Università degli Studi di Pavia

Dipartimento di Ingegneria Industriale e dell'informazione

Scuola di Dottorato in
Ingegneria Elettronica, Informatica ed Elettrica

PhD in Computer Engineering
XXXII Ciclo

Applied Data Science Approaches in FinTech: Innovative Models for Bitcoin Price Dynamics

Supervisors:	Candidate:
Prof. Marco Porta	Iman H. Abu Hashish
Prof. Tullio Facchinetti	

A.A. 2018/2019

To whom I ask for two stars,

and come back bringing me the sky.

To my parents.

Acknowledgments

This thesis would not have seen the light without the support of my, countless, Professors through the past three years, to whom I will be forever grateful.

To Dr. Omar Al-Kadi who helped me in taking the first step towards this journey.

To Prof. Motta who has been supporting me way before I started my PhD and two years after.

To Prof. Giudici who gave me the time, knowledge and support to follow the path I have been always passionate about.

To Prof. Porta and Prof. Facchinetti, a beautiful end to such a hard journey.

To my family, dad, mom, all 7 brothers and sisters and Tabboush the cat.

Thank you.

A special thank you goes to my best friend, fiancé and soon to be husband, Moumen. I would not have survived without you.

Contents

List of Figures

List of Tables

Acronyms

ACF Autocorrelation Function

AI Artificial Intelligence

APAC Asia-Pacific

API Application Programming Interface

ARIMA Autoregressive Integrated Moving Average

ATM Automated Teller Machine

BNN Bayesian Neural Networks

BTC Bitcoin

CHIPS Clearing House Interbank Payments System

DLT Distributed Ledger Technology

DSAA IEEE International Conference on Data Science and Advanced Analytics

EM Expectation Maximization

FinTech Financial Technology

FSI Financial Services Industry

GA Genetic Algorithm

GARCH Generalized Autoegressive Conditional Heteroskedasticity

GB Gradient Boosting

GBR Gradient Boosting Regressor

GFC Global Financial Crisis

GLM Generalized Linear Models

GRU Gated Recurrent Unit

HMM Hidden Markov Model

IASC International Association for Statistical Computing

ICML International Conference on Machine Learning

IoT Internet of Things

ISI International Statistical Institute

IT Information Technology

KDD Knowledge Discovery in Databases

LSTM Long Short-Term Memory

M2M Machine-to-Machine

MA Moving Average

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MSE Mean Squared Error

NIPS Neural Information Processing Symposium

NLP Natural Language Processing

NN Neural Networks

NNAR Neural Network Autoregression

OHLC Open-High-Low-Close prices

P2P Peer-to-Peer

PACF Partial Autocorrelation Function

RF Random Forest

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

SVC Support Vector Classifier

SVM Support Vector Machine

SWIFT Society of Worldwide Interbank Financial Telecommunications

TDNN Time-Delay Neural Networks

VAR Vector Autoregressive

VIX The CBOE Volatility Index

Chapter 1

Introduction

The first step of introducing Data Science as a new field was done back in 1962, when John Tukey argued that it was fulfilling the requirements of defining a science [149]. The new research field has been looked at as the “Future of Data Analysis”. Ever since, the interest in developing the newly introduced field, at the time, has emerged.

During the years, many related terms have been presented and different definitions have been established. To name a few: datalogy, datamatics and datamaton [116, 117], exploratory data analysis [148], data analytics [39] and data mining [49]. In particular, the term “Data Science”, has been introduced in 1974 by Peter Naur [120], and defined as “the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences”. A more detailed explanation of Data Science evolution is provided in section 2.1.1.

From a more recent perspective, Data Science is perceived as an art [72], a fourth research methodology [41] and a fourth approach to scientific discovery [38] “in addition to experimentation, modeling, and computation”. Accordingly, it is safe to say that Data Science can be defined based on the way it is perceived [23]. At its simplest, it is “the science of data”. From a deeper perspective, it is “a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, such as statistics, informatics, computing, communication, management and sociology, to study data and its domain employing data science thinking," as proposed by Cao in [23].

Although the evolution of Data Science goes back to almost 57 years ago, the continuous innovations and technological advancements contributed in an on-going interest in Data Science in the current times. Particularly, living in a data-intensive environment, a natural consequence to such innovations, not only led to creating a new scientific agenda in the present research communities, but also founding data-driven start-ups, introducing a new data-driven economy and traditional business transformation, and establishing a new job market [23].

Generally speaking, Data Science can be applied to any given domain. For instance, Advertising, Aerospace and Astronomy, Bioinformatics, Consulting Services, Healthcare, Recommender Systems, Security, E-commerce, Banking, etc. [23]. It is without a doubt that the application of Data Science to any of the aforementioned domains is important. However, a domain of a particular interest is Economy and Finance, particularly, Financial Technology (FinTech).

Similar to Data Science, FinTech, as a field, is not new and its evolution goes way back to more than 150 years ago according to Arner et al. [7], that started from the introduction of the telegraph in 1838, followed by the successful laying of the first transatlantic cable in 1866, thus, making the first step towards constructing a fundamental infrastructure for financial globalization at the time. A more detailed explanation of FinTech evolution is provided in section 2.2.1.

As the name implies, FinTech is composed of two main fields, namely, technology and finance. Consequently, the continuous technological advancements have a direct effect on the way financial services are perceived and provided. Specifically, looking at financial services, from a Data Science perspective, FinTech holds many potentials for addressing possible related challenges through analyzing financial, and non-financial related data, that may lead to insightful information that can be exploited for improving such services or even create new ones.

Financial Data Science is a term proposed by Giudici [62] that describes the implementation of Data Science on technologically enabled financial innovations that are often driven by Data Science. Accordingly, possible applications include algorithmic trading, identification of trends and forecasting, peer-to-peer lending, robot advisers and cryptoassets-related analysis and regulation [23, 63].

More particularly, when Bitcoin, the most famous cryptocurrency, was introduced in 2008 [115], it has caught the attention of many researchers, given its potentials in offering low-cost, decentralized transfer of value anytime and anywhere in the world [63]. Consequently, implementing Data Science specifically on Bitcoin opens many opportunities for perceiving this newly presented non-traditional asset, through analyzing related pricing data to understand its respective market that has massively grown in popularity, prices and volatility [63].

Accordingly, the general objective of this work is to present applied Data Science approaches in FinTech by proposing innovative models that aim at studying and exploring Bitcoin price dynamics from descriptive and predictive perspectives. More specific objectives are presented in section 1.1 and the thesis structure is defined in section 1.2.

1.1 Thesis Objectives

The overall objective of this work is presenting applied Data Science approaches in FinTech by proposing novel descriptive and predictive models for Bitcoin price dynamics. Accordingly, to construct the specific objectives of this work, three different areas are considered, namely, Data Science, FinTech and Data Science for FinTech, as presented in sections 2.1, 2.2 and 2.3, respectively.

The objectives of this work fall within the three aforementioned fields. Taking the field of Data Science into account, the first objective of this thesis is to develop domain-specific models and algorithms that aim at learning, mining and discovering hidden knowledge in related data, that are not available in the body of knowledge. Within the field of FinTech, the second objective of this thesis is to address the emergence of cryptocurrencies, a genuine financial innovation [134], specifically, Bitcoin by providing empirical evidences and developing related theories. Finally, considering Data Science for FinTech, the third and final objective of this thesis is to propose innovative descriptive and predictive models aiming at studying two specific research areas, namely, Bitcoin price dynamics and Bitcoin price prediction.

Specifically, within the research area of Bitcoin price dynamics, the objectives of this work are summarized as follows:

1. acquiring empirical evidences on whether Bitcoin prices from different exchange markets are strongly connected as in an integrated and efficient market;

2. exploring whether such interactions are affected by exogenous prices of classical assets;

3. shedding more light on the non-conclusive properties of Bitcoin that have been found previously in the literature;

4. modeling such dependencies through the dynamics of their latent causes, attributed to time switches between different market regimes.

Moreover, within the research of Bitcoin price prediction, the objective of this work is to develop an innovative and efficient predictive model that addresses intra-daily prices and achieves more accurate prediction results than those found in the literature.

To address the objectives within Bitcoin price dynamics, two innovative models are proposed, namely, a Network VAR Model and a Hidden Markov Model, explained in detail in sections 3.1 and 3.2, respectively. Additionally, a Hybrid Hidden Markov Model and Genetic Algorithm-Optimized Long Short Term Memory Network model is proposed in 3.3 that tackles the objective within the research area of Bitcoin price prediction.

1.2 Thesis Structure

The structure of the thesis is organized as follows:

- **Chapter 2** follows a thorough literature review, considering three different fields, namely, Data Science in section 2.1, FinTech in 2.2 and Data Science for FinTech 2.3. A focus on Data Science evolution is presented in section 2.1.1 along with its challenges and opportunities in section 2.1.2. Similarly, the evolution of FinTech is introduced in section 2.2.1 followed by its drivers, challenges and opportunities in section 2.2.2. Finally, Data Science for FinTech encompasses the related researches in Bitcoin Price Dynamics as explained in section 2.3.1, as well as those available for Bitcoin Price Prediction as demonstrated in section 2.3.2.

- In **Chapter 3**, the proposed models for achieving the aforementioned objectives are introduced in detail. Specifically, section 3.1 explains the Network VAR model, followed by section 3.2 that describes the proposed Hidden Markov Model, adopted for daily Bitcoin prices modeling in section 3.2.1, as well as for intra-daily Bitcoin prices in section 3.2.2. Finally, section 3.3 introduces the theory behind Genetic Algorithms in section 3.3.1 and Long Short Term Memory networks in section 3.3.2, and proposes the Hybrid Hidden Markov Model and a Genetic Algorithm-Optimized LSTM Network in 3.3.3.

- Based on the proposed models, **Chapter 4** demonstrates the implementation process of the three models, from the process of data collection up to presenting the related descriptive and predictive results. The implementation of the Network VAR model is presented in 4.1, followed by the implementation of the Hidden Markov Model in section 4.2 and finally the implementation of the Hybrid Hidden Markov Model and Genetic Algorithm-Optimized LSTM Network in 4.3.

- **Chapter 5** concludes the thesis by summarizing its contributions and presenting respective future work.

- Finally, **Chapter 6** lists the submissions and publications carried out through the past three years, including abstracts and full papers in sections 6.1 and 6.2, respectively.

Chapter 2

Literature Review

Following a thorough literature review, this chapter summarizes the evolution of Data Science and Financial Technology (FinTech) as documented in the body of knowledge, followed by challenges and opportunities in these fields, as well as related researches and possible applications, aiming at positioning the contributions of this thesis within the already-existing researches, from both a general and a specific perspective.

Accordingly, sections 2.1 and 2.2 present the evolution of Data Science and FinTech along with possible challenges and opportunities, respectively, while section 2.3 illustrates the related researches for Data Science approaches in FinTech, highlighting the main contributions of this research area in the literature.

2.1 Data Science Overview

To begin with, this section addresses the available literature in Data Science starting from its evolution as explained in section 2.1.1 up to the possible challenges and opportunities, that naturally follow such an evolution, in section 2.1.2.

2.1.1 The Evolution of Data Science

"For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I

have had a cause to wonder and doubt." John Tukey, 1962.

In 1962, John Tukey [149], argued in his article "The Future of Data Analysis" that his perception of being a statistician, whose main interest is inferences, has changed due to the evolution of mathematical statistics at the time. He stated that his interests were more centered around data analysis; starting from planning the ways for gathering data, up to developing analysis procedures and interpretation techniques of the results of such procedures.

According to Tukey, data analysis is a field that consists of inferential techniques, incisive procedures and allocation, but in a larger and a more varied manner, where important statistical contributions are to be discovered to influence the practice of data analysis in the future, thus, seeking novelty in data analysis. He argued that such novelties can be achieved by seeking new questions to be asked, tackling old problems in a more realistic framework, establishing useful properties of observations and finding and evading lying constraints.

Having illustrated that, the confusion in positioning the fields of statistics and data analysis was evident at the time. To tackle this, Tukey listed three tests to define a science, namely, "1. intellectual content, 2. organization into an understandable form and, 3. reliance upon the test of experience as the ultimate standard of validity." Accordingly, data analysis passes the previously mentioned tests, making it a science. As for the contribution of statistics in the data analysis field, Tukey argued that it is subjective to statisticians and to the standards they follow; whether pure mathematics or the actual analysis of data. Consequently, the perception of the future of data analysis, at the time, depended on the willingness of statisticians to take a step forward to deal with the "rocky road" of real problems rather than the "smooth road" of unrealistic assumptions and arbitrary criteria for creating "a great science to all fields of science and technology."

At this point, the term "Data Science" as a whole have not been discussed yet. Few years after, Peter Naur and his colleagues introduced innovative terms in [116] and [117], that were used in their local environment, and presented them for general adoption. The terms introduced were:

1. *datalogy*: the science of the nature and use of data;
2. *datamatics*: the part of datalogy which deals with the processing of data by automatic means;
3. *datamaton*: an automatic device for processing data.

Naur argued that the use of such terms would contribute in gaining clarity as they implicitly include important aspects in data representation.

Introducing these terms has been the first step for Naur to establish the computing field as an academic subject [119] in Denmark. Moreover, Naur suggested that the term "datalogy" should be used as a replacement to computer science. Indeed, in 1969, Copenhagen University adopted the term and computer science has been practiced under the name of "datalogy, the study of data and data processes" at the time, and developed its own forms under the name of the "Copenhagen Tradition" [144]. Later on, Naur presented these works in his textbook "Concise Survey of Computer Methods" in 1974 [118] which, according to [23], the term "Data Science" has been first officially mentioned and defined.

In the summary of Naur's book [120], the definition starts with "data: a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some processes.", followed by "data science: the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences". Moreover, Naur stated that the use of data science lies in the application of data processes in building models that tackle realistic problems to create new, yet unknown, data that can directly help humans in decision-making related matters. Indeed, this conforms the future of data analysis that was perceived by Tukey back in 1962.

In 1977, Tukey wrote the "Exploratory Data Analysis" book [148] under the principle of "It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it." Tukey focused on introducing exploratory data analysis, not to show its use, but to expose several techniques for analyzing data effectively by exploiting simple arithmetic and easy-to-draw pictures. He argued that confirmatory and exploratory data analysis should work side by side, given that confirmatory data analysis has been always looked at as “mere descriptive statistics”, regardless of the benefits it may have provided. Thus, Tukey emphasized the importance of exploring the data by discovering beyond whatever descriptive appearances that have been found, to provide new insights.

