0.0118	0.0118
Residual	98791.7008	353	279.863175						
Total	99970.3802	362	276.161271						

diff_Time_P-o	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Poloniex	.0032993	.0185169	0.18	0.859	-.033118 .0397166
UBV_Poloniex	.0387594	.0259303	1.49	0.136	-.0123279 .0897567
ENBV_Poloniex	-.0001831	.0009776	-0.21	0.835	-.0019091 .0015429
UNBV_Poloniex	-.0023303	.0014362	-1.62	0.106	-.0051548 .0004943
EBV_Hitbtc	.0002689	.0023423	0.11	0.909	-.0043377 .0048755
UBV_Hitbtc	.0003523	.0025523	0.14	0.890	-.0046673 .0053719
ENBV_Hitbtc	-.0000371	.0003078	-0.12	0.904	-.0006424 .0005681
UNBV_Hitbtc	.0000965	.0003887	0.25	0.804	-.0006678 .0008609
Binance	-6.86e-06	.000075	-0.09	0.927	-.0001544 .0001407
_cons	.5709798	2.29204	0.25	0.803	-3.936918 5.078978

Source	SS	df	MS	Number of obs =	F(9, 353)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	1.1038001	9	.122644678	363	0.0774	0.0774	0.0774	0.0774	0.0774
Residual	24.7923496	353	.070233285						
Total	25.8961517	362	.071536331						

diff_LLR_P-o	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Poloniex	-.0001975	.0002933	-0.67	0.501	-.0007745 .0003794
UBV_Poloniex	.0004216	.0004109	1.03	0.305	-.0003863 .0012294
ENBV_Poloniex	7.94e-06	.0000139	0.57	0.569	-.0000194 .0000353
UNBV_Poloniex	-9.44e-06	.0000228	-0.42	0.678	-.0000542 .0000353
EBV_Hitbtc	-.000024	.0000371	-0.65	0.517	-.0000977 .0000489
UBV_Hitbtc	.0000092	.0000404	0.23	0.823	-.0000125 .0001116
ENBV_Hitbtc	3.49e-06	4.89e-06	0.72	0.475	-6.10e-06 .0000131
UNBV_Hitbtc	-8.81e-06	6.16e-06	-1.43	0.153	-.0000209 3.30e-06
Binance	-6.10e-07	1.19e-06	-0.51	0.608	-2.95e-06 1.73e-06
_cons	-.0030286	.0363106	-0.08	0.934	-.0744409 .0683836

Bitstamp and Bitfinex

Source	SS	df	MS	Number of obs =	F(9, 353)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	435.414074	9	48.3793416	363	0.9315	0.0103	0.0103	0.0103	0.0103
Residual	42032.046	353	119.070952						
Total	42467.4601	362	117.313426						

diff_Max_Bi-x	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Bitstamp	-1.97e-06	3.76e-06	-0.52	0.601	-9.36e-06 5.42e-06
UBV_Bitstamp	1.90e-06	5.69e-06	0.33	0.738	-9.27e-06 .0000131
ENBV_Bitstamp	4.84e-07	4.72e-07	1.03	0.306	-4.44e-07 1.41e-06
UNBV_Bitstamp	1.74e-06	6.88e-07	2.53	0.012	3.84e-07 3.09e-06
EBV_Bitfinex	9.68e-07	1.69e-06	0.57	0.567	-2.35e-06 4.29e-06
UBV_Bitfinex	5.62e-07	2.22e-06	0.25	0.803	-3.87e-06 4.99e-06
ENBV_Bitfinex	-1.58e-07	1.74e-07	-0.91	0.365	-5.03e-07 1.85e-07
UNBV_Bitfinex	-1.02e-06	2.87e-07	-3.54	0.000	-1.58e-06 -4.52e-07
Binance	6.13e-08	8.86e-08	0.69	0.489	-1.13e-07 2.35e-07
_cons	-.0058953	.0023807	-1.13	0.258	-.0073775 .0019869

diff_Time_B-x	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Bitstamp	.0003852	.0012975	0.30	0.767	-.0021665 .0029337
UBV_Bitstamp	-.0017671	.001743	-1.01	0.311	-.005195 .0016607
ENBV_Bitstamp	-.0000592	.0000323	-0.18	0.855	-.0000976 .0000645
UNBV_Bitstamp	.0003379	.0003531	0.96	0.339	-.0003565 .0010223
EBV_Bitfinex	.0004705	.0010585	0.44	0.657	-.0016112 .0025523
UBV_Bitfinex	-.0003728	.0013288	-0.28	0.779	-.0029962 .0022405
ENBV_Bitfinex	-.0000422	.0001239	-0.34	0.732	-.0001945 .0001099
UNBV_Bitfinex	.0000529	.0001545	0.34	0.732	-.0002509 .0003567
Binance	-.000042	.0000574	-0.73	0.465	-.0001549 .000071
_cons	.6016834	1.59405	0.38	0.706	-2.53347 3.736713

Source	SS	df	MS	Number of obs =	F(9, 353)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	1.33880275	9	.148755061	363	0.4209	0.0109	0.0109	0.0109	0.0109
Residual	66.0329053	353	.187062055						
Total	67.371708	362	.186103691						

diff_LLR_Bi-x	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Bitstamp	.0000448	.0000514	0.87	0.385	-.0000564 .0001459
UBV_Bitstamp	.0000136	.0000691	0.20	0.844	-.0001222 .0001495
ENBV_Bitstamp	-.0000136	.0000129	-1.04	0.299	-.0000388 .0000116
UNBV_Bitstamp	3.35e-06	.000014	0.24	0.811	-.0000242 .0000309
EBV_Bitfinex	-.0000328	.000042	-0.78	0.436	-.0001153 .0000498
UBV_Bitfinex	-.0000476	.0000527	-0.90	0.367	-.0001512 .0000566
ENBV_Bitfinex	5.69e-06	4.89e-06	1.16	0.245	-3.92e-06 .0000153
UNBV_Bitfinex	1.14e-06	6.12e-06	0.19	0.853	-.0000109 .0000132
Binance	-5.52e-07	2.29e-06	-0.24	0.808	-5.03e-06 3.92e-06
_cons	.0233361	.0631018	0.37	0.712	-.1092939 .1475962

Bitstamp and Kraken

Source	SS	df	MS	Number of obs	=	363
Model	.076197473	9	.008466386	F(9, 353)	=	7.85
Residual	.380732824	353	.001078563	Prob > F	=	0.0000
				R-squared	=	0.1668
				Adj R-squared	=	0.1455
Total	.456930297	362	.001262238	Root MSE	=	.03284

diff_Max_Bi-n	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Bitstamp	2.04e-06	4.34e-06	0.48	0.631	-6.29e-06 .0000104
USV_Bitstamp	-4.38e-07	5.32e-06	-0.08	0.934	-.0000109 .00001
ENBV_Bitstamp	4.80e-07	6.59e-07	0.73	0.467	-8.15e-07 1.77e-06
UNBV_Bitstamp	1.24e-06	8.63e-07	1.44	0.150	-4.52e-07 2.94e-06
EBV_Kraken	1.05e-07	.0000112	0.01	0.993	-.0000022 .0000022
USV_Kraken	.0000223	.0000133	1.68	0.095	-3.88e-06 .0000486
ENBV_Kraken	-6.19e-07	1.50e-06	-0.41	0.680	-3.56e-06 2.33e-06
UNBV_Kraken	-8.29e-06	2.03e-06	-4.09	0.000	-.0000123 -4.30e-06
Binance	8.91e-08	1.73e-07	0.51	0.607	-2.50e-07 4.30e-07
_cons	-.0060867	.0047748	-1.27	0.203	-.0154774 .0033039

Source	SS	df	MS	Number of obs	=	363
Model	232.462774	9	25.8291971	F(9, 353)	=	0.11
Residual	80015.5262	353	226.672879	Prob > F	=	0.9994
				R-squared	=	0.0029
				Adj R-squared	=	-0.0225
Total	80247.989	362	221.679528	Root MSE	=	15.056

diff_Time_B-n	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Bitstamp	-.0005421	.0019417	-0.28	0.780	-.004361 .0032767
USV_Bitstamp	.0010207	.0024389	0.49	0.622	-.0035938 .0055992
ENBV_Bitstamp	.0000541	.0003019	0.18	0.858	-.0005396 .0006479
UNBV_Bitstamp	.0000213	.0003956	-0.54	0.590	-.0009912 .0005446
EBV_Kraken	.0011314	.005154	0.22	0.826	-.0080001 .012678
USV_Kraken	.001291	.0061138	0.21	0.833	-.0107331 .0133151
ENBV_Kraken	-.0001712	.0006863	-0.25	0.803	-.001521 .0011786
UNBV_Kraken	-.00002217	.00093	-0.24	0.812	-.0020007 .0014072
Binance	.0000207	.0000795	0.26	0.795	-.0001356 .000177
_cons	-.3841455	2.189936	-0.18	0.861	-4.68914 3.920849