Evidently, early definitions for data science and the related terms that were introduced to define the discipline, do not fully convey the current understanding of the corresponding research domain. However, in the same year at which Tukey introduced “Exploratory Data Analysis”, a step forward was taken by The International Statistical Institute (ISI) [83] where a new section was founded under the name of the International Association for Statistical Computing (IASC) [82], whose mission is to “convert data into information and knowledge”, through statistical computing; statistics in the communication and computer age.

The emerging interest in data analysis and its endless capabilities, at the time, resulted in introducing new related terms on a continuous basis. In 1987, the terms “Descriptive and Prescriptive Models” were introduced in [141] to depict “Knowledge Engineers’” interest in inferences and decision-making by describing the actual behavior versus the optimal strategies to adjust with a complex problem, respectively, for the development of knowledge bases. Not only new terms and definitions emerged, but also new communities. In 1989, the first workshop on Knowledge Discovery in Databases (KDD) took place [88], aiming at addressing the growth of databases and the need to create a corresponding knowledge base, using the available techniques provided by different fields, namely, expert systems, machine learning, intelligent databases, knowledge acquisition and statistics. This workshop was the first step to many coming KDD conferences [89]. Ever since, the research attractions on such fields became even more emerging, leading to the use of new terms such as “data analytics” [39] and “data mining” [49].

In addition to KDD, several well-known conferences were established, such as International Conference on Machine Learning (ICML) and Neural Information Processing Symposium (NIPS). Moreover, a dedicated conference on data science under the name of IEEE International Conference on Data Science and Advanced Analytics (DSAA) [43]. Consequently, it is evident that the continuous interest in these multidisciplinary fields has contributed in making data science the fastest growing and most popular computing, statistics, and interdisciplinary communities [23].

Fig. 2.1 summarizes the evolution of data science over time by listing representative researches, communities and research groups.

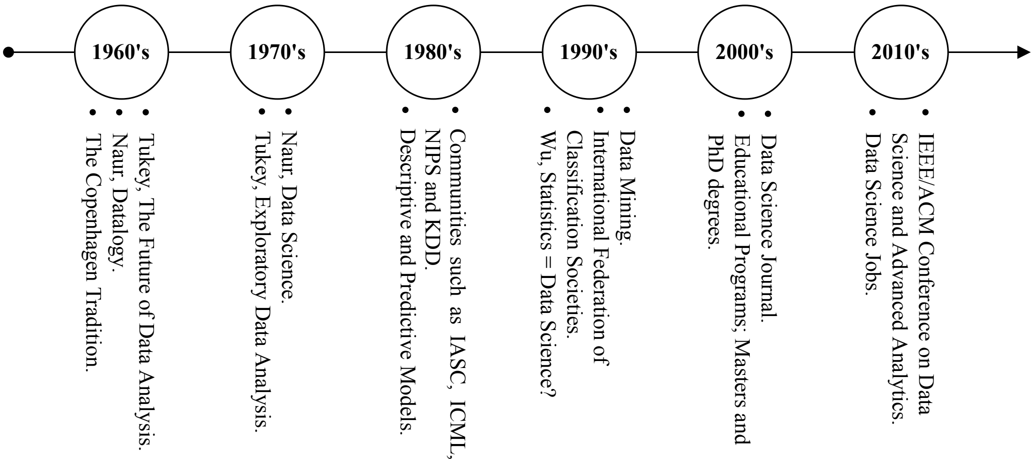

1960's
1970's
1980's
1990's
2000's
2010's
• Tukey, The Future of Data Analysis.
• Naur, Datalogy.
• The Copenhagen Tradition.
• Naur, Data Science.
• Tukey, Exploratory Data Analysis.
• Communities such as IASC, ICML, NIPS and KDD.
• Descriptive and Predictive Models.
• Data Mining.
• International Federation of Classification Societies.
• Wu, Statistics = Data Science?
• Data Science Journal.
• Educational Programs; Masters and PhD degrees.
• IEEE/ACM Conference on Data Science and Advanced Analytics.
• Data Science Jobs.

Figure 2.1: History of Data Science - Inspired from [23].

2.1.2 Data Science Challenges and Opportunities

Given the fact that data science is a new multidisciplinary field that has been built over already existing disciplines, it is difficult to address its related challenges and opportunities from a general point of view. For instance, many published researches that tackle challenges and opportunities of data science, considering one discipline at a time e.g. statistics [78, 77, 54, 158], data mining [50, 25, 19, 26], machine learning [102, 161], etc., can be found, however, few researches tackle the subject of data science as a whole [22].

To address this, Cao developed in [20, 22, 21, 23] a research map that captures the main challenges and research directions of data science as an emerging interdisciplinary field. Fig. 2.2 illustrates Cao's proposed research map of data science.

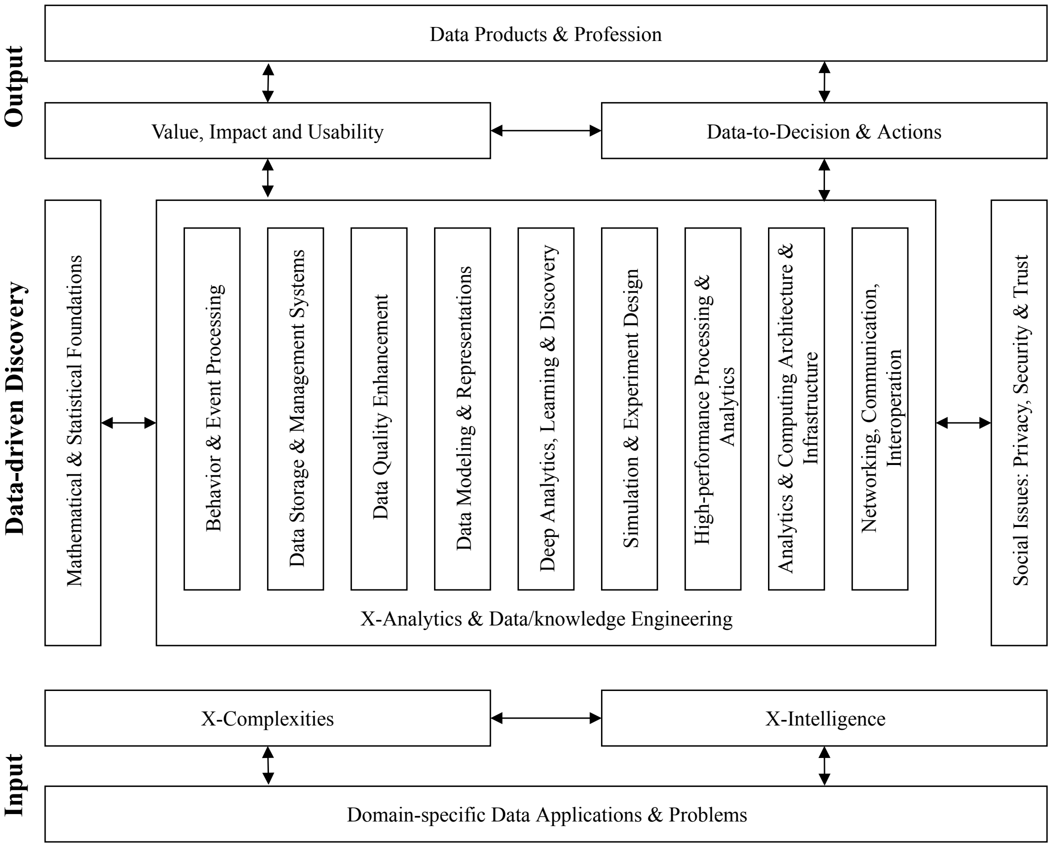

Output
Data Products & Profession
Value, Impact and Usability
Data-to-Decision & Actions
Data-driven Discovery
Mathematical & Statistical Foundations
Behavior & Event Processing
Data Storage & Management Systems
Data Quality Enhancement
Data Modeling & Representations
Deep Analytics, Learning & Discovery
Simulation & Experiment Design
High-performance Processing & Analytics
Analytics & Computing Architecture & Infrastructure
Networking, Communication, Interoperation
X-Analytics & Data/knowledge Engineering
Social Issues: Privacy, Security & Trust
X-Complexities
X-Intelligence
Input
Domain-specific Data Applications & Problems

Figure 2.2: Cao's Research Map of Data Science [21].

As Cao was developing the research map of data science, illustrated in fig. 2.2, he introduced the term "X-Generations" which depicts the new generations of complexities, intelligences and opportunities, and includes X-Complexities, X-Intelligence, X-Opportunities, X-Analytics and X-Informatics [23].

In a previous research [22], Cao argued that a data science problem is a complex system that has to address multiple complexities that have not been well-addressed or even addressed at all. Thus, X-Complexities refer to "diverse, widespread complexities that may be embedded in data, behavior, domain, societal aspects, organizational matters, environment, human involvement, network, and learning and decision-making". Thus, Cao argues that addressing such complexities using data science is an essential objective of data science, as it has outstanding potentials in exploring the embedded X-Intelligences, within these complexities, that consists of "data intelligence, behavior intelligence, domain intelligence, human intelligence, network intelligence, organizational intelligence, and environmental intelligence". A natural consequence of addressing X-Complexities and X-Intelligences is introducing X-Opportunities that can be specified in terms of X-Analytics which "refers to various opportunities discoverable by applying and conducting analytics on domain-particular data," and X-Informatics which "refers to the creation and application of informatics for specific domain problems" [23].

With that said, challenges and opportunities as proposed by Cao in [20, 22, 21, 23], can be summarized as follows:

- Data/business understanding challenges: identifying, specifying, representing, and quantifying X-Complexities and X-Intelligences, leading to effective methodologies and technologies.

- Mathematical and statistical foundations challenges: exploring whether, how and why already existing theoretical foundations are lacking the ability needed to deal with X-Complexities and X-Intelligences.

- X-Analytics and data/knowledge engineering challenges: developing domain-specific analytic theories, tools and systems that are not yet available in the body of knowledge, which involves:
  - Behavior and event processing: developing behavioral models that capture the evolution of behaviors and events of individuals and groups in the physical world.
  - Data storage and management systems: designing management systems that are able to deal with large amounts of data in real time.
  - Data quality enhancement: improving existing data quality issues, such as noise, uncertainty and missing values.

  - Data modeling, learning and mining: modeling, learning and mining data embedded in X-Complexities and X-Intelligences.
  - Data analytics, learning and discovery: analyzing, learning and discovering hidden knowledge in domain-specific data by introducing innovative models and algorithms.
  - Simulation and experimental design: simulating complexities, intelligences and processes in data and designing experiments to explore respective impacts.
  - High performance processing and analytics: processing and analyzing online, large scale, real time, internet and cloud-based data.
  - Analytics and computing architecture and infrastructure: facilitating previous challenges by providing new effective computing architecture and infrastructure.
  - Networking, communication and interoperation: supporting networking, communication and interoperation in distributed data science teams during solving data science problems.

- Quality and social issues challenges: dealing with social issues such as privacy, security, and trust, as well as enabling related data science tasks.
- Data value, impact and utility challenges: identifying and evaluating the value, impact and utility of domain-specific data.
- Data-to-decision and action-taking challenges: developing decision support systems, along with theories, to enable data-driven decision-making.

Having illustrated that, the main contributions of this thesis are within “X-Analytics and Data/Knowledge Engineering”, specifically, “Data Modeling, Learning and Mining” and “Deep Analytics, Learning and Discovery”. A more detailed literature review on the chosen domain and its related challenges can be found in section 2.2.

2.2 The Rise of Financial Technology

This section provides a thorough overview on FinTech and its evolution as explained in section 2.2.1, and discusses its relevant drivers, challenges and opportunities in section 2.2.2.

2.2.1 The Evolution of FinTech

"FinTech, the word which originates from the marriage of 'finance' and 'technology'." Zavolokina et al., 2017.

It is without a doubt that the FinTech industry is rapidly growing. In 2018, it was represented with a $111.8 billion in investments globally [95]. Obviously, such growth and development did not happen overnight. In 2015, American Banker, an award-winning daily trade newspaper [1], re-published an article [2], that was originally published in 1993, to shed light upon the first time the term "FinTech" was ever used. Accordingly, it is safe to say that the origin of the term goes back to the early 1990's, when Citicorp, a predecessor to the current Citigroup [31], initiated a banking research project called "FinTech" as the original name of the Financial Services Technologies Consortium [3]. Notwithstanding, the relationship between finance and technology goes way back and the rise of FinTech is only a natural consequence of the growth of these two disciplines over the years, together with other drivers to be explained in section 2.2.2.

To explain the evolution of FinTech, Douglas W. Arner, Jànos Barberis and Ross P. Buckley published three detailed papers [7, 9, 8] that explore FinTech and its evolution for a time period of almost 150 years. In [7, 9], the authors classified the evolution of FinTech to three (and a half) main eras, considering both developed and developing countries, namely, FinTech 1.0, FinTech 2.0, FinTech 3.0 and FinTech 3.5 to describe the evolution in the Asia-Pacific (APAC) region and Africa.