Source	SS	df	MS	Number of obs	=	363
Model	10.6110835	9	1.17900928	F(9, 353)	=	1.05
Residual	394.053838	353	1.12194555	Prob > F	=	0.3990
				R-squared	=	0.0261
				Adj R-squared	=	0.0013
Total	406.664922	362	1.12338376	Root MSE	=	1.0592

diff_LLR_Bi-n	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Bitstamp	-3.56e-06	.0001366	-0.03	0.979	-.0002722 .0002651
USV_Bitstamp	.0000885	.0001716	0.52	0.606	-.0000449 .0004259
ENBV_Bitstamp	.0000147	.0000212	0.69	0.489	-.0000271 .0000565
UNBV_Bitstamp	.0000346	.0000278	1.24	0.214	-.0000201 .0000894
EBV_Kraken	.0000513	.0003626	0.14	0.887	-.0006418 .0007445
USV_Kraken	.0001476	.0004301	0.34	0.732	-.0006983 .0009936
ENBV_Kraken	-.0000265	.0000483	-0.55	0.583	-.0001215 .0000684
UNBV_Kraken	-.0001138	.0000854	-1.74	0.083	-.0002420 .0000149
Binance	3.98e-06	5.59e-06	0.71	0.477	-7.01e-06 .000015
_cons	-.1598592	.1540008	-1.04	0.300	-.4627337 .1430152

Kraken and Bitfinex

Source	SS	df	MS	Number of obs	=	363
Model	.04439541	9	.004932823	F(9, 353)	=	6.71
Residual	.253368106	353	.000724754	Prob > F	=	0.0000
				R-squared	=	0.1462
				Adj R-squared	=	0.1244
Total	.303763516	362	.000839126	Root MSE	=	.02711

diff_Max_Kr-x	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Kraken	7.63e-07	9.30e-06	0.08	0.935	-.0000175 .000019
USV_Kraken	3.39e-06	.0000115	0.29	0.768	-.0000192 .000026
ENBV_Kraken	-7.08e-07	1.40e-06	-0.51	0.614	-3.44e-06 2.05e-06
UNBV_Kraken	-1.20e-06	1.83e-06	-0.66	0.512	-4.81e-06 2.40e-06
EBV_Bitfinex	7.62e-07	2.77e-06	0.27	0.784	-4.70e-06 6.22e-06
USV_Bitfinex	6.44e-07	3.42e-06	0.19	0.851	-6.08e-06 7.37e-06
ENBV_Bitfinex	4.57e-08	2.60e-07	0.18	0.861	-4.64e-07 5.57e-07
UNBV_Bitfinex	-5.39e-07	3.71e-07	-1.45	0.148	-1.57e-06 1.91e-07
Binance	6.59e-08	1.45e-07	0.45	0.651	-2.20e-07 3.52e-07
_cons	-.002354	.0038135	-0.62	0.537	-.0098541 .005146

Source	SS	df	MS	Number of obs	=	363
Model	1075.50003	9	119.500003	F(9, 353)	=	0.44
Residual	94503.365	353	273.380637	Prob > F	=	0.9146
				R-squared	=	0.0110
				Adj R-squared	=	-0.0142
Total	97578.865	362	269.554876	Root MSE	=	16.334

diff_Time_Kr-x	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Kraken	-.0005575	.0056699	-0.10	0.922	-.0117085 .0105935
USV_Kraken	-.0060423	.0070149	-0.86	0.390	-.0198385 .0077538
ENBV_Kraken	.0002952	.0008941	0.30	0.765	-.0014246 .0019349
UNBV_Kraken	.0011633	.0011182	1.13	0.259	-.000936 .0034625
EBV_Bitfinex	-.0002683	.0016926	-0.16	0.874	-.0035972 .0030606
USV_Bitfinex	.0025593	.0020858	1.23	0.221	-.0015429 .0066615
ENBV_Bitfinex	-.0000161	.0001587	-0.10	0.919	-.0003282 .0002959
UNBV_Bitfinex	-.0003967	.0002460	-1.71	0.089	-.0008322 .0000398
Binance	-.0000215	.0000887	-0.24	0.809	-.000196 .0001531
_cons	.6587562	2.326154	0.37	0.712	-3.716107 5.433619

Source	SS	df	MS	Number of obs	=	363
Model	.945177693	9	.105019744	F(9, 353)	=	1.28
Residual	29.0672846	353	.082343583	Prob > F	=	0.2488
				R-squared	=	0.0315
				Adj R-squared	=	0.0068
Total	30.0124623	362	.082907355	Root MSE	=	.28696

diff_LLR_Kr-x	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
EBV_Kraken	-.0000339	.0000984	-0.33	0.738	-.0002265 .0001606
USV_Kraken	-.0000441	.0001217	-0.36	0.717	-.0002835 .0001954
ENBV_Kraken	.0000013	.0000148	0.88	0.381	-.0000162 .0000421
UNBV_Kraken	3.50e-06	.0000194	0.18	0.857	-.0000147 .0000417
EBV_Bitfinex	6.15e-07	.0000294	0.02	0.983	-.0000572 .0000584
USV_Bitfinex	.0000647	.0000362	1.79	0.075	-6.47e-06 .0001359
ENBV_Bitfinex	-1.59e-06	2.75e-06	-0.58	0.563	-7.01e-06 3.82e-06
UNBV_Bitfinex	-7.72e-06	3.93e-06	-1.96	0.050	-.0000154 1.68e-08
Binance	-1.64e-06	1.54e-06	-1.08	0.283	-4.68e-06 1.37e-06
_cons	.0500091	.040371	1.24	0.216	-.0293889 .129407


TABLE 11– THE CER CYBER SECURITY SCORE: CRYPTOCURRENCY EXCHANGES

#	Exchange	CMC rank	Server Security	User Security	Crowdsourced Security	Historical	CSS
1	Kraken	☆26	8.82	7.81	10.00	10.00	9.06
2	Coinbase Pro	☆29	8.18	7.81	10.00	10.00	8.74
3	Binance	☆3	7.70	7.81	10.00	10.00	8.50
4	BitMEX	☆1	8.10	6.56	10.00	10.00	8.50
5	itBit	☆71	7.62	7.50	10.00	10.00	8.41
6	Bittrex	☆50	8.56	5.31	10.00	8.50	8.23
7	BitBay	☆86	8.46	7.50	5.00	10.00	8.13
8	Cryptonex	☆27	8.42	7.50	5.00	10.00	8.11
9	HitBTC	☆15	6.86	7.81	10.00	10.00	8.08
10	Bitlish	☆63	9.08	8.75	0.00	10.00	7.94
11	KuCoin	☆74	8.60	10.00	0.00	10.00	7.90
12	B2BX	☆57	8.08	6.88	5.00	10.00	7.84
13	OOBTC	☆44	7.06	8.75	5.00	10.00	7.63
14	Upbit	☆17	8.02	10.00	0.00	10.00	7.61
15	bitFlyer	☆52	9.28	5.31	0.00	10.00	7.49
16	Exrates	☆23	8.16	8.75	0.00	10.00	7.48
17	Livecoin	☆82	7.72	10.00	0.00	10.00	7.46
18	Bitfinex	☆13	9.10	10.00	0.00	6.50	7.45
19	GBX	☆78	7.64	10.00	0.00	10.00	7.42
20	YoBit	☆73	7.56	10.00	0.00	10.00	7.38
21	BitMax	☆42	8.18	7.50	0.00	10.00	7.29
22	Coinone	☆64	8.02	7.81	0.00	10.00	7.26
23	Bilaxy	☆72	8.38	6.56	0.00	10.00	7.24
24	P2PB2B	☆46	8.06	7.50	0.00	10.00	7.23
25	DigiFinex	☆6	7.34	5.31	5.00	10.00	7.22
26	Coindeal	☆93	6.98	10.00	0.00	10.00	7.09
27	lBank	☆10	7.32	8.75	0.00	10.00	7.06
28	OEX	☆21	7.24	8.75	0.00	10.00	7.02
29	Mercatox	☆100	7.62	7.50	0.00	10.00	7.01
30	Vebitcoin	☆96	6.82	10.00	0.00	10.00	7.01
31	Coinhub	☆65	6.78	10.00	0.00	10.00	6.99
32	CEX.IO	☆90	7.46	7.50	0.00	10.00	6.93
33	Coinbe	☆83	6.66	10.00	0.00	10.00	6.93
34	Coinsuper	☆28	6.66	10.00	0.00	10.00	6.93
35	CoinsBank	☆40	8.42	4.38	0.00	10.00	6.91
36	ABCC	☆58	6.56	10.00	0.00	10.00	6.88
37	BiteBTC	☆51	7.36	7.50	0.00	10.00	6.88
38	Hotbit	☆67	7.30	7.50	0.00	10.00	6.85
39	BITBOX	☆81	7.58	6.56	0.00	10.00	6.84
40	Bgogo	☆69	8.22	4.38	0.00	10.00	6.81
41	Gate.io	☆48	8.22	8.75	0.00	6.50	6.81
42	UEX	☆59	7.18	7.50	0.00	10.00	6.79
43	Trade By Trade	☆80	6.34	10.00	0.00	10.00	6.77
44	BitMart	☆20	7.36	6.56	0.00	10.00	6.73
45	Coinall	☆79	8.06	4.38	0.00	10.00	6.73
46	BCEX	☆30	7.64	5.63	0.00	10.00	6.72
47	BitForex	☆16	6.04	10.00	0.00	10.00	6.62
48	IDAX	☆14	6.22	8.75	0.00	10.00	6.51
49	OKEEx	☆4	7.56	4.38	5.00	6.50	6.48
50	CHAOEX	☆45	5.74	10.00	0.00	10.00	6.47