The authors argue that the development of FinTech followed a bottom-up approach, starting from FinTech 1.0 which took place from 1866 up to 1967. Despite the fact that finance was indeed interconnected with technology at the time, the

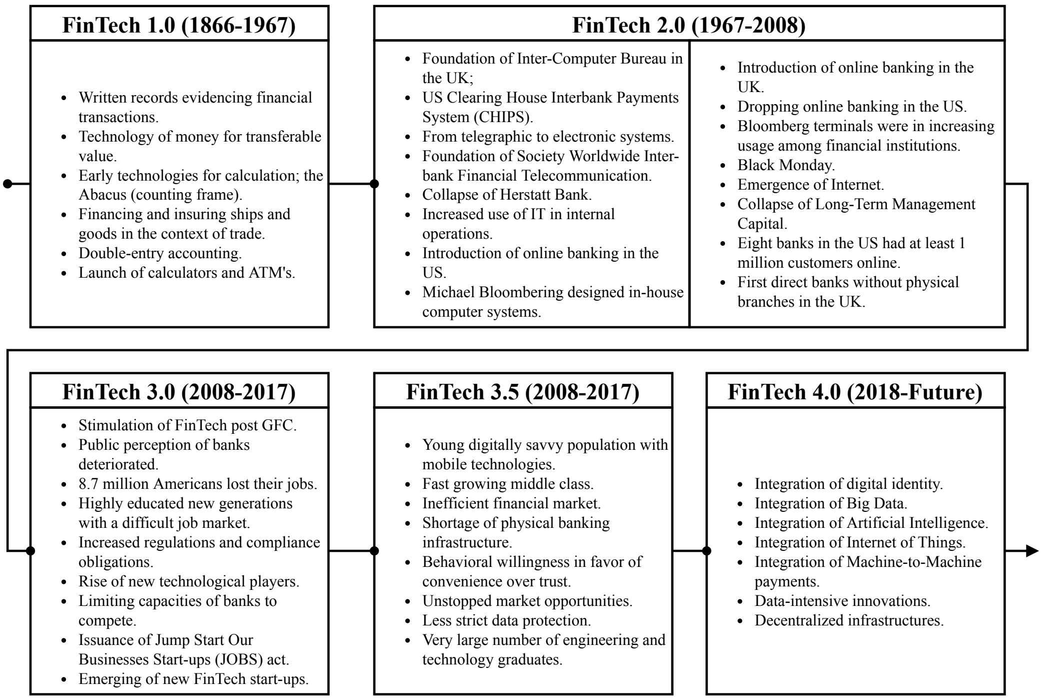
Financial Services Industry (FSI) was still an analogue industry [7]. To name a few, written records demonstrating financial transactions mark the earliest forms of Information Technology (IT), money as a technology evidencing transferable value, introduction of the Abacus (counting frame), financing and insuring goods and ships within the context of trade, and the development of double-entry accounting. Additionally, post-World War II, rapid developments took place and the first era of FinTech ended with the implementation of a global telex network that served as an infrastructure to the following era [9].

Introducing calculators and Automated Teller Machines (ATM) announced the commencement of FinTech 2.0. In this era, that took place from 1967 to 2008, not only was the FSI globalized but also digitalized [7]. On the one hand, numerous developments were achieved in different financial areas such as payments, securities and consumer areas. Additionally, many institutions were established such as Inter-Computer Bureau in the UK, US Clearing House Interbank Payments System (CHIPS), Fedwire and the Society of Worldwide Interbank Financial Telecommunications (SWIFT), evidencing the dominance of regulated financial firms that exploit information technologies to provide financial products and services. On the other hand, some financial crises took place which also contributed in the evolution of FinTech. For instance, the collapse of Herstatt Bank in Germany in 1974 as it failed to deliver US dollars to banks in New York due to time zone differences triggered the attention of regulators to set new guidelines to handle such risks that can be a natural consequence of the adoption of new payment systems. Similarly, Black Monday took place in 1987 when the stock markets around the world crashed due to, possibly, program trading where securities are bought and sold automatically according to pre-set price levels, which proves that at the time, the world was indeed connected through technology [9].

In 2008, one of the major financial crises took place, namely, the Global Financial Crisis (GFC) that started with bursting of the US housing bubble [59]. Unlike FinTech 1.0 where financial firms were dominating the FSI, new non-financial actors found an opportunity in implementing new technologies to financial services and become a part of the FSI providers [7], leading to the start of FinTech 3.0 in the developed countries up to present (2017), the year in which [8] was published. A shaken image of banks, damaged bank profitability and competitiveness, risen bank costs and unemployment are a few consequences of the GFC. Accordingly, new innovations started emerging to address such consequences. The authors in [8] argue that FinTech 3.0 would not have been founded had the GFC happened before 2008. This is due to the fact that the financial innovations of 2008 highly rely on smart phones and Application Programming Interface (API). To name but a few, Peer-to-Peer (P2P) lending, crowdfunding, algorithmic trading, etc. The importance of this era lies in the inclusion of non-financial actors in providing financial services to the public, as well as, the speed of development and innovations. Since 2008, FinTech has been evolving, in both developed and developing countries, due to varying causes. Consequently, the authors in [7, 9, 8] introduced FinTech 3.5 for the APAC region and Africa. For instance, underdeveloped banking and the spread of smart phones were the main two drivers for adopting FinTech in Africa [8]. As for the APAC region, the drivers can be summarized by disbelief in government-owned banking systems due to corruption, less IT spending by traditional banks, high usage of smart phones and limited branch network distribution [9]. Moreover, the authors of [7, 9] extended their work and introduced FinTech 4.0 in [8], starting from 2018 to future. The main characteristic of this era is the integration of digital identity, Internet of Things (IoT), Machine-to-Machine (M2M) payments, data-intensive innovations, and decentralized infrastructures to the FSI.

Fig. 2.3 illustrates the four eras of FinTech as suggested by the authors in [7, 9, 8], by listing a number of important events and characteristics within each era.

Having illustrated that, Thomas Puschmann elaborated on the classification of FinTech evolution suggested by Arner et al. in [7, 9] and argued that the evolution of FinTech can be categorized in three different areas that include five phases described as follows [129]:



FinTech 1.0 (1866-1967)

- Written records evidencing financial transactions.
- Technology of money for transferable value.
- Early technologies for calculation; the Abacus (counting frame).
- Financing and insuring ships and goods in the context of trade.
- Double-entry accounting.
- Launch of calculators and ATM's.

FinTech 2.0 (1967-2008)

- Foundation of Inter-Computer Bureau in the UK;
- US Clearing House Interbank Payments System (CHIPS).
- From telegraphic to electronic systems.
- Foundation of Society Worldwide Inter-bank Financial Telecommunication.
- Collapse of Herstatt Bank.
- Increased use of IT in internal operations.
- Introduction of online banking in the US.
- Michael Bloombering designed in-house computer systems.

- Introduction of online banking in the UK.
- Dropping online banking in the US.
- Bloomberg terminals were in increasing usage among financial institutions.
- Black Monday.
- Emergence of Internet.
- Collapse of Long-Term Management Capital.
- Eight banks in the US had at least 1 million customers online.
- First direct banks without physical branches in the UK.

FinTech 3.0 (2008-2017)

- Stimulation of FinTech post GFC.
- Public perception of banks deteriorated.
- 8.7 million Americans lost their jobs.
- Highly educated new generations with a difficult job market.
- Increased regulations and compliance obligations.
- Rise of new technological players.
- Limiting capacities of banks to compete.
- Issuance of Jump Start Our Businesses Start-ups (JOBS) act.
- Emerging of new FinTech start-ups.

FinTech 3.5 (2008-2017)

- Young digitally savvy population with mobile technologies.
- Fast growing middle class.
- Inefficient financial market.
- Shortage of physical banking infrastructure.
- Behavioral willingness in favor of convenience over trust.
- Unstopped market opportunities.
- Less strict data protection.
- Very large number of engineering and technology graduates.

FinTech 4.0 (2018-Future)

- Integration of digital identity.
- Integration of Big Data.
- Integration of Artificial Intelligence.
- Integration of Internet of Things.
- Integration of Machine-to-Machine payments.
- Data-intensive innovations.
- Decentralized infrastructures.

Figure 2.3: Evolution of FinTech.

1. Internal digitization: which includes the first three phases and focuses on the digitization of internal processes. The first phase took place until the 1960's and the goal of FinTech was to gain efficiency in support processes, while the second phase took place from 1960 to 1980 where the main focus was on back-office process. Finally, the third phase took place from 1980 to 2010 where the integration of IT was fully implemented in internal systems, unlike the first and second phases where the integration was non-existent or only partially existent, respectively.

2. Provider-oriented digitization: which contains the fourth phase that took (and will take) place from 2010 to 2020. The main focus is to integrate providers through outsourcing in different areas such as IT, payment systems, investments, etc. and thus, an external integration of financial services providers.

3. Customer-oriented digitization: which contains the fifth and final phase as suggested by Puschmann, where it will take place starting from 2020. The main focus will be centered around customers to create new ecosystems by the

integration of external non-financial services providers.

Lastly, it is important to note that the term "FinTech" can be viewed either as the integration of IT and finance to provide financial services efficiently or as the companies, firms or start-ups that provide such financial services [160]. These two views are used interchangeably in the literature.

2.2.2 FinTech Drivers, Challenges and Opportunities

Having illustrated the evolution of FinTech, it is now evident that finance and technology were, and still are, interconnected since almost 150 years ago, where the development of one discipline directly reflects on the other. However, it was also noted that this was not the only factor that drove the emergence and evolution of FinTech, where crises that occurred throughout the years played a major role as well, especially in the developed countries post GFC. Accordingly, this section focuses on the drivers of FinTech, possible challenges and opportunities, as well as research gaps and directions.

As previously explained, the interconnection between finance and technology is not novel, while the hype and emerging interest are, especially in the current times. As Arner et al. argues in [7, 9, 8], FinTech 3.0, from 2008 to present (2017) in [8], was crucially different than FinTech 1.0 and FinTech 2.0. for several reasons; the occurrence of GFC, the rapid developments in IT, and the diverse identities of financial services providers.

Indeed, Zavolokina et al. [160] confirmed the diverse identities of primary actors who would influence the evolution of FinTech. Their study shows that IT companies were dominant at the beginning of 1987-1989 and the end of 2001-2002. While financial institutions made an appearance in 1990 and 1998. Moreover, retailers were present in 1988, 1999 and 2001. Additionally, the presence of financial institutions peaked in 2004-2005 and 2007-2008. Moreover, the authors argued that in the period between 2010-2015, the diversity of identities has been constantly increasing where accelerators and consulting companies have become involved in FinTech at the time.

Within the same period of time, the presence of accelerators and consulting firms decreased, while the presence of financial institutions stabilized, and the presence of IT companies and start-ups significantly increased. The authors also confirmed the effect or financial crises on FinTech, illustrating the impact of the burst of the Dot-Com bubble [42] on diversifying the identities of financial services providers, which resulted in a shift from IT companies to financial institutions due to public's lack of trust. Finally, they illustrated the impact of GFC on increasing the diversities of financial services providers even more, as well as creating innovative topics in the context of FinTech.

Similarly, John Schindler [134] studied the drivers of FinTech to answer two questions, namely, "why FinTech is happening now?" and "why FinTech is getting much more attention than traditional innovation does?". Accordingly, the author explained the supply and demand framework and reflected it on possible drivers of FinTech to address the first question and introduced the concept of depth of financial innovation to address the second one. Consequently, from a supply point of view, the drivers of FinTech are; the use of technology which enables firms to provide innovative products and services, the significant increase of regulatory burdens contributed in creating innovative alternatives and macroeconomic conditions, which resulted in pressuring financial institutions to increase profits and cut costs. From a demand point of view, the increasing use of smart phones created new opportunities to fill new demands for new services and products to "match the mobile lifestyle", as well as demographics, specifically, millennials, created a new demand for such services. Additionally, Schindler pointed out the importance of the depth level of financial innovation, where a deeper innovation directly indicates a more profound innovation, thus, the ability to build further innovations over it. He then explained that there are three levels of financial innovation depth summarized as follows:

- Surface innovations: indicate innovations that do not change the fundamental nature of a financial service or product. Most of financial innovations fall in this level, such as online banking.

- Genuine innovations: indicate innovations that change the fundamental nature of a financial service or product, thus, creating new financial services and products. A small number of financial innovations fall in this level, such as P2P lending and cryptocurrencies.

- Foundational innovations: indicate significant innovations to the infrastructures and other foundations of the financial system. A rare number of financial innovations fall in this level, such as Distributed Ledger Technology (DLT) and the Blockchain.

Thus, it is evident that the potentials these innovations have are transforming, and will still transform, the financial system, explaining the hype around FinTech.

Furthermore, more recent researches were published agreeing with the previously mentioned drivers of FinTech, adding increasing levels of distrust towards financial institution, falling barriers to enter the digital disruption, attractive profit pools and increased awareness of regulators [127, 10].

Considering the numerous drivers of FinTech, it is still a challenging environment. As the name implies, the importance of having a digitally savvy talent is a must. However, the lack of talent in the current market is indeed one of the challenges faced in FinTech [10] where the authors argue that, even with the availability of such talents, employers will face a significant competition to hire them. Additionally, according to [145], 71% of millennials would rather go to their dentists than deal with bankers. Even though that the long history of disappointments in financial institutions faced by the public, specifically millennials, is indeed considered as a driver to FinTech, customers are still skeptic toward FinTech start-ups, even those that are regulated [10]. Likewise, regardless of the fact that the increased awareness of regulators towards FinTech is indeed an important driver [127], the regulatory burdens are still high [10].

Thus, taking FinTech drivers and challenges under consideration, many opportunities can be created in the research community to tackle such issues and overcome them. Recently, the editorial board of “The Review of Financial Studies” journal [147] provided possible research directions in FinTech [70], summarized as follows:

- Balancing theory and empirical work: the emergence of big data and data science provided the important potentials needed to analyze massive amounts of structured, semi-structured and unstructured data in different domains, giving researchers the opportunity to develop descriptive and predictive models through empirical studies. Likewise, developing the theories behind such descriptions and/or predictions is just as important and gives researchers possible grounds and explanations to think about and reflect on.

- International dimensions: while explaining the evolution of FinTech in section 2.2.1, Arner et. al [7, 9, 8] argued that the evolution of FinTech differs in developing countries and introduced FinTech 3.5. Indeed, the authors in [70] confirmed and argued that, given that the FSI is among the most developed industries in the US, fewer opportunities for innovation are available. Thus, international collaboration with developing countries create an opportunity for researchers in finance.

- Interdisciplinary collaborations: the interdisciplinary nature of FinTech clearly indicates the need to have knowledge in both finance and technology. Accordingly, collaborations between researchers in finance and computer science complement the missing knowledge in either discipline.

- Links to existing research: the authors argue that many of the considered issues in FinTech are previously tackled by many researchers. Accordingly, it is important to build new researches over already-existing ones rather than reinventing the wheel.

- Loss of trust in the current system: as previously mentioned, the increasing distrust in financial institutions triggered the evolution of FinTech [7, 9, 8, 160, 134, 127, 10]. For instance, Bitcoin was introduced as a decentralized P2P payment system [115] and was intended to take away the power from central banks [70, 55]. Accordingly, the authors argue that there is still a possibility to create a mutually beneficial situation and integrate new technological innovations without completely giving up on traditional financial institutions, which creates important opportunities for the research community.

- Rightsizing regulations: two important research questions are suggested by the authors within this area, namely, "how to regulate new FinTech entities relative to financial institutions?" and "whether the new forms of financing introduced by FinTech pose the same need for regulation as the traditional ones". Indeed, finding an appropriate trade-off in regulations for both FinTech start-ups and traditional financial institutions is a must, without undermining the importance of either one.

- A new market equilibrium: the authors argue that there are two possible options for traditional financial institutions; either they will be completely replaced with new FinTech firms, or they will adopt FinTech and become more advanced digitally. Accordingly, studying the effect of either option is a high priority research topic that can be tackled by developing new theories and conducting empirical studies.

- Welfare matters: according to [150], 1.7 billion of the population is still unbanked either by a traditional financial institution or a mobile money provider. Therefore, the authors argue that there is a greater mission for FinTech rather than the evolution of the financial sector, that is the overall welfare of its consumers. Thus, the research community is encouraged to consider such issues and propose appropriate solutions accompanied by relevant theories and empirical studies.

Having illustrated the evolution of FinTech in section 2.2.1 and its related drivers, challenges and opportunities, it should be noted that the chosen domain, to conduct applied data science approaches on, is FinTech. From a general point of view, the main contributions of this thesis fall within developing innovative descriptive and predictive models that address the emergence of cryptocurrencies, a genuine financial innovation [134], specifically, Bitcoin. A more detailed relative literature review is presented in section 2.3.

2.3 Data Science for Fin-Tech

"..if we don't innovate successfully, we're toast." Ian Narev, CEO of Commonwealth Bank, 2016.

In the previous sections 2.1 and 2.2, the evolutions of data science and FinTech, along with their respective challenges and opportunities, have been investigated. Thus, it is now apparent that these two interdisciplinary fields intersect in providing unprecedented opportunities and potentials to address domain-related issues that could not have been tackled before.

One of the previously mentioned challenges of FinTech, in section 2.2.2, is exploiting available data through the development of descriptive and predictive models, as a result of the emergence of data science, to gain insights and contribute in evolving current related theories. Indeed, in [23] Cao argues that data science has a major role in FinTech by analyzing related financial data to address potential problems and possible risks. To name a few, data science approaches can be applied in portfolio management optimization, price movements analyses, market trends identification and forecasting, Bitcoin and cryptocurrency analyses, fraud detection, market movement predictions, etc., in order to improve financial services. Similarly, Giudici in [62] introduced the concept of financial data science and described it as "the application of data science on the technologically enabled financial innovations (FinTech)". Indeed, according to a recent survey [56], data-oriented applications are considered crucial to the development of FinTech and meeting the new demands of the FSI.

On a similar note, Giudici argues in [63] that FinTech solutions are mainly driven by three technologies, namely, 1. Big Data Analytics with possible applications in

P2P lending, 2. Artificial Intelligence (AI) with possible applications in robo-advisory and 3. Blockchain technologies with possible applications in cryptoassets; confirming, on one hand, the importance of data science in FinTech for creating such solutions that are affecting the nature of the financial industry, and on the other hand, the potentials of the Blockchain technology that would also contribute in the evolution of FinTech and the future of the FSI.

In [104], the Blockchain is simply defined as "a technology to handle blocks in a chain", where each block is digitally signed with a hash value that links these blocks together forming the "chain". Although the concept of the Blockchain is not new [104] and its development is based on earlier technologies [110, 111], it has only gained its popularity when Bitcoin was introduced back in 2008 [115]. Within such a concept, the Blockchain acts as a DLT that records and verifies Bitcoin transactions while providing anonymity, security, immutability and a mutual trust between peers, given its nature in recording such transactions in a tamper-proof manner. Consequently, the research interest in the Blockchain as a DLT and in Bitcoin has been rapidly emerging in different domains, mainly technology, followed by economy, finance and accounting as the research developed [80]. Moreover, a particular interest in Bitcoin price characteristics has been emerging. Possible reasons may include, to name a few, the overall view of Bitcoin as an investment rather than a currency [162], the nature of Bitcoin's underlying infrastructure that can contribute in enabling illegal businesses [53, 133], a solution to the lack of trust in traditional financial institutions [133] and the growing use of cryptocurrency, specifically, Bitcoin the market's leader, resulting in a rise in trading volume as well as volatility [35].

Thus, generally, the objective of this thesis is to implement data science approaches within the FinTech domain, focusing on Blockchain technologies and its main application, i.e., cryptocurrencies. In particular, Bitcoin (BTC), the very first example of utilizing the Blockchain [104]. More specifically, to understand Bitcoin price dynamics through the development of descriptive and predictive data-driven models and providing empirical evidences. Accordingly, a thorough literature review is provided in sections 2.3.1 and 2.3.2, considering main researches on Bitcoin price dynamics through descriptive models and Bitcoin price prediction through predictive models, respectively.

2.3.1 Bitcoin Price Dynamics

As a result to a thorough systematic literature review on cryptocurrencies as financial assets [35], it has been reported that price dynamics is one of the most popular research areas in this field, together with market efficiency and cryptocurrency structure. Specifically, the research on Bitcoin, including its price characteristics, has risen from 5 to 485 papers over the period of 2011 to 2016, corresponding to the rise of Bitcoin prices within the same period [80]. Accordingly, following this emerging stream of research, this section explores researches concerning the study of cryptocurrency market prices, specifically, Bitcoin price dynamics, either from an endogenous or exogenous point of view.

From a theoretical viewpoint, Dwyer [44] examined the economics and financial properties of cryptocurrencies and argued that, the existence of a quantity limit along with the use of P2P networks, can create an economic equilibrium in which cryptocurrencies, including Bitcoin, have a positive value.

Trying to understand price dynamics from an empirical perspective, Bouoiyour et al. [16] applied a technique called empirical mode decomposition to assess Bitcoin prices formation, and argued that, although Bitcoin is considered as a speculative asset, it is extremely driven by long-term fundamentals. However, Corbet et al. [36] conducted multivariate statistical approach and studied the relationships between three cryptocurrencies; Bitcoin, Litecoin and Ripple, and their links to traditional financial assets, using a variance decomposition approach, and showed that the studied cryptocurrencies are strongly interconnected with each other by demonstrating similar patterns in returns and volatility while being relatively isolated from other financial assets such as gold, S&P 500 index, the CBOE Volatility Index (VIX) and GSCI. They also found that the volatility of cryptoassets is substantially higher than that of traditional assets, thus, showing a diversification benefit. Dyhrberg [45], Bouri et al. [17], reported similar conclusions, confirming the isolation of cryptoassets from traditional assets, pointing that such isolation emerges in the short-run rather than in the long-run, making the diversification benefit not conclusive. As a further support, a more recent paper by Ciaian et al. [30] applied autoregressive distributed lag models to daily data of Bitcoin and other sixteen cryptocurrencies and reported that they are interdependent but still independent from exogenous factors. They also found that such interdependent relationships are stronger in the short-run rather than in the long-run, consistently with the findings of [16]. Reporting similar conclusions, the authors in [46, 93, 139] with the sole difference of linking Bitcoin prices with S&P 500, in a weak manner.

Further arguments in favour of the endogenous nature of price dynamics have been provided by Blau [14], who studied the dynamics of Bitcoin prices using the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models and found that price volatility does not depend on speculative trading. Moreover, Polasik et al. [126] provided a regression analysis of the investment characteristics of Bitcoin and reported that Bitcoin returns are mainly driven by endogenous causes, such as sentiments on cryptocurrencies, or the total number of transactions. Finally, Viglione [153] found a positive relationship in a cross-country correlation between the level of technology and Bitcoin prices.

Understanding price interconnectedness is important, not only to describe the relationships between different asset prices, but also to understand whether prices in different markets quickly react to each other. In other words, whether cryptoasset markets are efficient. Brandvold et al. [18] was the first to address this matter by studying the price discovery process in Bitcoin markets. Using data from seven exchanges, over the period of April 2013 to February 2014, the authors found that Mt. Gox exchange market (bankrupting shortly after the sample period) and BTC-e exchange market were the price setters at the time. Pagnottoni et al. [121] extended the work in [18] and noted the increased role of Chinese exchanges. On a similar line of work, Urquhart [151] analyzed Bitcoin price return data from August 2010 to July 2016 to address the same issue, however, the efficient market hypothesis could not be confirmed. Consequently, Nadarajah et al. [114] revealed that an odd integer power transformation of Bitcoin price returns can be concluded as “weakly efficient”, and thus, the evidence on Bitcoin market efficiency is not conclusive.

Having illustrated that, and within this specific stream of research, the specific contributions of this thesis are two-fold; further acquiring empirical evidences on whether Bitcoin prices from different exchange markets are strongly connected as in an integrated and efficient market, and whether such interactions are affected by exogenous prices of classical assets. Thus, shedding more light on the non-conclusive properties of Bitcoin that have been found previously in the literature. Accordingly, an innovative, mostly descriptive, Network Vector Autoregressive (VAR) model has been developed to address these issues and the results have been published in [65]. This work will be explained in details in section 4.1.

While the developed model in the first contribution of this thesis, namely the Network VAR model, models the dependencies between the observed markets only, a further contribution is considered by modeling the same dependencies through the dynamics of their latent causes, attributed to time switches between different market regimes. To this aim, a Hidden Markov Model (HMM) is developed and the results have been submitted [64]. Similarly, this work will be explained in details in section 4.2.

2.3.2 Bitcoin Price Prediction

Market prediction in cryptocurrency is another popular research area given the challenges that need further considerations from the research community [40]. Within this section, a number of the most recent relative researches are discussed, specifically focusing on the proposed predictive model, frequency of the data used, related features and, finally, the predictive power of such models. Accordingly, the considered researches are classified, based on the data frequency used, into daily and intra-daily predictions, and discussed as follows.

Daily Bitcoin Price Prediction

Autoregressive Integrated Moving Average model (ARIMA) and Long Short-Term Memory network (LSTM) are observed to be two of the widely used models to predict Bitcoin prices.

For instance, Roy et al. proposed few statistical models to predict Bitcoin prices for the next 10 days, namely, Autoregressive (AR), Moving Average (MA) and ARIMA in [132]. They extracted daily Bitcoin prices from July 2013 to August, specifically, open, high, low, close prices (OHLC) along with volume and market capitalization. Accordingly, they reported that ARIMA achieved the highest accuracy of 90% while MA achieved the lowest accuracy of 87%. The same set of features has been extracted in [112] to predict the highest price of the day for a collection of 15 cryptocurrencies. However, only the high price has been fitted to ARIMA and all other features have been excluded. Accordingly, the authors reported an average accuracy of 86.4% for 95% of the considered cryptocurrencies with an average of 97.8% for Bitcoin. The authors argue that increasing the number of observations would contribute in achieving higher accuracy rates using ARIMA.

Similarly, Azari investigated the efficiency of ARIMA in predicting Bitcoin daily closing prices in [12]. Accordingly, he argued that fitting ARIMA with a 3-year long dataset of Bitcoin closing prices yields in a large Mean Squared Error (MSE) due to high volatility rates over such a long period. However, dividing the dataset, where each subset has its own unique trends, can help in decreasing the MSE of the predictions.

With reference to LSTM networks, the authors in [142], gathered a dataset consisting of market-related, sentiment-related, Blockchain-related features, along with stock market indices to understand the possible variables that play a role in predicting Bitcoin prices by fitting such features into LSTM network. The authors argue that, regardless the fact that Bitcoin prices are hard to predict, the aforementioned features are indeed crucial.

Moreover, Wu et al. [157] introduced a hybrid model of LSTM, paired with Autoregressive(2) to predict daily Bitcoin prices, using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). The proposed method achieved a Root Mean Squared Error (RMSE) of 247.33, compared to a conventional LSTM, which achieved an RMSE of 256.41. Exploiting sentiment-related features, the authors in [91] proposed a hybrid model that consists of sentiment analysis techniques, Natural Language Processing (NLP) and LSTM to predict the direction of Bitcoin price changes and achieved an accuracy of 67.6%.

Highlighting the efficiency of LSTM networks in [81], Support Vector Machine (SVM), Linear Regression, Neural Networks (NN), LSTM and rolling LSTM networks have been implemented using 17 different features that are related to macroeconomics, global exchange rates and Blockchain information. Accordingly, an RMSE of 59.04 and a Mean Absolute Percentage Error (MAPE) of 0.044 have been reported for rolling LSTM model, which outperforms the aforementioned considered models. On a similar line of work, McNally et al. [109] compared the performance of LSTM, Recurrent Neural Network (RNN) and ARIMA models in predicting the direction of Bitcoin price changes, confirming the efficiency of LSTM which achieved a prediction accuracy of 52.78%, thus, outperforming Bayesian-optimized RNN and ARIMA, which achieved an accuracy rate of 50.25% and 50.05%, respectively. Likewise, a better performance for LSTM was reported in [98] compared to Generalized Regression Neural Networks (GRNN), with an RMSE of 2.75×10^3 compared to 8.80×10^3, respectively, for predicting Bitcoin prices.

Continuing on this line, a number of researches conducted a comparison between the two popular models in Bitcoin prices prediction, namely ARIMA and LSTM. Karakoyun et al. [86] reported a MAPE of 11.86% for ARIMA, which significantly decreased to 1.4% when fitting the same dataset into LSTM networks. Similarly, Press [128] showed the efficiency of a novel modified LSTM compared to ARIMA, when using them in building a scalable online platform to predict a real-time stream of daily Bitcoin prices.

Further models have been considered for Bitcoin prices predictions. To name a few, Munim et al. [113] proposed Neural Network Autoregression (NNAR) models to predict next-day Bitcoin prices, achieving an RMSE of 0.069. Moreover, the authors in [143] proposed α-Sutte Indicator that achieved the lowest MSE compared to both ARIMA and NNAR. Other models include Bayesian Neural Network (BNN) as in [85], Random Forest (RF) [71], Neuro Fuzzy techniques [11], Logistic and Linear Regression [99, 4], Time-Delay Neural Networks (TDNN) [74], Averaged One-Dependence Estimators [90], Support Vector Classifier (SVC) and Gradient Boosting Regressor (GBR) [154], VAR [106] and, finally, Binomial Generalized Linear Models (GLM) [105].

Intra-daily Bitcoin Price Prediction

It is evident that the frequency of the data is an important factor to be considered while choosing the appropriate model for Bitcoin prices prediction. As previously mentioned, Binomial GLM was used to predict daily Bitcoin prices in [105], along with SVM and RF, achieving an accuracy of 0.9879, 0.2716 and 0.9498, respectively. However, when the same models are used for intra-day data, the accuracy rates decreased significantly. Specifically, for Binomial GLM which achieved, when fitted to 10-second data an accuracy of 0.085, and an accuracy of 0.539 when fitted to 10-minute data.