51	Korbit	☆75	5.74	10.00	0.00	10.00	6.47
52	CoinMex	☆84	7.12	5.63	0.00	10.00	6.46
53	GDAC	☆66	6.36	7.81	0.00	10.00	6.43
54	Huobi	☆5	6.02	8.75	0.00	10.00	6.41
55	Exmo	☆62	7.38	7.50	0.00	7.50	6.39
56	GOPAX	☆95	6.58	6.88	0.00	10.00	6.39
57	Poloniex	☆56	7.72	7.50	0.00	6.50	6.36
58	Liquid	☆31	6.44	6.88	0.00	10.00	6.32
59	CoinEx	☆98	6.22	7.50	0.00	10.00	6.31
60	DragonEX	☆33	5.40	10.00	0.00	10.00	6.30
61	BTCBOX	☆76	6.14	7.50	0.00	10.00	6.27
62	Coineal	☆35	5.64	8.75	0.00	10.00	6.22
63	LocalTrade	☆87	7.04	4.38	0.00	10.00	6.22
64	Gemini	☆60	6.58	5.63	0.00	10.00	6.19
65	CoinEgg	☆53	5.84	7.81	0.00	10.00	6.17
66	Kryptono	☆68	5.92	7.50	0.00	10.00	6.16
67	Bitbank	☆34	6.60	5.31	0.00	10.00	6.15
68	IDCM	☆25	5.38	8.75	0.00	10.00	6.09
69	Simex	☆39	5.76	7.50	0.00	10.00	6.08
70	Bitinka	☆55	5.64	7.81	0.00	10.00	6.07
71	CoinBene	☆9	5.34	8.75	0.00	10.00	6.07
72	Fatbtc	☆36	6.34	5.63	0.00	10.00	6.07
73	MBAex	☆85	5.74	7.50	0.00	10.00	6.07
74	Bit-Z	☆11	6.32	5.63	0.00	10.00	6.06
75	Sistemkoin	☆47	6.72	4.38	0.00	10.00	6.06
76	Bitrue	☆77	6.24	5.63	0.00	10.00	6.02
77	Bitsane	☆94	5.42	7.81	0.00	10.00	5.96
78	LATOKEN	☆43	6.82	7.81	0.00	6.50	5.96
79	FCoin	☆24	4.64	10.00	0.00	10.00	5.92
80	Bibox	☆18	4.56	10.00	0.00	10.00	5.88
81	Bitstamp	☆41	6.54	7.81	0.00	6.50	5.82
82	LakeBTC	☆92	5.24	7.50	0.00	10.00	5.82
83	CoinTiger	☆38	4.82	8.75	0.00	10.00	5.81
84	EXX	☆19	5.34	6.56	0.00	10.00	5.72
85	RightBTC	☆32	4.64	8.75	0.00	10.00	5.72
86	ZB.COM	☆8	5.24	6.88	0.00	10.00	5.72
87	Instant Bitex	☆97	4.88	7.50	0.00	10.00	5.64
88	Cashierest	☆54	3.98	10.00	0.00	10.00	5.59
89	TOPBTC	☆37	4.74	7.50	0.00	10.00	5.57
90	Allcoin	☆49	5.34	5.31	0.00	10.00	5.52
91	CoinZest	☆61	3.84	10.00	0.00	10.00	5.52
92	Neraex	☆89	5.60	4.38	0.00	10.00	5.50
93	DOBI	☆22	4.48	7.81	0.00	10.00	5.49
94	Coinsquare	☆88	5.26	5.31	0.00	10.00	5.48
95	InfinityCoinExchange	☆91	5.54	4.38	0.00	10.00	5.47
96	ZBG	☆12	4.94	5.31	0.00	10.00	5.32
97	Coinbit	☆7	3.80	7.81	0.00	10.00	5.15
98	Bitthumb	☆2	6.24	7.81	0.00	1.50	4.67
99	Coincheck	☆70	5.96	5.63	0.00	3.00	4.48
100	Zaif	☆99	5.72	5.31	0.00	3.00	4.31

TABLE 12— TRADING FEES

Binance

Level	30d Trade Volume(BTC)	&	BNB holdings	Maker	Taker	 Maker	Taker
General	< 100 BTC	or	≥ 0 BNB	0.1000%	0.1000%	0.0750%	0.0750%
VIP 1	≥ 100 BTC	&	≥ 50 BNB	0.0900%	0.1000%	0.0675%	0.0750%
VIP 2	≥ 500 BTC	&	≥ 200 BNB	0.0800%	0.1000%	0.0600%	0.0750%
VIP 3	≥ 4500 BTC	&	≥ 1000 BNB	0.0700%	0.0900%	0.0525%	0.0675%
VIP 4	≥ 10000 BTC	&	≥ 2000 BNB	0.0600%	0.0800%	0.0450%	0.0600%
VIP 5	≥ 20000 BTC	&	≥ 3500 BNB	0.0500%	0.0700%	0.0375%	0.0525%
VIP 6	≥ 40000 BTC	&	≥ 6000 BNB	0.0400%	0.0600%	0.0300%	0.0450%
VIP 7	≥ 80000 BTC	&	≥ 9000 BNB	0.0300%	0.0500%	0.0225%	0.0375%
VIP 8	≥ 150000 BTC	&	≥ 11000 BNB	0.0200%	0.0400%	0.0150%	0.0300%

Coinbase

Pricing Tier	Taker Fee	Maker Fee
<\$100K	0.25%	0.15%
100K - 1M	0.20%	0.10%
1- 10M	0.18%	0.08%
10 -50M	0.15%	0.05%
50 - 100M	0.10%	0.00%
100 - 300M	0.08%	0.00%
300 - 500M	0.07%	0.00%
500M - 1B	0.06%	0.00%
\$1B+	0.05%	0.00%

Kraken

30- Day Volume	Maker	Taker
\$0 - \$50,000	0.16%	0.26%
\$50,001 - \$100,000	0.14%	0.24%
\$100,001 - \$250,000	0.12%	0.22%
\$250,001 - \$500,000	0.10%	0.20%
\$500,001 - \$1,000,000	0.08%	0.18%
\$1,000,001 - \$2,500,000	0.06%	0.16%
\$2,500,001 - \$5,000,000	0.04%	0.14%
\$5,000,001 - \$10,000,000	0.02%	0.12%
\$10,000,000+	0.00%	0.10%

Bitfinex

Order Execution

EXECUTED IN THE LAST 30 DAYS (USD EQUIVALENT)

MAKER FEES

TAKER FEES

\$0.00 or more traded	0.100%	0.200%
\$500,000.00 or more traded	0.080%	0.200%
\$1,000,000.00 or more traded	0.060%	0.200%
\$2,500,000.00 or more traded	0.040%	0.200%
\$5,000,000.00 or more traded	0.020%	0.200%
\$7,500,000.00 or more traded	0.000%	0.200%
\$10,000,000.00 or more traded	0.000%	0.180%
\$15,000,000.00 or more traded	0.000%	0.160%
\$20,000,000.00 or more traded	0.000%	0.140%
\$25,000,000.00 or more traded	0.000%	0.120%
\$30,000,000.00 or more traded	0.000%	0.100%
\$300,000,000.00 or more traded	0.000%	0.090%
\$1,000,000,000.00 or more traded	0.000%	0.085%
\$3,000,000,000.00 or more traded	0.000%	0.075%
\$10,000,000,000.00 or more traded	0.000%	0.060%
\$30,000,000,000.00 or more traded	0.000%	0.055%

Bitstamp

ALL TRADING PAIRS (CUMULATIVE)

Fee %	30 days USD volume
0.25%	< \$20,000
0.24%	< \$100,000
0.22%	< \$200,000
0.20%	< \$400,000
0.15%	< \$600,000
0.14%	< \$1,000,000
0.13%	< \$2,000,000
0.12%	< \$4,000,000
0.11%	< \$20,000,000
0.10%	> \$20,000,000

Hitbtc

Trading fees based on volume			
Tier	30-days Trading Volume (BTC)	Maker Fee	Taker Fee
0	≥ 0 BTC	0.1%	0.2%
1	≥ 100 BTC	0.08%	0.2%
2	≥ 200 BTC	0.06%	0.2%
3	≥ 500 BTC	0.04%	0.2%
4	≥ 1000 BTC	0.02%	0.2%
5	≥ 1500 BTC	0%	0.2%
6	≥ 2000 BTC	0%	0.18%
7	≥ 3000 BTC	0%	0.16%
8	≥ 4000 BTC	0%	0.14%
9	≥ 5000 BTC	0%	0.12%
10	≥ 6000 BTC	-0.01%	0.1%
11	≥ 60000 BTC	-0.01%	0.09%
12	≥ 200000 BTC	-0.01%	0.085%
13	≥ 600000 BTC	-0.01%	0.075%
14	≥ 2000000 BTC	-0.01%	0.06%
15	≥ 6000000 BTC	-0.01%	0.055%

Poloniex

Maker	Taker	Trade Volume (trailing 30 day avg)
0.08%	0.20%	< \$1m USD
0.02%	0.15%	< \$20m USD
0.00%	0.10%	≥ \$20m USD

10.2 FIGURES

FIGURE 1 – WEEKDAY VOLUME

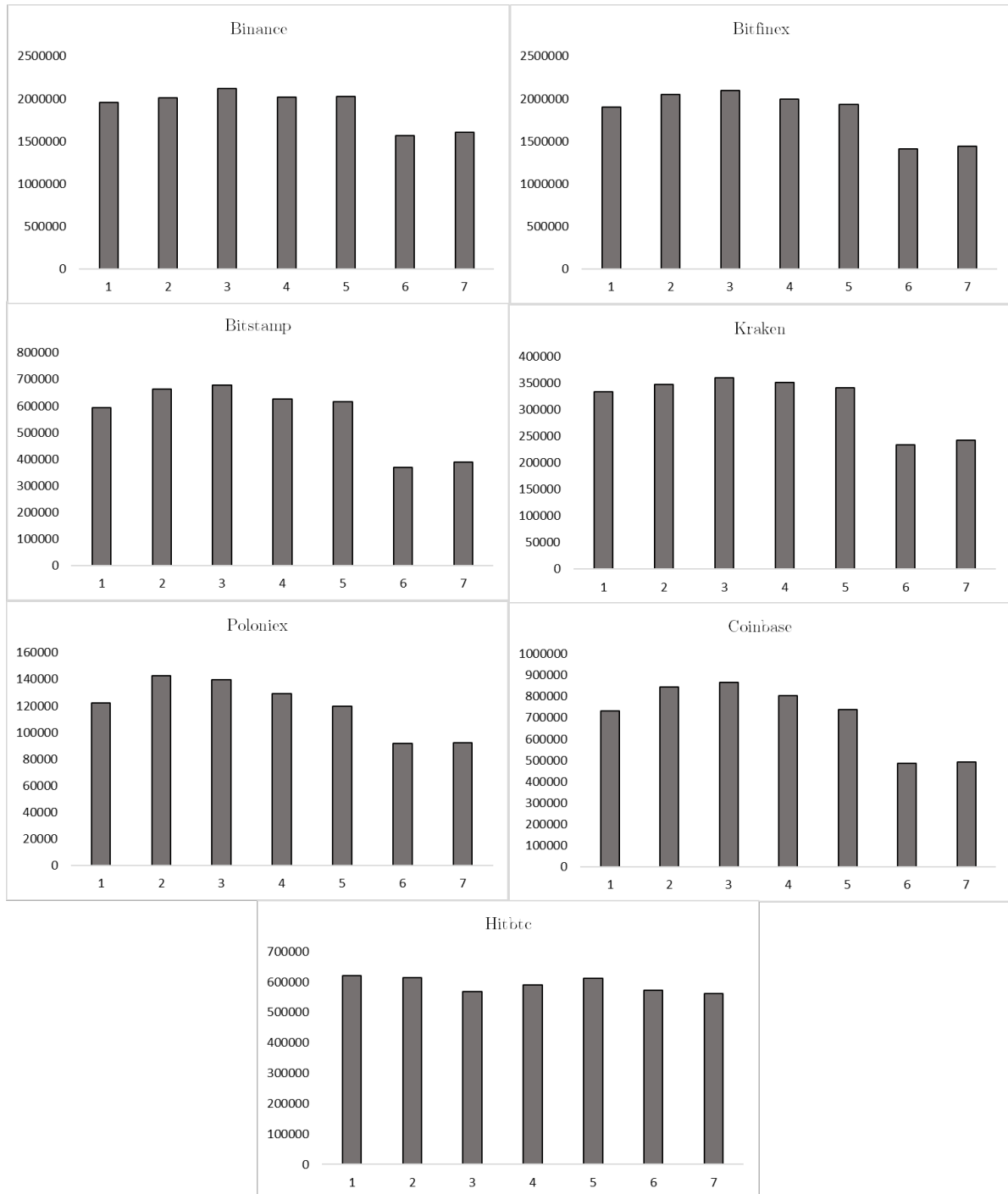


FIGURE 2 – TRADE SIZE DISTRIBUTIONS

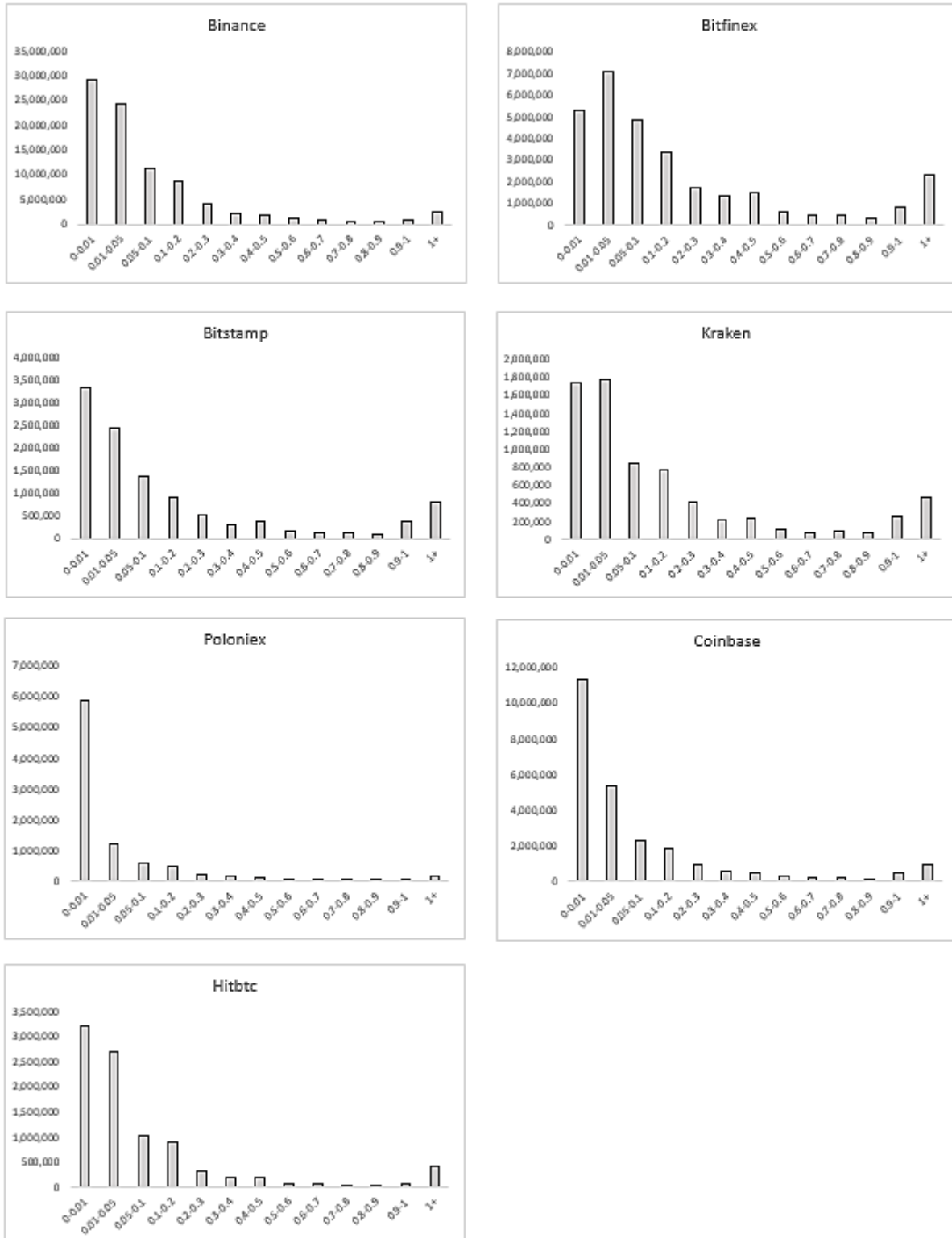
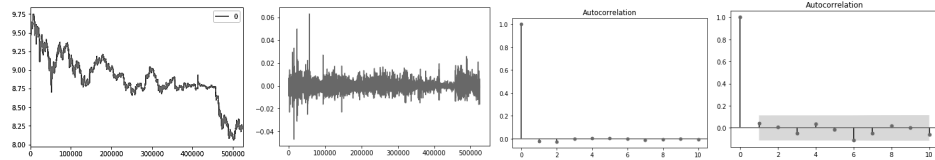


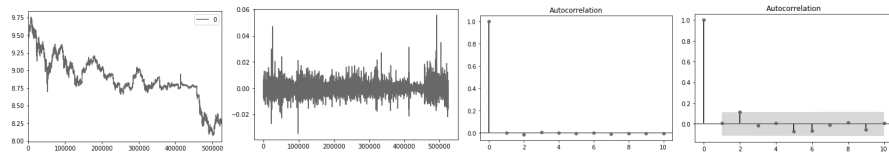
FIGURE 3 – STATIONARITY PLOTS

All outputs are collected from Stata. The plots below show prices in log-levels, differenced levels, the output of ACF plot for all 525,600 observations and a sample of 300 observations. The sample is merely included to show how the ACF plot should look, but due to the large number of observations the confidence bands are extremely tight and not visible on the original ACF plot.