Nevertheless, LSTM networks are still dominating the considered models while predicting Bitcoin intra-daily prices. Cerda et al. [27] proposed the use of LSTM networks to predict intra-daily prices using 5-minute Bitcoin prices from July 2018 to December 2018, along with sentiment-related features extracted from crypto-influencers posts from Twitter, achieving and RMSE of 10.87. Within the same stream of work, the authors in [138] investigated the abilities Google trends and Telegram messaging platform have in predicting cryptocurrencies prices. To address this, they used LSTM to fit hourly data of Bitcoin and Ethereum prices using pricing-related features, Google trends features, and sentiment-related features extracted from Telegram. Accordingly, for Bitcoin, using pricing-related features, namely, price and trading volumes, LSTM achieved an accuracy of 0.62. Extending the model using Google trends slightly increased the accuracy up to 0.64. However, extending the model using Telegram features, increased the accuracy significantly up to 0.76. Finally, implementing the LSTM model with all the features combined achieved an accuracy of 0.63. Moreover, the authors in [100] extended LSTM networks by utilizing Word2Vec models to predict Bitcoin price fluctuations hourly. The proposed hybrid model achieved a predictive accuracy of 54.5% analyzing Reddit posts from crypto-communities, 12 hours in the future.

Comparing the predictive performance of LSTM in predicting intra-daily Bitcoin prices, Phaladisailoed et al. [124] used 1-minute data from January 2012 to January 2018 to implement several models, namely, Theil-Sen Regression, Huber Regression, LSTM, and Gated Recurrent Unit (GRU) in order to discover the most efficient model to predict Bitcoin prices. Results showed the power of GRU by achieving the lowest MSE of 0.00002 among the previously mentioned models. Despite that, comparing LSTM to Gradient Boosting (GB) as in [96], for predicting cryptocurrency prices using 10-minute data, indicated a performance improvement when using LSTM instead of GB.

Further models have been considered to predict intra-daily prices as well. For instance, VAR model was implemented in [136] using 5-minute data from September 2014 to August 2018. Moreover, Multi-Linear Regression as in [84], Generative Temporal Mixture model as in [75, 76], RNN [94], Bayesian Regression [135], Random Sampling Method as proposed in [137], and, finally, Linear and Logistic Regression [73].

Consequently, the third contribution of this thesis within Bitcoin price prediction aims at developing an innovative and efficient predictive model that addresses intra-daily prices and achieves more accurate prediction results than those found in the literature. Specifically, by extending the previously proposed HMM, as mentioned in 2.3.1, and proposing a hybrid model for Bitcoin prices prediction using HMM and LSTM networks. The results have been published in [5] and will be explained in detail in section 4.3.

Finally, to conclude this chapter, a summary of the considered papers in this section is provided in tables 2.1 and 2.2, for daily and intra-daily Bitcoin prices prediction, respectively.

Table 2.1: Summary of Reviewed Papers in Cryptocurrency Daily Prices Prediction.

Paper	*Features Considered*	*Model Used*	*Evaluation Metric*	*Predictive Power*
[132]	OHLC, Volumes, Market Capital	ARIMA	Accuracy	90.31%
[112]	OHLC, Market Capital	ARIMA	Accuracy	86.42%
[12]	Closing Price	ARIMA	MSE	16000
[142]	Stock Indices, Sentiments, Blockchain	LSTM	MAE	N.A.
[157]	ACF, PACF	AR-LSTM	RMSE	247.33
[91]	Popularity, Subjectivity, S&P500, Crypto Prices	NLP-LSTM	Accuracy	67.6%
[81]	Blockchain	Rolling LSTM	RMSE	59.04

continued on next page

Table 2.1 – continued from previous page

Paper	*Features Considered*	*Model Used*	*Evaluation Metric*	*Predictive Power*
[109]	OHLC, Blockchain, Moving Avg.	LSTM	Accuracy	52.78%
[98]	Crypto Prices	LSTM	RMSE	2570
[86]	Crypto Prices	LSTM	MAPE	1.40%
[128]	Crypto Prices	LSTM	N.A.	N.A.
[113]	Crypto Prices	ARIMA	RMSE	0.042
[143]	Crypto Prices	α-sutte	MSE	121362.34
[85]	Blockchain, Macroeconomic, Exchange Rates	Bayesian Neural Network	RMSE	0.0244
[71]	OHLC, Volumes	Random Forest	RMSE	193
[11]	Crypto Prices	Neuro Fuzzy	RMSE	0.0376
[99]	Crypto Prices, N-gram	Logistic Regression	Accuracy	61.9%
[4]	Sentiments, Google Trends, Tweets	Linear Regression	N.A.	N.A.

continued on next page

Table 2.1 – continued from previous page

Paper	*Features Considered*	*Model Used*	*Evaluation Metric*	*Predictive Power*
[74]	OHLC, Volumes	TDNN	MSE	42×10^{-6}
[90]	Sentiments, Prices, Transactions	Averaged One-Dependence Estimators	Accuracy	79.57%
[69]	Crypto Prices	SVM	Accuracy	62.31%
[106]	Crypto Prices, Google Trends, Volatility, S&P500, Volumes, Social Media	VAR	N.A.	N.A.
[105]	Network, Market	Binomial GLM	Accuracy	98%

Table 2.2: Summary of Reviewed Papers in Cryptocurrency Intra-daily Prices Prediction.

Paper	*Data Freq.*	*Features Considered*	*Model Used*	*Evaluation Metric*	*Predictive Power*
[105]	10-minute	Network, Market	Random Forest	Accuracy	57%
[27]	5-minute	Opening Price, Closing Prices, Sentiments	LSTM	N.A.	N.A.
[138]	Hourly	Crypto Prices, Sentiments, Telegram	LSTM	Accuracy	63%
[100]	Hourly	Market, Social Media	Word2Vec-LSTM	Accuracy	54.5%
[124]	1-minute	OHLC, Volumes	LSTM	MSE	2×10^{-5}
[96]	10-minute	OHLC, Volumes	LSTM	F1 Score	0.63-0.68
[136]	5-minute	Volumes, Volatility, Tweets	VAR	N.A.	N.A.
[84]	1-minute	Crypto Prices, Tweets	Multi-Linear Regression	R^2 Score	44%
[75, 76]	Hourly	Spread, Ask/Bid Volumes, Depth, Slope	Generative Temporal Mixture	RMSE	0.025
[94]	Tick	Crypto Prices	RNN	N.A.	N.A.
[135]	1-second	Crypto Prices, Order Book	Bayesian Regression	N.A.	N.A.
[137]	1-minute	OHLC	Random Sampling Method	Accuracy	47%
[73]	Every Transaction	Crypto Prices, Transactions	Linear Regression	MSE	1.94

Chapter 3

Methodology

Prior to presenting the results achieved within the contributions of this thesis, a thorough explanation of the proposed models, from a theoretical perspective, is provided in this chapter. Consequently, section 3.1 demonstrates the proposed Network VAR model for understanding Bitcoin price dynamics from an endogenous as well as exogenous points of view. Section 3.2 presents the adopted Hidden Markov Model (HMM) for understanding how Bitcoin prices switch between different regimes, going from "bull" to "stable" and "bear" behaviors. Finally, section 3.3 illustrates the proposed hybrid model for predicting Bitcoin prices using HMM and Genetic Algorithm (GA) optimized LSTM networks.

3.1 Network VAR Models

Let y_t^i be the price of Bitcoin in a specific exchange market $i\,(i = 1, \ldots, I)$, at time $t\,(t = 1, \ldots, T)$. We assume that y_t^i is a function of:

1. an autoregressive component, that expresses the dependency on the past prices of the same exchange y_{t-1}^i;

2. a cross-sectional component, that expresses the contemporaneous dependency on the prices of other exchanges y_t^j and

3. a stochastic residual.

Formally, for each price i and time t we assume that the following holds:

$$y_t^i = \sum_{p=1}^{p_0} \alpha_p^i y_{t-p}^i + \sum_{j \neq i} \beta^{ij} y_t^j + \epsilon_t^i, \tag{3.1}$$

where p is a time lag (with a maximum lag $p_0 < t$), α_p^i and β^{ij} are unknown coefficients to be estimated from the considered data, and ϵ_t^i are standard Gaussian residuals, which are independent across time and exchanges.

Equation 3.1 models Bitcoin price dynamics as a structural VAR, in which the price in each exchange market depends on its p past values, through the idiosyncratic autoregressive component $\sum_{p=1}^{p_0} \alpha_p^i y_{t-p}^i$ and, in addition, it depends on the contemporary values of the other markets, through the systemic component $\sum_{j \neq i} \beta^{ij} y_t^j$.

The previous model can be expressed in a more compact matrix form, as follows:

$$Y_t = \sum_{p=1}^{p_0} A_p Y_{t-p} + B_0 Y_t + E_t, \tag{3.2}$$

where Y_t is an I-dimensional vector containing the prices of all exchanges at time t, Y_{t-p} is the same vector, lagged at time $t-p$, A_p is a $p \times I$ matrix that contains the autoregressive coefficients, B_0 is a $I \times I$ symmetric matrix with null diagonal elements containing the contemporaneous coefficients and, finally, E_t is a vector of standard Gaussian residuals independent across time.

For estimation purposes, the model in 3.2 can be transformed in a reduced form, thus becoming:

$$Y_t = \Gamma_1 Y_{t-1} + ... + \Gamma_p Y_{t-p} + U_t, \tag{3.3}$$

with

$$\begin{cases} \Gamma_1 = (\mathbb{I} - B_0)^{-1} A_1, \\ \dots \\ \Gamma_p = (\mathbb{I} - B_0)^{-1} A_p, \\ U_t = (\mathbb{I} - B_0)^{-1} E_t. \end{cases} \tag{3.4}$$

The previous formulation allows the estimation of the vectors of modified autoregressive coefficients $\Gamma_1, ..., \Gamma_p$, using time series data on Bitcoin prices contained in the stacked vector $\{Y_1, \dots, Y_t, \dots, Y_T\}$.

However, we are not interested in estimating Γ_p, but in separately estimating its components $\{A_1, ..., A_p\}$ and B_0, disentangling the autoregressive part from the contemporaneous one. In this sense, once B_0 is obtained, $\{A_1, ..., A_p\}$ can be derived from 3.4.

To estimate B_0, note that $(\mathbb{I} - B_0)U_t = E_t$, so that $U_t = B_0 U_t + E_t$. This implies that, for each exchange i;

$$U_t^i = \sum_{j \neq i} \beta^{ij} U_t^j + \epsilon_t^i, \tag{3.5}$$

meaning that the off-diagonal elements of B_0 can be obtained by regressing each modified residual, derived from the application of 3.3, on those of the other exchanges.

Note that the regression model in 3.5 is based on the transformation derived in equation 3.4, which makes the modified residuals correlated. The direction of such correlation is, however, unknown. In the application of 3.5 it is thus not clear which price residual assumes the form of a response variable, and which one is of an explanatory regressor.

A simple solution to this problem would be to estimate all possible regressions, that is, to regress each of Bitcoin prices on all the others. However, this procedure would be, besides illogical, computationally inefficient. To solve this issue, we approximate each pair of regression coefficients β^{ij} and β^{ji} is proposed, representing two opposite causality directions, with their partial correlation coefficient, which is undirected, but univocally determined by them.

Formally, let $\Sigma = Corr(U)$ be the correlation matrix between the modified residuals, and let Σ^{-1} be its inverse, with elements σ^{ij}. The partial correlation coefficient $\rho_{ij|S}$ between the residuals U^i and U^j, conditional on the remaining residuals $(U^s, s = 1, \ldots, S)$, where $S = I \setminus \{i, j\}$, can be obtained by:

$$\rho_{ij|S} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}. \tag{3.6}$$

It can be shown that:

$$|\rho_{ij|S}| = \sqrt{\beta^{ij} \cdot \beta^{ji}}, \tag{3.7}$$

which means that the absolute value of the partial correlation coefficient between U^i and U^j, given all the other residuals, can be obtained as the geometric average between the coefficients β^{ij} and β^{ji} defined by equation 3.5 setting, respectively, i rather than j as response variables. Equation 3.7 justifies the replacement of β^{ij} and β^{ji} with their corresponding partial correlation coefficient $\rho_{ij|S}$.

From an economic viewpoint, the partial correlation coefficient expresses how the Bitcoin price of an exchange i is affected by the contemporaneous price of other exchanges, or of classical assets, $j \neq i$, keeping other prices fixed. An important advantage that derives from the employment of partial correlations lies in the possibility of employing correlation network models based on the conditional independence relationships described by partial correlations.

More precisely, let us assume that the vectors U_t are independently distributed, according to a multivariate normal distribution $\mathcal{N}_I\left(0, \Sigma\right)$, where Σ represents the correlation matrix, that is assumed to be non-singular.

A correlation network model can be represented by an undirected graph G such that $G = (V, E)$, with a set of nodes $V = \{1, \ldots, I\}$, and an edge set $E = V \times V$ that describes the connections between the nodes. G can be represented by a binary adjacency matrix E with elements e_{ij}, each of them providing the information of whether a pair of vertices in G is, symmetrically, linked between each other ($e_{ij} = 1$) or not ($e_{ij} = 0$). If the nodes V of G are put in correspondence with the random variables $U_1, \ldots, U_I$, the edge set E induces conditional independence on U via the so-called Markov properties [101].

Let Σ^{-1} be the inverse of Σ, whose elements can be indicated as $\{\sigma^{ij}\}$. Whittaker [156] proved that the following equivalence holds:

$$\rho_{ij|S} = 0 \Longleftrightarrow U_i \perp U_j | U_{V \setminus \{i,j\}} \Longleftrightarrow e_{ij} = 0$$

where the symbol $\perp$ indicates conditional independence.

From a graph-theoretic viewpoint, the previous equivalence means that a link between two exchange prices is present if, and only if, the corresponding partial correlation coefficient is significantly different from zero. From a financial viewpoint, the previous equivalence implies that, if the partial correlation between two measures is equal to zero, the corresponding price residuals are conditionally independent and, therefore, the corresponding exchanges do not directly impact each other. Lastly, from a statistical viewpoint, it is also possible to test the null hypothesis that two exchanges are conditionally independent, testing whether the corresponding partial correlation coefficient is equal to zero, by means of the statistical test described by Whittaker [156] or by Giudici [61].

Summing up, a full correlation network model among Bitcoin exchange prices can be estimated on the basis of the pairwise partial correlation coefficients between the modified residuals.