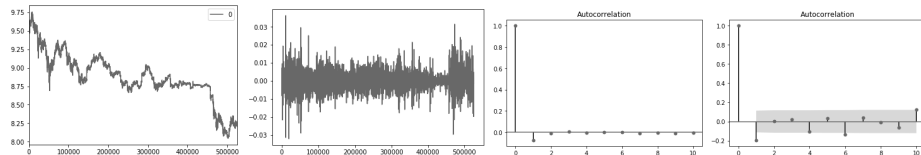
Binance



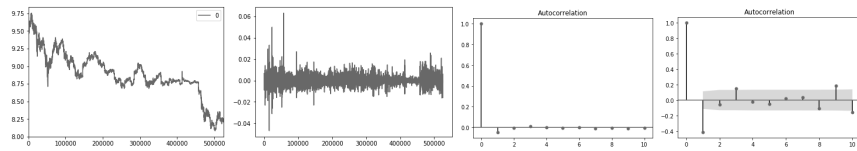
Bitfinex



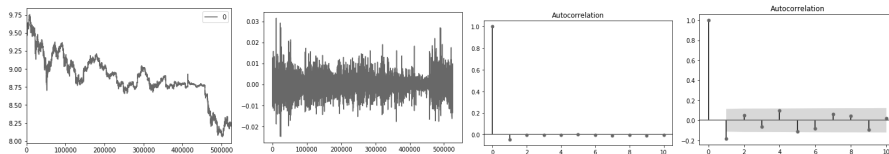
Bitstamp



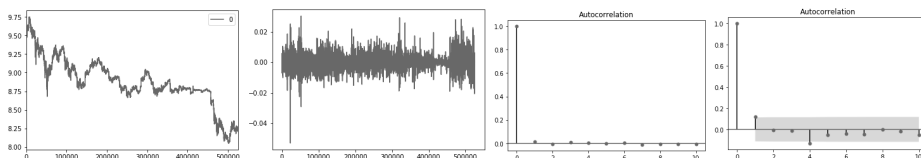
Kraken



Poloniex



Coinbase



Hitbtc

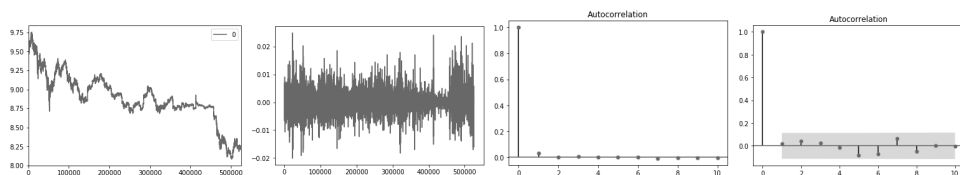
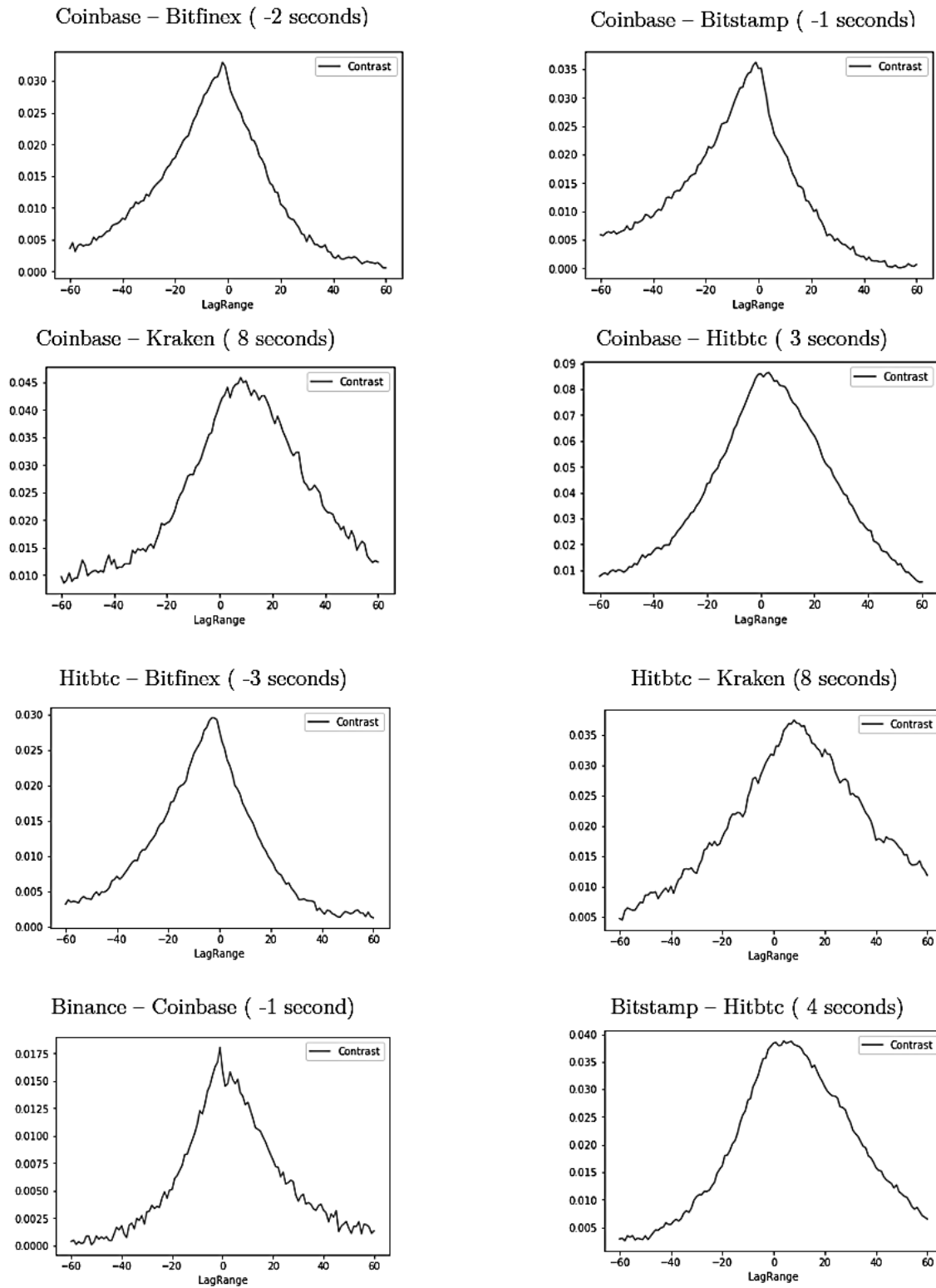
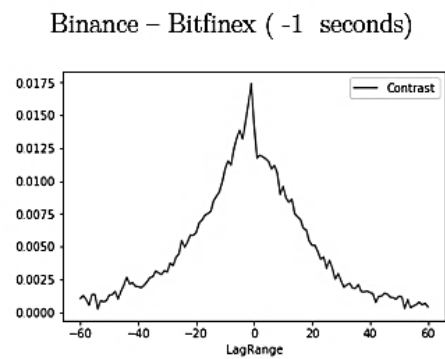
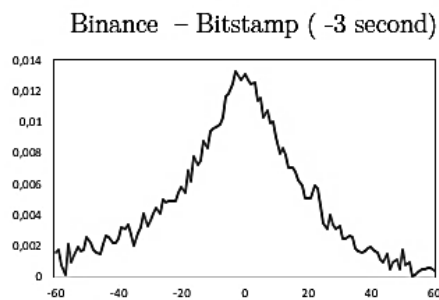
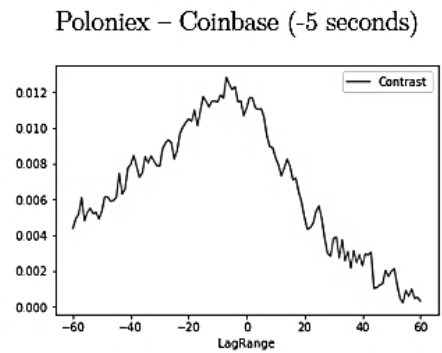
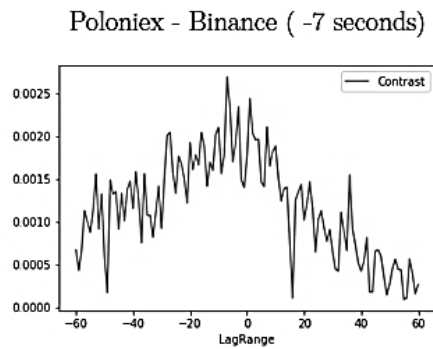
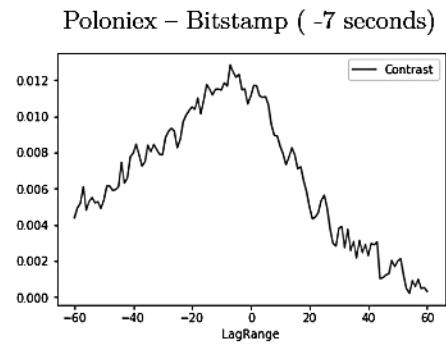
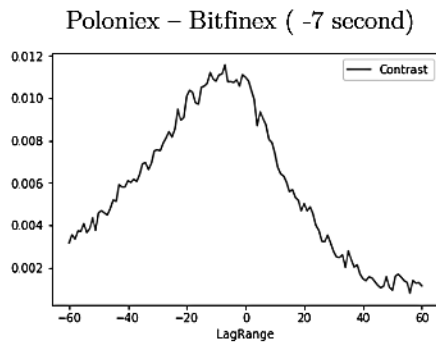
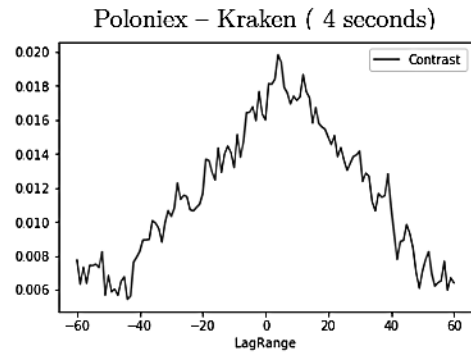
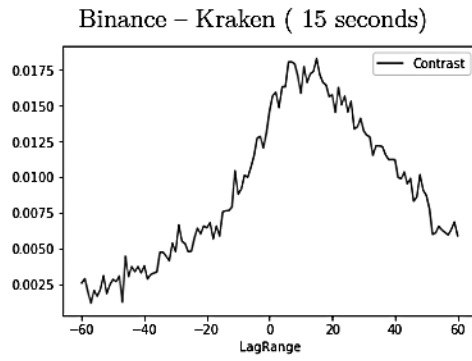