The implementation of the proposed model is illustrated, in detail, in section 4.1.

3.2 Hidden Markov Models

Hidden Markov Models are generative probabilistic models in which a sequence of observations Y is generated by a sequence of internal hidden states S [60].

A discrete hidden Markov model assumes that each observation Y_t, for ($t =$

$1, \ldots, T)$ is generated by a stochastic process whose state S is a discrete random variable, hidden to the observer. The probability of observing Y_t at any given time t can be described by a statistical distribution, conditional on S_t, usually known up to a parameter θ. It also assumes that the time transition between subsequent states, $(S_1, \ldots, S_T)$ follows a Markov chain, typically of first order.

More formally, the previous assumptions mean that the joint distribution of the observed time series $Y_{1:T}$, and of the corresponding hidden states $S_{1:T}$, can be factorized as:

$$P(Y_{1:T}, S_{1:T}) = \prod_{t=1}^{T} P(Y_t|S_t)P(S_t|S_{t-1}) \tag{3.8}$$

in which $P(S_1|S_0) = P(S_1)$ is the unconditional distribution of the initial state.

To further specify the probability distribution in 3.8, the following components need to be defined:

1. the conditional distribution $P(Y_t|S_t)$, that links the observed variables with the hidden states;

2. the state transition matrix which defines the conditional probabilities $P(S_t|S_{t-1})$ and

3. the probability distribution for the initial state $P(S_1)$.

Hidden Markov models are usually assumed to be time invariant, which implies that the conditional distributions and the state transition matrices do not depend on t.

To simplify, an HMM is defined by A, B and π, and implicitly, by the number of observations N, as well as the number of hidden states M. In the model, A represents the state transition probability $M \times M$ matrix, B represents the observation probability $M \times N$ matrix, and π is the initial state distribution. Thus, an HMM can be defined as:

$$\lambda = (A, B, \pi) \tag{3.9}$$

HMM are used to solve three fundamental problems [140], that can be summarized as follows:

- Problem 1: given the model $\lambda = (A, B, \pi)$, and a sequence of observations Y, determine the likelihood of the observed data to the given model;
- Problem 2: given the model $\lambda = (A, B, \pi)$, and a sequence of observations Y, determine the optimal sequence of hidden states underlying the Markov process;
- Problem 3: given a sequence of observations Y, estimate the model's parameters A, B and π.

Accordingly, the purpose of adopting HMM is estimating its respective parameters (Problem 3), given a sequence of observations Y, followed by calculating the likelihood of the data (Problem 1), and finally, determining the optimal sequence of the hidden states (Problem 2).

The proposed HMM is adopted twice, mainly for descriptive modeling: using daily and intra-daily Bitcoin prices as illustrated in sections 3.2.1 and 3.2.2, respectively, along with their respective implementations in sections 4.2 and 4.3.

3.2.1 Hidden Markov Models for Daily Bitcoin Prices

Within the context of daily Bitcoin prices, only endogenous factors are taken into consideration, namely, daily Bitcoin closing prices in different exchange markets. Exogenous factors have been excluded since they do not affect or weakly affect the dynamics of Bitcoin prices as indicated by the obtained results from the implementation of the Network VAR model. The results are explained in detail in section 4.1.2. Accordingly, given a sequence of observations Y where each observation Y_t is a vector of market prices Y_t^i, $(i = 1, \ldots, I; t = 1, \ldots, T)$, one for each of the I considered Bitcoin exchanges. We assume that at any given time point t, the vector Y_t follows an HMM, specified by the joint probability distribution:

$$P(S^i_{1:T}, Y^i_{1:T}) = P(S^i_1)P(Y^i_1|S^i_1)\prod_{t=2}^{T} P(S^i_t|S^i_{t-1})P(S^i_t|Y^i_t) \tag{3.10}$$

Moreover, given the multivariate nature of Y_t, it is also assumed that each conditional distribution $P(Y_t|S_t)$ is a multivariate Gaussian, with a mean vector and an unknown variance-covariance matrix Σ, which will be estimated using the available data, along with the transition matrix of the hidden states. The initial state will be instead considered as a given constant value.

To apply the proposed model on the available data, several computational, widely known, algorithms are needed in order to solve the previously mentioned problems, as well as to compare and test different structures while modeling. Consequently, the first and second problems, namely, determining the likelihood of the data and the optimal sequence of hidden states, can be computed by using Viterbi and the Forward-Backward algorithms, while estimating the model's parameters can be done by using an iterative Baum-Welch Expectation-Maximization algorithm (EM) [107], assuming a predefined number of hidden states M.

Alternatively, to compare different model structures, such as models with a varying number of hidden states, or models with different variance-covariance matrices, a new algorithm needs to be considered. To achieve this aim, the likelihood ratio tests as proposed in [69] are adopted, which enable comparing a diagonal covariance matrix model with a full covariance matrix model, given a different number of hidden states.

3.2.2 Hidden Markov Models for Intra-daily Bitcoin Prices

In the context of adopting HMM using intra-daily Bitcoin prices, the same approach described in section 3.2.1 is followed. However, unlike the previous case where the analysis is conducted on daily closing prices, the data considered here are 2-minute ask and bid Bitcoin prices, aiming at exploring the descriptive behavior of the mid-market price within one exchange market. A more detailed explanation is illustrated in section 4.3.

Let y_t^i be the best price of each side i, where $i = \{ask, bid\}$ at time t, where $t = (1, 2, \ldots, T)$. We assume that the vectors Y^i are independent among each other, and each following a Markov process, specified by the joint probability distribution, independently across i:

$$P(S_{1:T}^i, Y_{1:T}^i) = P(S_1^i)P(Y_1^i|S_1^i)\prod_{t=2}^{T} P(S_t^i|S_{t-1}^i)P(S_t^i|Y_t^i) \tag{3.11}$$

Similarly, we assume that each distribution $P(Y_t|S_t)$ is a multivariate Gaussian.

3.3 Genetic Algorithm Optimized LSTM Networks

For the purpose of predicting Bitcoin mid-market price, we propose a hybrid of Hidden Markov Models and Long Short Term Memory networks optimized with Genetic Algorithms to fine-tune the network's parameters. Accordingly, section 3.3.1 showcases an overview on Genetic Algorithms while section 3.3.2 introduces Long Short Term Memory networks. Finally, section 3.3.3 illustrates the proposed model in detail.

3.3.1 Genetic Algorithms

Genetic Algorithms (GAs) are a type of optimization algorithms that are used to find the optimal solution(s) to a target problem [24], by mimicking the biological processes of evolution and natural selection. As GA are inspired from biological processes, terminologies such as chromosomes, populations, crossover and mutations are also adopted. Each potential solution to the optimization problem is represented by a chromosome, expressed in the form of binary strings [122]. Accordingly, GA start with initializing a population; where a chromosome is randomly selected for "fitness" evaluation. The fitness of a chromosome is calculated in accordance with a predefined fitness function; the target function to be optimized. Thus, the performance of each chromosome is evaluated and only chromosomes with excellent performance are selected for reproduction.

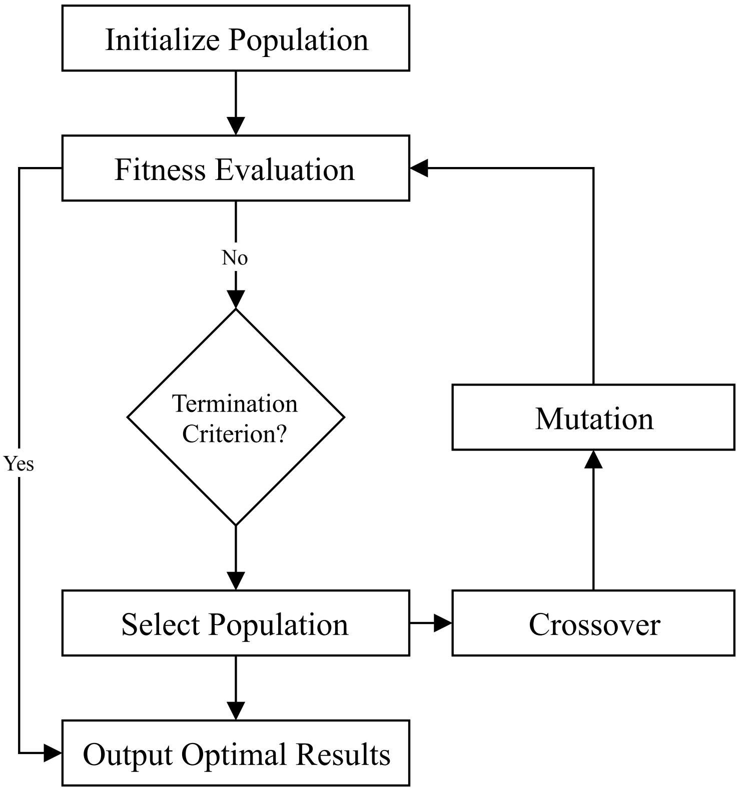


Figure 3.1: Basic Structure of Genetic Algorithms - adopted from [122].

Following the basic structure of GA as illustrated in fig. 3.1, once chromosomes are selected, new chromosomes are created from the selected ones during the crossover process. While in the mutation process, individual bits in the new chromosomes are randomly manipulated, by being swapped or turned off, to introduce diversity in chromosomes. Selection, crossover and mutations are repeated until a termination criterion is satisfied and the superior chromosomes with high performance are generated.

The motivation behind using GA is that they are powerful and more efficient than random search and exhaustive search algorithms [92] for optimization purposes. Moreover, GA do not require any information other than a solution representation and a fitness function with accordance to a given problem. This makes their applicability suitable for general problems, and especially appealing for our specific problem; that is optimizing the LSTM network parameters.

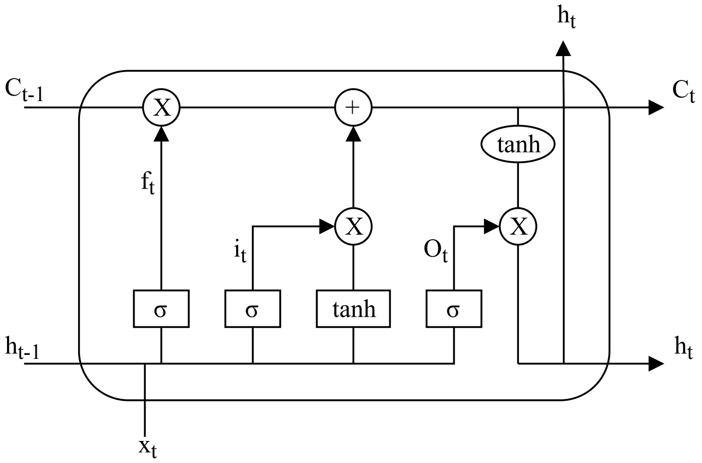


Figure 3.2: LSTM Cell Structure - adopted from [103].

3.3.2 Long Short Term Memory Networks

Long Short Term Memory networks are a special kind of Recurrent Neural Networks that were first introduced in 1997 [79]. They are specifically designed to overcome common problems in Recurrent Neural Networks, i.e., vanishing gradients, exploding gradients, and long-term dependencies, as they are able to remember information for more than 1000-time steps [29].

Fig. 3.2 illustrates the structure of an LSTM cell. The cell of an LSTM network has mainly three gates; input gate, forget gate and output gate. Using these gates, LSTM has the ability to remove or add information to the cell state. Each gate is composed of a sigmoid layer and a point-wise multiplication operation, which outputs a number between 0 and 1 that indicates how much information should be passed or thrown away [95].

LSTM starts by deciding which information is going to be deleted from the cell state $C_{\text{t-1}}$, also known as the memory, by considering the last output from the previous LSTM cell $h_{\text{t-1}}$, and the next input at current time x_{t} through the sigmoid function of the forget gate f_{t}, which in turn, outputs a number between 0 and 1 for each value on the cell state to either completely get rid of or keep, respectively. This process is defined by:

$$f_t = \sigma(W_f \cdot [h_{\text{t-1}}, x_t] + b_f) \tag{3.12}$$

where W_f represents the weights of the forget gate neurons and b_f represents the biases of the forget gate.

The next step is to define the information that is going to be added to the cell state by the input gate i_{t}, where a *tanh* layer generates new candidates $\tilde{C_t}$ to be added to the cell state using the input gate, represented by:

$$i_t = \sigma(W_i \cdot [h_{\text{t-1}}, x_t] + b_i) \tag{3.13}$$

where W_i represents the weights of the input gate neurons and b_f represents the biases of the input gate, and the new candidates $\tilde{C_t}$ are defined by equation 3.14 along with their respective weights and biases as follows:

$$\tilde{C_t} = tanh(W_C \cdot [h_{\text{t-1}}, x_t] + b_C) \tag{3.14}$$

Followed by updating the cell state from $C_{\text{t-1}}$ to C_{t}, taking into account the information that was thrown away previously by the forget gate and adding the new candidates that were decided by the input gate through:

$$C_t = f_t * C_{\text{t-1}} + i_t * \tilde{C_t} \tag{3.15}$$

Finally, the output gate O_{t} decides which information is going to be output, from the cell state, through the sigmoid layer as follows:

$$O_t = \sigma(W_o \cdot [h_{\text{t-1}}, x_t] + b_o) \tag{3.16}$$

Once that is decided, the *tanh* function of the output layer transforms the cell state values between -1 and 1, which are then multiplied by the resulting O_{t} in equation 3.16 to assure considering only the previously decided pieces of information, yielding the final output of the current LSTM cell h_{t}, illustrated as follows:

$$h_t = O_t * tanh(C_t) \tag{3.17}$$

3.3.3 Hybrid HMM and GA-Optimized LSTM Networks

With reference to Hidden Markov Models as explained in section 3.2.2, along with the Genetic Algorithms in section 3.3.1 and Long Short Term Memory networks in section 3.3.3, an innovative hybrid model is proposed and illustrated in fig 3.3.

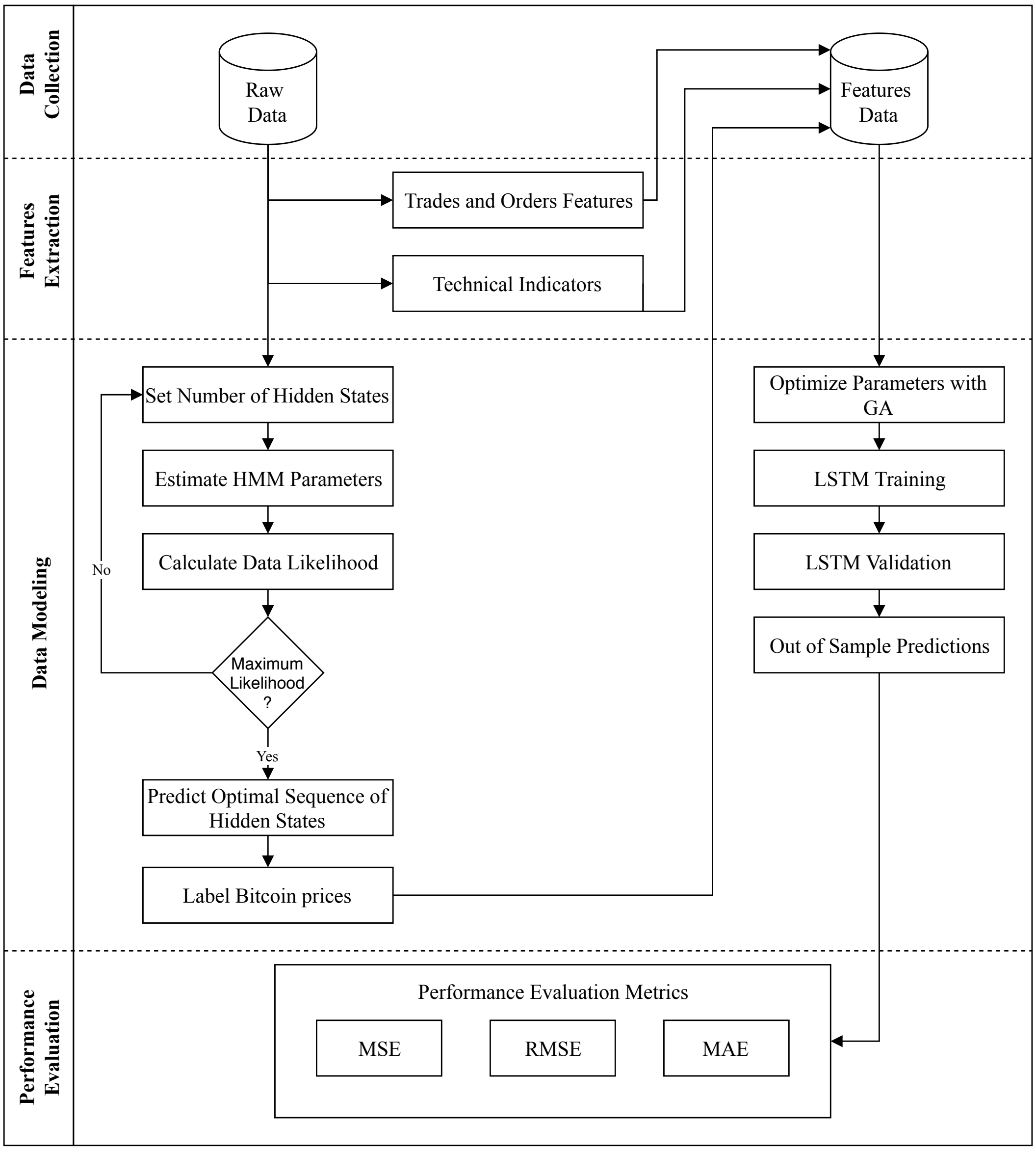


Figure 3.3: The Hybrid HMM and GA-Optimized LSTM Networks Model for Bitcoin Price Prediction.

The proposed model is composed of four main phases as follows:

- Data Collection: as illustrated in fig. 3.3, this phase contains two databases; raw data, which is represented by raw Bitcoin-related data and features data, which is represented by a combination between the previously mentioned raw data and further extracted related features from the following phases.
- Features Extraction: within this phase, raw data are processed and used to calculate a number of Bitcoin related features, that are believed to be beneficial to the prediction process in the following phases.
- Data Modeling: the proposed model is composed of two modeling approaches; descriptive through HMM and predictive through GA-optimized LSTM. As illustrated, the descriptive modeling follows the previously explained approach, in section 3.2.2, to create a new feature called “state” to better describe Bitcoin prices through insightful, yet hidden information that cannot be directly observed. Consequently, the “state” feature is then added to the features data, which in turn is considered in predictive modeling along with other related features contained in the database. Moreover, GA are exploited to fine-tune the LSTM network parameters, which is then used to predict Bitcoin prices.
- Performance Evaluation: once the predictions are generated following the proposed model, the performance is evaluated through several metrics, namely, Mean Squared Error, Root Mean Squared Error and Mean Absolute Error. The selected performance metrics are calculated as follows:

  1. Mean Squared Error (MSE):

  $$MSE = \frac{1}{N_{sample}} \sum_{t}^{T} (mmp(t) - \widehat{mmp(t)})^2 \tag{3.18}$$

  2. Root Mean Squared Error (RMSE):

  $$RMSE = \sqrt{\frac{1}{N_{sample}} \sum_{t}^{T} (mmp(t) - \widehat{mmp(t)})^2} \tag{3.19}$$

3. Mean Absolute Error (MAE):

$$MAE = \frac{1}{N_{sample}} \sum_{t}^{T} |mmp(t) - \widehat{mmp(t)}| \tag{3.20}$$

where N_{sample} is the size of the sample considered and mmp is the mid-market price of Bitcoin as well as the target variable to be predicted as explained in detail in section 4.3.2.

A more detailed explanation of the proposed model and its implementations is provided in section 4.3.

Chapter 4

Implementation

In this chapter, a detailed explanation of the contributions' implementation is provided, following the previously proposed models explained in Chapter 3.

This chapter is divided in three sections. Section 4.1 introduces the first contribution of this thesis within the research area of Bitcoin price dynamics, aiming at modeling the dependencies and interactions of Bitcoin prices within different exchange markets and possible relationships to classical assets, using a Network VAR model. Within the same area, section 4.2 tackles similar issues, from a different viewpoint, aiming at modeling such dependencies through the dynamics of their latent causes, attributed to time switches between different market regimes through Hidden Markov Models. Finally, section 4.3 proposes an innovative hybrid model for Bitcoin price prediction, using Hidden Markov Models and Genetic Algorithm-optimized LSTM networks.

4.1 A Network VAR Approach

For the purpose of modeling and explaining the evolution of Bitcoin prices, a novel Network Vector Autoregressive model is proposed. Specifically, this work extends previous researches by Brandvold et al. [18] and Corbet et al. [36] to acquire further empirical evidences on the non-conclusive properties of Bitcoin prices, namely, whether Bitcoin prices from different exchange markets are interconnected, and whether such interactions are affected by exogenous prices of traditional assets. Simply put, the objective of this work is to answer two research questions:

1. Is Bitcoin still an investment diversifier?

2. Are Bitcoin exchange markets efficiently integrated?

To address these questions, the proposed approach is based on a VAR model with an extension based on network models. The proposed model is believed to improve the performance of pure VAR models, given that network models introduce a contemporaneous contagion component, that can be exploited to describe the contagion effect between Bitcoin prices. Although the proposed model is mainly descriptive, it has been also extended to predict Bitcoin prices using the information contained in the multivariate interdependencies among Bitcoin exchange prices from one hand, and between Bitcoin prices and traditional assets prices from the other hand. This work has been published in [65]. [1]

Starting to explain the implementation approach, section 4.1.1 introduces the collected data, followed by section 4.1.2 which illustrates the empirical application of the proposed model along with the obtained descriptive results. Finally, section 4.1.3 shows the predictive results of the proposed model.

4.1.1 Data Collection

Given the purpose of this work, and without loss of generality, the chosen cryptocurrency to be addressed is Bitcoin due to its importance and popularity. Specifically, daily closing prices (USD) are considered.

As discussed, the first objective of this work is to assess whether Bitcoin prices in different exchange markets are correlated, thus, presenting endogenous price variations. Accordingly, a set of representative exchanges have been chosen whose price data is made available for a sufficiently long period of time. In particular, eight exchange markets of different geographic locations have been considered,

[1] It has been also cited in [35], [159], [155], [123], [67], [34], [52], [108], [28], [13], [87], [97], [152], [68], [6], [66], [51], [125] and [146].

representing 60% of the total daily volume trades. Table 4.1 demonstrates the considered exchanges along with their respective market shares, retrieved at the time (beginning of 2018) [48]. Bitcoin closing daily price data have been collected from each of these exchanges for a period of time from May 18th, 2016 to April 30th, 2018.

Table 4.1: Considered Exchange Markets by Daily Trading Volume.

Exchange Market	*Market Share*
Bitfinex	42%
Bitstamp	5%
Bittrex	0.5%
Coinbase	6%
Gemini	2%
HitBTC	3%
ItBit	1%
Kraken	0.5%

Moreover, in order to understand whether Bitcoin price variations can be explained by exogenous factors, daily data of the most important classical assets have been collected within the same period of time. Particularly considering Gold, Oil, S&P500 along with two exchange rates: USD/Yuan and USD/EUR. Similarly, closing daily prices have been considered. The collected data were obtained from Coinbase Pro API [32], previously known as GDAX, as well as from Bloomberg Terminal [15].

Unlike classical markets, cryptocurrency exchange markets are open 24 hours a day, 7 days a week. Taking this into account, the prices of classical assets during markets closure, are replaced with the last closing price at closing time. Accordingly, the prepared dataset is composed of 13 variables and 713 data points.

Having illustrated that, fig. 4.1 shows the evolution of Bitcoin prices within the considered period of time. Accordingly, the well-known rise of Bitcoin prices in 2017 is evident, where prices have increased from a minimum of $430 to a maximum of $20,000, followed by a high volatility in 2018. Slight differences between prices within different exchanges are noted, as the lines are not perfectly aligned. To better understand that, table 4.2 presents several summary statistics based on the collected dataset.

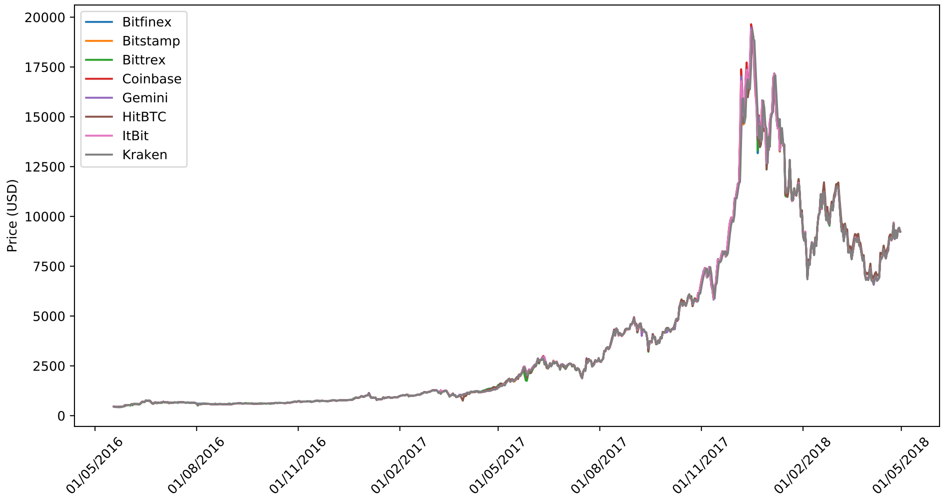


Figure 4.1: Time Series Plot of Bitcoin Prices for the Considered Period of Time.

Table 4.2: Summary Statistics for Closing Prices for the Considered Markets.

Market	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
Bitfinex	3899.5	4274.64	435.61	19187.12
Bitstamp	3899.04	4286.02	439.62	19187.78
Bittrex	3893.83	4269.86	421.11	19261.10
Coinbase	3919.05	4318.98	438.38	19650.01
Gemini	3910.38	4306.36	437.57	19475.90
HitBTC	3916.19	4297.17	436.36	19095.30
ItBit	3907.13	4300.32	438.61	19357.97
Kraken	3890.18	4272.55	433.50	19356.91
Gold	1275.57	52.34	1128.42	1366.38
Oil	48.67	3.16	39.51	54.45
S&P500	2414.78	212.308	2000.54	2872.87
USDEUR	0.88	0.04	0.80	0.96
USDYuan	6.67	0.19	6.26	6.96

Table 4.2 confirms the slight differences in Bitcoin prices for the considered exchanges. The mean and standard deviations are somewhat different, so are the maximum value statistics. Compared to classical assets, namely, Gold and Oil, Bitcoin volatility is about 80 times and 1400 times higher, respectively. Similarly, compared to S&P500, Bitcoin volatility is about 20 times higher. Moreover, it is evident that the considered exchange rates are much less volatile than Bitcoin prices.

This is in accordance with the results available in [36].

4.1.2 Descriptive Results

Once the data have been collected and finalized, the model proposed in section 3.1 was implemented using the R programming language [131]. Accordingly, this section presents the descriptive empirical findings based on the implementation of a network VAR model on the prepared dataset.

Starting from fig 4.2, the correlation between the closing prices for the considered markets is calculated and illustrated in the form of a heatmap. Positive correlations are represented in shades of blue while negative correlations are represented in shades of red, where darker colors depict higher correlations in absolute values.

On one hand, fig. 4.2 shows that the correlations between different exchanges are quite high, revealing that markets are highly correlated and synchronized, thus, resulting in a strong endogenous source of price variations. On the other hand, the correlations between Bitcoin prices and real assets, namely, Gold and Oil, are low. Such results are in line with the results presented in [36], considering Bitcoin as a potential diversification asset. Moreover, the figure shows positive correlations with S&P500 as well. However, unlike the results reported in [36], the correlation is negative with the considered exchange rates.

Prior to coming to a conclusion based on the reported results, it is believed that correlation should be “netted” from spurious effects. Indeed, it is well-known that pairwise correlation may be inflated by correlations that may arise from a common relationship with a third variable. To tackle this, partial correlations solve such issues by calculating the correlation between the residuals from a Linear Regression model of each of the two variables with all the remaining ones, thus, measuring the “net” or “additional” correlation.