FIGURE 4 – CROSS CORRELATION FUNCTIONS





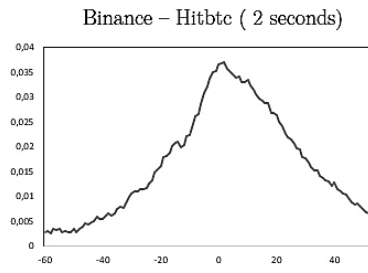
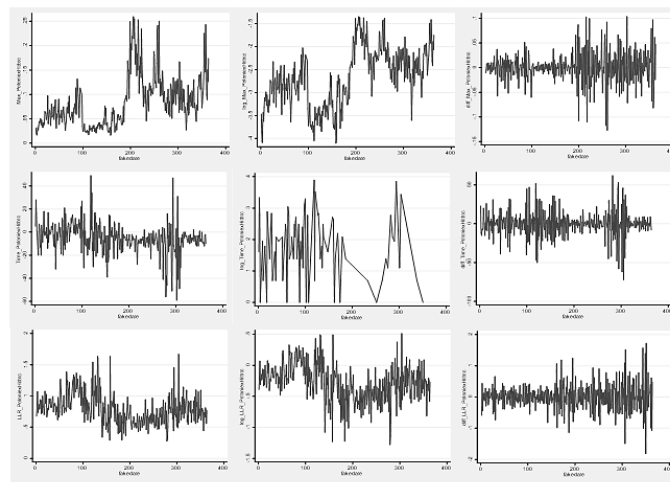
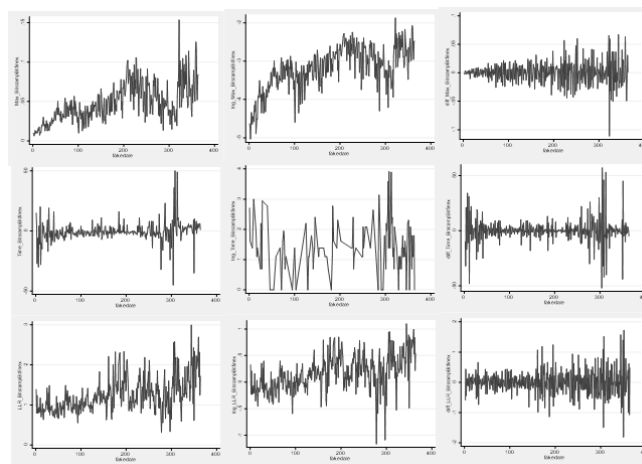


FIGURE 5 – PLOTS OF DEPENDENT VARIABLES

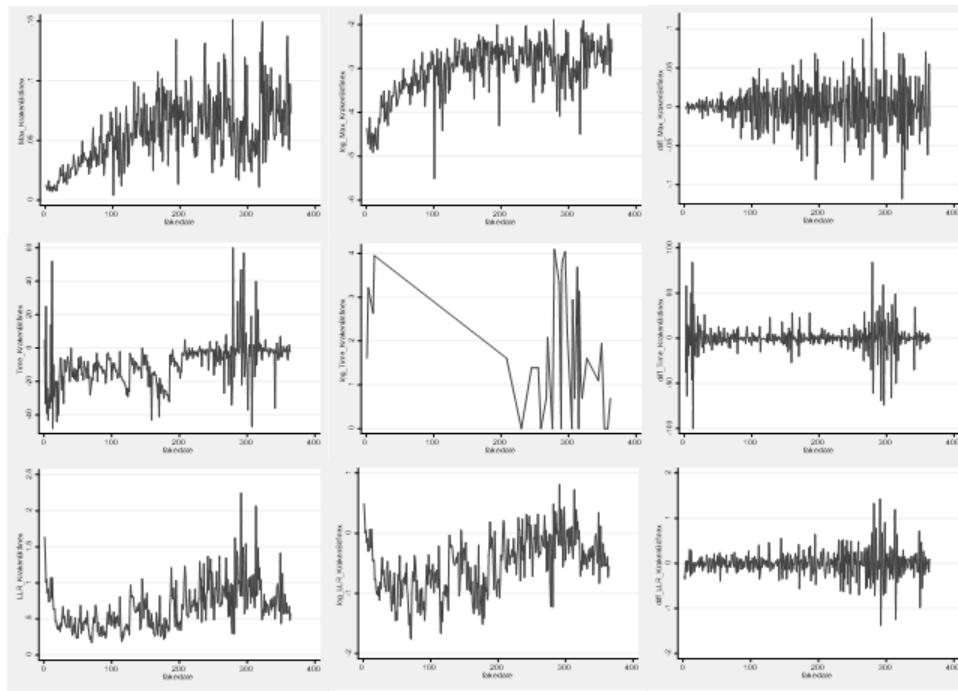
Poloniex and Hitbtc



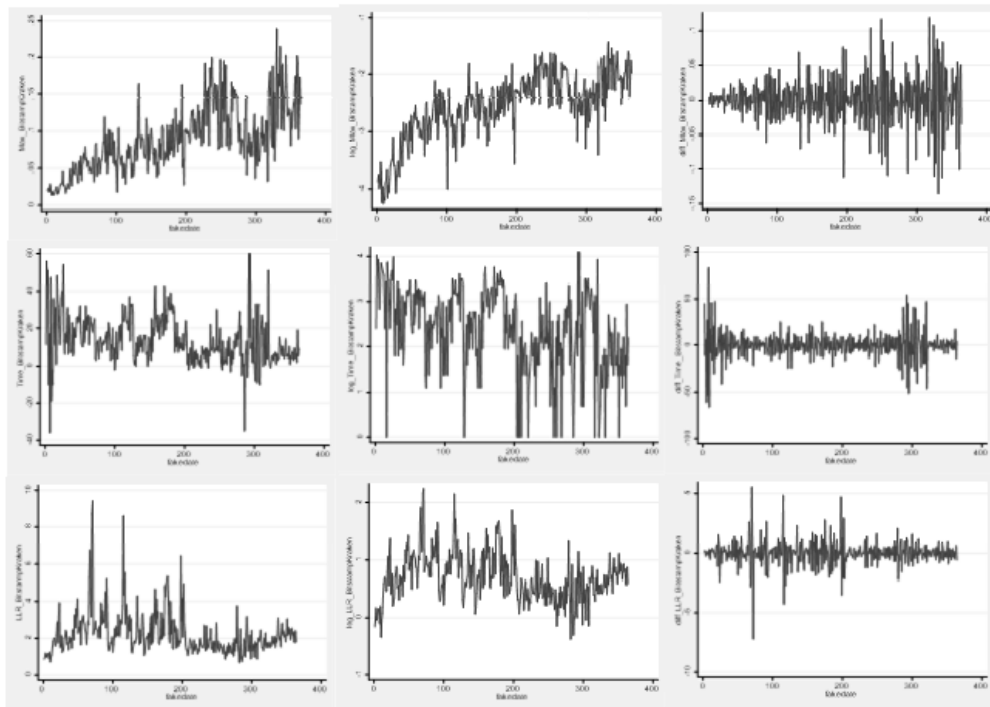
Bitstamp and Bitfinex



Kraken and Bitfinex



Bitstamp and Kraken



10.3 PYTHON CODE

Additional details available upon request.

STATIONARITY

```
from matplotlib import pyplot
from numpy import log
%matplotlib inline

#check plot
binance=df['Pbinance']
binance.plot()
pyplot.show()
binance.hist()

#if not-stationary, calculate log-values and plot
binanceX=binance.values
binancelog=pd.DataFrame(log(binanceX))
binancelog.plot()

pyplot.hist(binancelog)
pyplot.show()

#if non-stationary, find the order of integration I(x)
binancediff=binancelog.diff().dropna()
pyplot.plot(binancediff)

df_old=None
#Check ACF plot
from statsmodels.graphics.tsaplots import plot_acf
plot_acf(new,lags=10)
pyplot.show()

#Augmented dickey fuller
from statsmodels.tsa.stattools import adfuller

result=adfuller(binancediff[0], regression='nc', autolag='AIC')
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
print('Critical Values:')
for key, value in result[4].items():
    print('\t%s: %.3f' % (key, value))
```

VECTOR AUTOREGRESSIVE MODEL

```
import numpy as np
import pandas
import statsmodels.api as sm
from statsmodels.tsa.api import VAR, DynamicVAR

varsample=johansensample

varsample=np.log(varsample).diff().dropna()

varmodel= VAR(varsample)

varmodel.select_order(15)

#results with maxlag 15 and BIC criterion selects number of lags
varresults = varmodel.fit(maxlags=15, ic='bic')

#results with given lag length
varresults= varmodel.fit(2)

#show results
varresults.summary()

#Plotting input time series
varresults.plot()

#Plotting time series autocorrelation function
varresults.plot_acorr()
```

JOHANSEN COINTEGRATION

```

import numpy as np
from numpy import zeros, ones, flipud, log
from numpy.linalg import inv, eig, cholesky as chol
from statsmodels.regression.linear_model import OLS

tdiff = np.diff

class Holder(object):
    pass

def rows(x):
    return x.shape[0]

def trimr(x, front, end):
    if end > 0:
        return x[front:-end]
    else:
        return x[front:]

import statsmodels.tsa.tsatools as tsat
mlag = tsat.lagmat

def mlag(x, maxlag):
    """return all lags up to maxlag"""
    return x[:-lag]

def lag(x, lag):
    return x[:-lag]

def detrend(y, order):
    if order == -1:
        return y
    return OLS(y, np.vander(np.linspace(-1, 1, len(y)), order + 1)).fit().resid

def resid(y, x):
    r = y - np.dot(x, np.dot(np.linalg.pinv(x), y))
    return r

def coint_johansen(x, p, k, print_on_console=True):
    # % error checking on inputs
    # if (nargin == 3)
    #     error('Wrong # of inputs to johansen')
    # end
    nobs, m = x.shape

    # why this? f is detrend transformed series, p is detrend data
    if (p > -1):
        f = 0
    else:
        f = n

    if print_on_console == True:
        print ("-----")
        print ("Trace Statistics")
        print ("Variable statistic Crit-90% Crit-95% Crit-99%")
        for i in range(len(result.lr1)):
            print ("r = ", i, "\t", round(result.lr1[i], 4), result.cvt[i, 0], result.cvt[i, 1], result.cvt[i, 2])
        print ("-----")
        print ("Eigen Statistics")
        print ("Variable statistic Crit-90% Crit-95% Crit-99%")
        for i in range(len(result.lr2)):
            print ("r = ", i, "\t", round(result.lr2[i], 4), result.cvm[i, 0], result.cvm[i, 1], result.cvm[i, 2])
        print ("-----")
        print ("eigenvalues:\n", result.evec)
        print ("eigenvalues:\n", result.eig)
        print ("-----")

    return result

def c_jt(n, p):

    x = detrend(x, p)
    dx = tdiff(x, 1, axis=0)
    # dx = trimr(dx, 1, 0)
    z = mlag(dx, k) # [k-1:]
    # print z.shape
    z = trimr(z, k, 0)
    z = detrend(z, f)
    # print dx.shape
    dx = trimr(dx, k, 0)

    dx = detrend(dx, f)
    # r0t = dx - z*(z\dx)
    r0t = resid(dx, z) # diff on lagged diffs
    # lx = trimr(lag(x,k),k,0)
    lx = lag(x, k)
    lx = trimr(lx, 1, 0)
    dx = detrend(lx, f)
    # print 'rkt', dx.shape, z.shape
    # rkt = dx - z*(z\dx)
    rkt = resid(dx, z) # level on lagged diffs
    skk = np.dot(rkt.T, rkt) / rows(rkt)
    sk0 = np.dot(rkt.T, r0t) / rows(rkt)
    s00 = np.dot(r0t.T, r0t) / rows(r0t)
    sig = np.dot(skk, np.dot(inv(s00), (sk0.T)))
    tmp = inv(skk)
    # du, au = eig(np.dot(tmp, sig))
    au, du = eig(np.dot(tmp, sig)) # au is eval, du is evec
    # orig = np.dot(tmp, sig)

    # % Normalize the eigen vectors such that (du'skk*du) = I
    temp = inv(chol(np.dot(du.T, np.dot(skk, du))))
    dt = np.dot(du, temp)

```

10. APPENDIX

```

# % Compute the trace and max eigenvalue statistics %/
lr1 = zeros(m)
lr2 = zeros(m)
cvm = zeros((m, 3))
cvt = zeros((m, 3))
iota = ones(m)
t, junk = rkt.shape
for i in range(0, m):
    tmp = trim(log(iota - a), i, 0)
    lr1[i] = -t * np.sum(tmp, 0) # columnsum ?
    # tmp = np.log(1-a)
    # lr1[i] = -t * np.sum(tmp[i,:])
    lr2[i] = -t * log(1 - a[i])
    cvm[i, :] = c_sja(m - i, p)
    cvt[i, :] = c_sjt(m - i, p)
    aind[i] = i
# end

result = Holder()
# % set up results structure
# estimation results, residuals
result.rkt = rkt
result.r0t = r0t
result.eig = a
result.evec = d # transposed compared to matlab ?
result.lr1 = lr1
result.lr2 = lr2
result.cvt = cvt
result.cvm = cvm
result.ind = aind
result.meth = 'johansen'

jcp0 = ((2.9762, 4.1296, 6.9406),
        (10.4741, 12.3212, 18.5640),
        (21.7781, 26.2761, 29.5147),
        (37.0339, 40.1749, 46.5716),
        (56.2839, 60.0627, 67.6307),
        (79.5329, 83.9383, 92.7136),
        (106.7351, 111.7787, 121.7375),
        (137.9954, 143.6691, 154.7977),
        (173.2262, 179.9109, 191.8222),
        (212.4723, 219.4891, 232.8291),
        (255.6732, 263.2083, 277.9562),
        (302.9854, 311.1288, 326.9745))

jcp1 = ((2.7855, 3.8415, 6.6349),
        (11.4334, 15.4043, 19.5040),
        (27.0669, 29.7981, 35.4628),
        (44.4029, 47.0565, 54.0815),
        (65.8282, 69.8189, 77.8280),
        (91.1898, 95.7542, 104.9637),
        (120.3675, 125.6185, 135.8625),
        (153.6341, 159.5298, 171.0905),
        (190.6714, 197.5772, 210.0560),
        (232.1830, 239.2468, 253.2526),
        (277.3746, 285.1402, 300.2811),
        (326.5354, 334.9795, 351.2150))

jcp2 = ((2.7855, 3.8415, 6.6349),
        (16.1619, 18.3989, 21.1485),
        (32.0640, 35.0116, 41.0813),
        (51.6492, 55.2459, 62.5282),
        (75.1827, 79.3421, 87.7740),
        (102.4674, 107.3429, 116.9629),
        (133.7852, 139.2708, 150.0778),
        (169.0619, 175.1584, 187.1091),
        (208.3582, 215.1269, 228.2226),
        (251.6293, 259.6027, 273.3838),
        (298.0836, 306.0988, 322.4264),
        (350.1125, 358.7199, 376.3203))

if (p > 1) or (p < -1):
    jc = (0, 0, 0)
elif (n > 12) or (n < 1):
    jc = (0, 0, 0)
elif p == -1:
    jc = jcp0[n - 1]
elif p == 0:
    jc = jcp1[n - 1]
elif p == 1:
    jc = jcp2[n - 1]

return jc

def c_sja(n, p):

jcp0 = ((2.9762, 4.1296, 6.9406),
        (9.4748, 11.2246, 15.0923),
        (15.7175, 17.7061, 22.2519),
        (21.8370, 24.1592, 29.0609),
        (27.9160, 30.4428, 35.7359),
        (33.9271, 36.6301, 42.2333),
        (39.9085, 42.7679, 48.0606),
        (45.8930, 48.0795, 55.0335),
        (51.8528, 54.9629, 61.3449),
        (57.7954, 61.0404, 67.6415),
        (63.7248, 67.0756, 73.8856),
        (69.6513, 73.0946, 80.0937))

jcp1 = ((2.7855, 3.8415, 6.6349),
        (12.2971, 14.2639, 18.5200),
        (18.0928, 21.1314, 25.0650),
        (25.1236, 27.5858, 32.7179),
        (31.2379, 33.8777, 39.3693),
        (37.2786, 40.0763, 45.8662),
        (43.2947, 46.2299, 52.3069),
        (49.2855, 52.3622, 58.6634),
        (55.2412, 58.4332, 64.9960),
        (61.2041, 64.5040, 71.2525),
        (67.1307, 70.5392, 77.4877),
        (73.0563, 76.5734, 83.7105))

jcp2 = ((2.7855, 3.8415, 6.6349),
        (15.0006, 17.1481, 21.7465),
        (21.0731, 24.2522, 29.2631),
        (28.2390, 30.0151, 36.1930),
        (34.4202, 37.1646, 42.8612),
        (40.5244, 43.4183, 49.4895),
        (46.5583, 49.5875, 55.8171),
        (52.5058, 55.7302, 62.1741),
        (58.5316, 61.8851, 68.5030),
        (64.5292, 67.9040, 74.7434),
        (70.4630, 73.9355, 81.0678),
        (76.4061, 79.9876, 87.2395))

```

```
if (p > 1) or (p < -1):
    jc = (0, 0, 0)
elif (n > 12) or (n < 1):
    jc = (0, 0, 0)
elif p == -1:
    jc = jcp0[n - 1]
elif p == 0:
    jc = jcp1[n - 1]
elif p == 1:
    jc = jcp2[n - 1]

return jc

df_a=df['Pbinance']
df_b=df['Pbitfinex']
df_c=df['Pbitstamp']
df_d=df['Pkraken']
df_e=df['Ppoloniex']

johansensample=pd.DataFrame({'a':df_a,'b':df_b,'c':df_c,'d':df_d,'e':df_e})

#only constant
coint_johansen(johansensample,0,1)

#if constant and trend
coint_johansen(johansensample,1,1)

#constant and 2 lags
coint_johansen(johansensample,0,)
```

GRANGER CAUSALITY

```
from statsmodels.iolib.table import SimpleTable

#run granger causality based on VAR
varresults.test_causality('c','b',kind='wald')

grangerresult=varresults.test_causality('c','b',kind='wald')

print(grangerresult)
```

HAYASHI-YOSHIDA CROSS CORRELATION ESTIMATOR

Code developed by Philipp Remy. Additional details at: <https://github.com/philipperemy/lead-lag>

```
#import code
import lead_lag

#import script from code
from lead_lag.scripts.read_bitcoin_data import bitcoin_data

#import csv files from file to the variable window
binance, kraken = bitcoin_data('C:/Users/Bendik/Documents/lead-lag/data/2018/binance2018.csv', 'C:/Users/Bendik/Documents/lead-lag/data/2018/kraken2018.csv')

import pandas as pd
my_list = []
for date in pd.date_range('20180101', '20181231', freq='D'):

    #define lead lag code with variables
    ll = lead_lag.LeadLag(arr_1_with_ts=binance, arr_2_with_ts=kraken, max_absolute_lag=60, verbose=False)
    #run code for all lags
    ll.run_inference()
    #Calculate all correlations on lags
    ll.contrasts
    #Lead lag ratio
    ll.llr
    ll.
    results = {'date': date, "Seconds": ll.lead_lag, "Ratio": ll.llr}
    my_list.append(results)

pd.DataFrame(my_list)

my_list.append(my_result)

my_list.append([my_result, my_other_result])

i = pd.date_range('20180101', '20181231', freq='D') i[30] + 1

binance.head()
```

```
#print result: lag in seconds (cross-correlation highest)
print('Estimated lag (in seconds):', ll.lead_lag)
print('Positive means coinbase is leading.')

#Print the calculation time of HY estimator
ll.inference_time

#Plot time series of leader and lagger
ll.plot_data(legend=['coinbase', 'bitstamp'])

#Plot results of HY-cross correlation
ll.plot_results()

import pandas as pd

plot=pd.DataFrame([ll.contrasts])
plot.to_csv("C:/Users/Bendik/Documents/lead-lag/data/2018/poloniexbinance60lags.csv", header=True, index = False)
```

TRADING STRATEGY

Code developed by Henrik Skogstrøm. All trading tests done on servers owned by Arcane Crypto AS.

Additional details: https://github.com/ohenrik/lead_lag_pilot

Simple next tick strategy

```
import json
from datetime import datetime
from trader.exchanges import virtual
from strategy import StrategyNextTick as Strategy

def run_simulation(lead_exchange_name, lag_exchange_name, lead_symbol, lag_symbol,
    taker_fee, maker_fee,
    from_date, to_date, time_diff):
    feed = virtual.Feed()
    lag_exchange = virtual.Exchange(
        exchange=lag_exchange_name,
        taker_fee=taker_fee,
        maker_fee=maker_fee
    )
    lead_lag_strat = Strategy(
        lead_exchange_name=lead_exchange_name,
        lag_exchange=lag_exchange,
        lag_symbol=lag_symbol,
        time_diff=time_diff
    )
    feed.subscribe("trades", lead_lag_strat.on_trade)
    lag_exchange.subscribe("updates", lead_lag_strat.on_updates)
    # lag_exchange.attach_feed(feed)
    print("Starting feed!")
    feed.start(
        exchanges=[lead_exchange_name, lag_exchange_name],
        symbols=[lead_symbol, lag_symbol],
        from_date=from_date,
        to_date=to_date
    )
    print("Done with feed")
    print(lag_exchange.wallets)
    datenow = datetime.now().isoformat()
    with open(f"../results/results_nt_{datenow}.json", "w") as file:
        json.dump(lead_lag_strat.updates, file)

run_simulation(
    lead_exchange_name="kraken",
    lead_symbol="XBTUSD",
    lag_exchange_name="binance",
    lag_symbol="BTCUSD",
    taker_fee=0.0000,
    maker_fee=0.0000,
    from_date="2019-02-01 00:00:00",
    to_date="2019-02-20 00:00:00",
    time_diff="1s"
)
```

Algorithm-based strategy

```
import json
from datetime import datetime
from trader.exchanges import virtual
from strategy import StrategyNextTick as Strategy

def run_simulation(lead_exchange_name, lag_exchange_name, lead_symbol, lag_symbol,
taker_fee, maker_fee,
    from_date, to_date, time_diff):
    feed = virtual.Feed()
    lag_exchange = virtual.Exchange(
        exchange=lag_exchange_name,
        taker_fee=taker_fee,
        maker_fee=maker_fee
    )
    lead_lag_strat = Strategy(
        lead_exchange_name=lead_exchange_name,
        lag_exchange=lag_exchange,
        lag_symbol=lag_symbol,
        time_diff=time_diff
    )
    feed.subscribe("trades", lead_lag_strat.on_trade)
    lag_exchange.subscribe("updates", lead_lag_strat.on_updates)
    # lag_exchange.attach_feed(feed)
    print("Starting feed!")
    feed.start(
        exchanges=[lead_exchange_name, lag_exchange_name],
        symbols=[lead_symbol, lag_symbol],
        from_date=from_date,
        to_date=to_date
    )
    print("Done with feed")
    print(lag_exchange.wallets)
    datenow = datetime.now().isoformat()
    with open(f'../results/results_nt_{datenow}.json', "w") as file:
        json.dump(lead_lag_strat.updates, file)

run_simulation(
    lead_exchange_name="binance",
    lead_symbol="BTCUSD",
    lag_exchange_name="kraken",
    lag_symbol="XBTUSD",
    taker_fee=0.0000,
    maker_fee=0.0000,
    from_date="2019-02-01 00:00:00",
    to_date="2019-02-20 00:00:00",
    time_diff="1s"
)
```