Consequently, fig. 4.3 presents a heatmap illustrating all pairwise partial correlations between the closing prices for the considered markets. Additionally, for the sake of a clear interpretation, insignificant correlations whose p-value is greater than

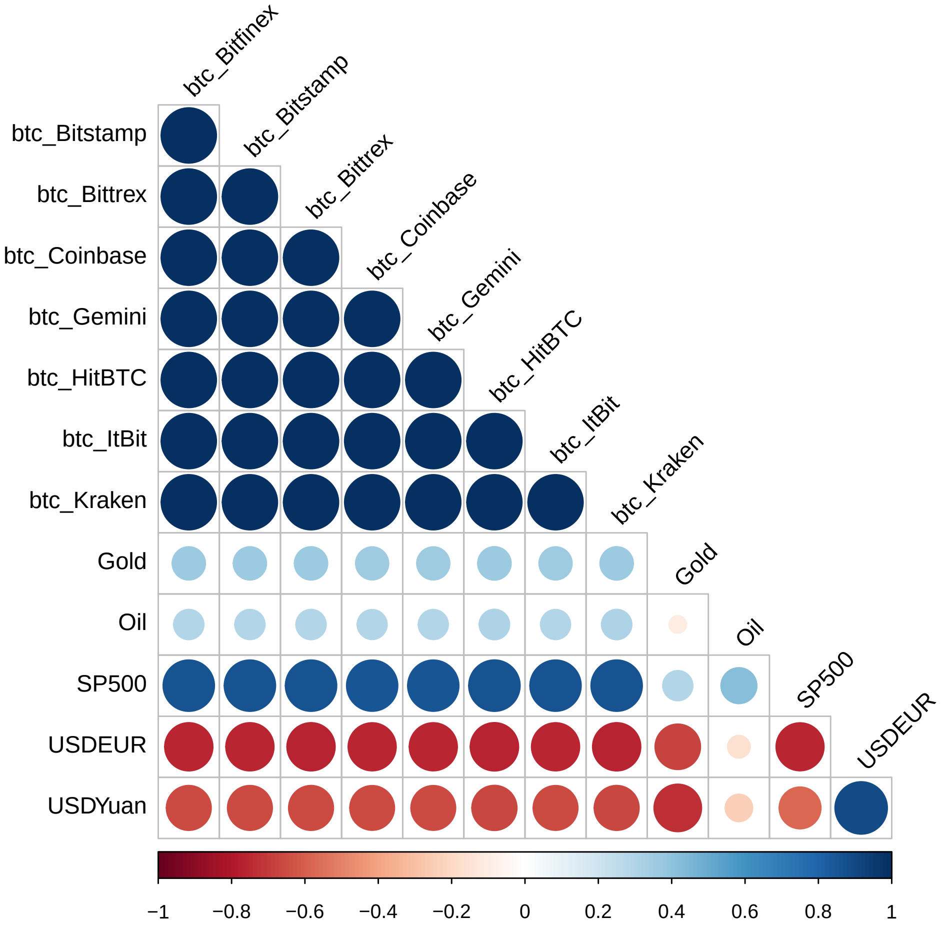


Figure 4.2: Correlation Matrix between Closing Prices for the Considered Markets.

0.05 are marked with a cross.

Looking at fig. 4.3, many correlations are deemed to be insignificant at the 5% level. In particular, considering the top of the heatmap, most of the correlations of the large exchange markets are indeed significant, namely, Bitfinex, Bitstamp and Coinbase. On the contrary, insignificant correlations are noted for the smallest exchanges, namely, Bittrex, ItBit and Kraken. Such results indicate that large markets are the most important endogenous driver of prices, confirming the economic intuition that larger trading volumes make the price, in accordance with [18].

Regarding the exogenous causes of Bitcoin price variations, Oil is noted to

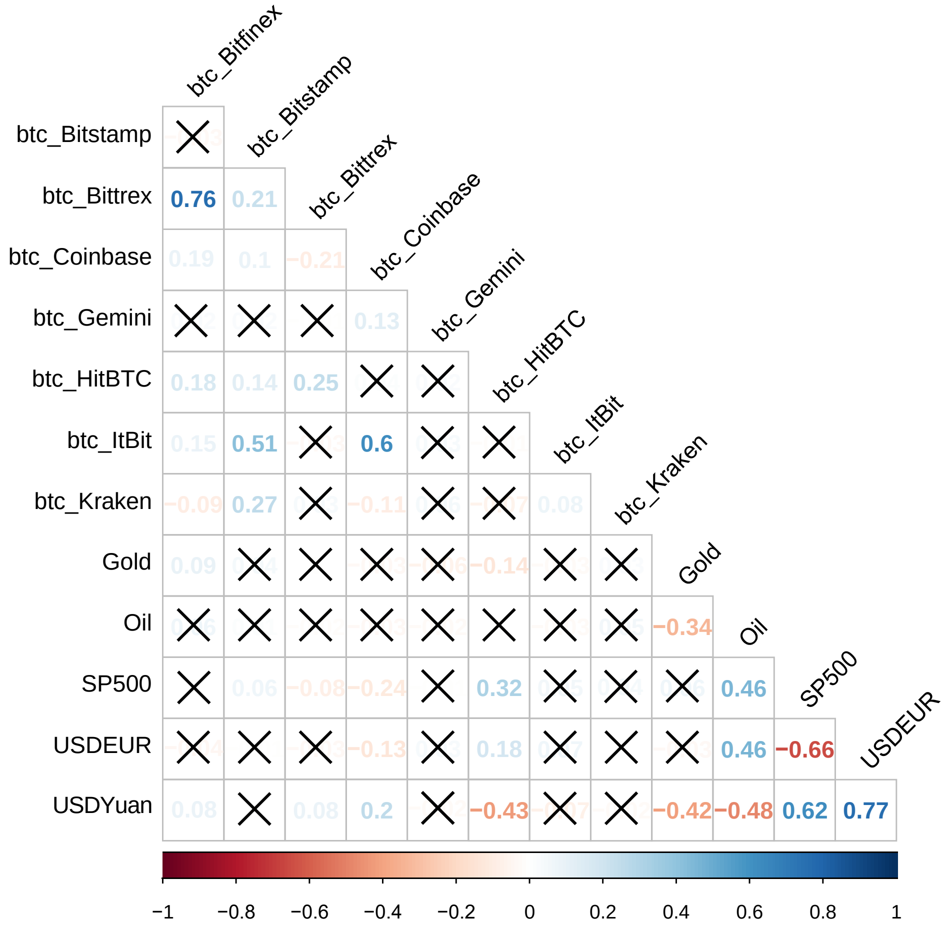


Figure 4.3: Partial Correlation Matrix between Closing Prices for the Considered Markets.

have insignificant partial correlation with Bitcoin prices, while Gold is significantly correlated with Bitcoin prices from only two exchanges out of the eight considered, and of a very low magnitude. This, indeed, confirms the presence of a very low correlation between Bitcoin prices and real asset prices. As for the partial correlation between Bitcoin prices and financial assets, fig. 4.3 shows that the partial correlation is either insignificant or very weak, with a noticeable exception for HitBTC exchange market.

Therefore, once the correlations are netted from spurious effects, their nature of being potential diversifiers with respect to classical assets is confirmed, in accordance with the finding of [36], with the exception of specific exchange markets, such as HitBTC, which may be presumably affected by the behavior of local traders.

In addition to its ability in clarifying “actual” correlations, partial correlation is able to help in describing the multivariate patterns of possible relationships between the considered prices, by means of a graphical network model [156, 101]. Such a model can be obtained by associating each asset price with a node in a graph, followed by drawing a link between two nodes, if and only if, the corresponding partial correlation is significantly different from zero. Accordingly, fig. 4.4 illustrates the graphical network model based on the considered data.

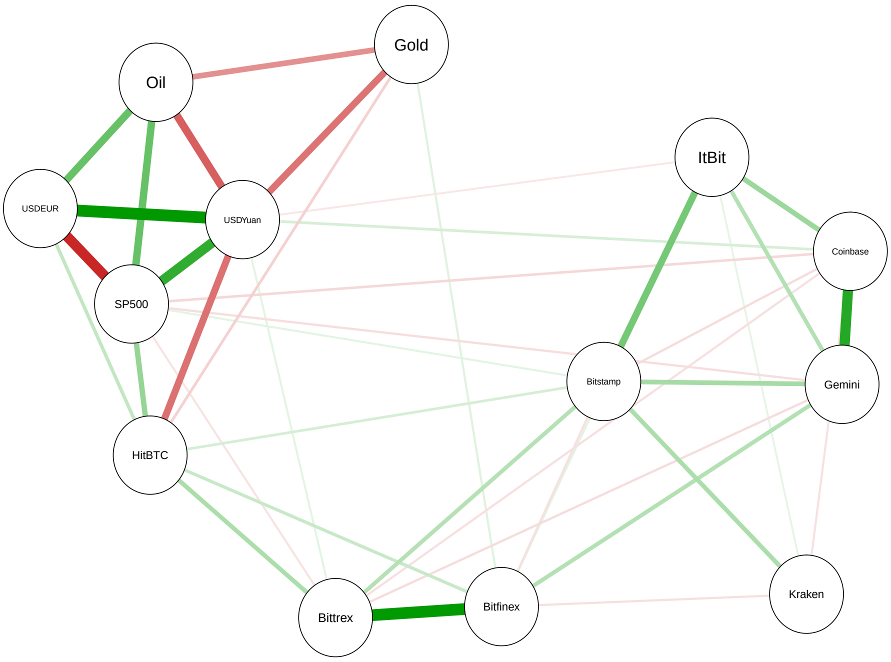


Figure 4.4: Graphical Correlation Network Model between Closing Prices for the Considered Markets.

Fig. 4.4 graphically confirms the previous discussed findings. In particular, it shows that Bitcoin prices and other classical assets prices form two rather distinct clusters of connections, which are highly interconnected inside. Moreover, the high centrality of the larger exchange markets, i.e., Bitfinex and Bitstamp, confirms their nature of driving Bitcoin prices in other exchanges as price setters. Additionally, a link between the two clusters is noted through HitBTC exchange market, which is affected by both classical and other Bitcoin markets. The found behavior of HitBTC may be due to the peculiar nature of investors, evidently acting simultaneously on both markets.

4.1.3 Predictive Results

Although the proposed model is mainly descriptive, it is further extended to predict Bitcoin daily closing prices, leveraging both endogenous and exogenous sources of correlations.

With respect to the previously calculated partial correlations, the B_0 matrix can be derived, along with the autoregressive parameters $A_1, \ldots, A_p$. Thus, the time dependent price of each asset i can be disentangled by separately estimating the autoregressive idiosyncratic component and the contemporaneous component, according to equation 3.2. Accordingly, table 4.3 shows the results of such an estimation.

Table 4.3: Comparison between the Estimation Components of Prices Obtained with the Proposed Model.

Market	*Contemporaneous Component*	*Autoregressive Component*
Bitfinex	452.31	0.74
Bitstamp	450.35	0.19
Bittrex	452.15	0.69
Coinbase	453.55	0.68
Gemini	450.01	0.20
HitBTC	451.98	0.37
ItBit	451.01	0.19
Kraken	451.02	0.28
Gold	97.51	0.99
Oil	4.03	0.95
S&P500	197.26	0.97
USDEUR	0.06	0.96
USDYuan	0.51	0.99

Looking at table 4.3, it is noted that the autoregressive components prevails only for the considered exchange rates, much more stable variables and less connected with other prices, compared to other assets. In all other cases, and in particular, for Bitcoin prices, the multivariate contemporaneous component prevails by far. Meaning that, most prices are driven by correlations with contemporaneous prices, rather than by their past behavior. Thus, the interconnectedness nature between different exchange markets can be necessary to estimate Bitcoin prices correctly.

In order to understand whether the proposed network VAR model is able to well-predict Bitcoin prices, the predictive performance needs to be assessed. Particularly, to understand and evaluate whether the introduction of the contemporaneous component improves the predictive accuracy compared to a pure autoregressive model.

Bearing that in mind, the proposed model is implemented using the prepared dataset, apart from the last fifty days, about 10% of the overall data. Accordingly, Bitcoin prices are predicted for the excluded days, with one day ahead rolling predictions, and then compared to the actual prices. Similarly, the obtained predictions are compared to another set of predictions obtained by the implementation of a pure autoregressive model, excluding the contemporaneous effect. In both cases, to make the model more realistic and useful in practice, the contemporaneous prices at time t have been replaced with those at $t-1$. Finally, to evaluate the predictive performance, RMSE of the predictions with respect to the actual values, is calculated for both models, as reported in table 4.4.

Table 4.4 reports the RMSE for the proposed model, a full structural VAR with a contemporaneous component compared to a pure autoregressive component. Accordingly, the RMSE for Bitcoin price predictions for the exchange markets, under the proposed model, average at about 11% of the mean prices, about 100 times higher than the corresponding values of Oil prices, about 50 times higher for S&P500 and USDYuan, and finally, about 20 times higher for Gold and USDEUR. The reported results are, indeed, in accordance with the economic intuition that it is more difficult to predict more volatile assets.

Comparing the reported RMSEs in table 4.4, it is suggested that less central, more

Table 4.4: Comparison between the RMSE of a Full Structural VAR Model and a Pure Autoregressive Model.

Market	*Full Structural VAR*	*Pure Autoregressive*
Bitfinex	267.37	293.49
Bitstamp	379.71	397.25
Bittrex	290.84	305.51
Coinbase	550.10	579.98
Gemini	792.22	786.58
HitBTC	288.63	342.50
ItBit	331.62	455.45
Kraken	718.51	676.35
Gold	6.74	7.19
Oil	0.55	0.58
S&P500	5.90	6.35
USDEUR	0.002	0.003
USDYuan	0.02	0.02

remote exchanges as illustrated in fig. 4.4, such as Gemini, Kraken and Coinbase, are the most difficult to predict. The reason behind this condition is that such exchanges are less connected to observed market prices and, presumably, more dependent on external perturbations related to changes in the regulatory environment, or simply, to changes in sentiment between cryptocurrency investors. Indeed, the minimum prediction errors are found in HitBTC, a highly connected market to many nodes, and in Bitfinex, the leading trading exchange and the strongest price setter.

Generally, table 4.4 shows that the proposed model outperforms a pure autoregressive model, thus, justifying its additional complexity with an increase in its predictive power. Of course, with the exception of the least two central exchanges: Gemini and Kraken. This further suggests that the proposed model better-predicts Bitcoin prices for exchange markets that are more interconnected to one another, and/or to classical markets.

4.2 A Hidden Markov Model for Regime Changes

Following the same line of research, this work aims at modeling and explaining the evolution of Bitcoin prices using Hidden Markov Models. However, while the previously implemented model, in section 4.1, focuses on modeling the dependencies between the observed exchange markets directly, the implemented model in this section addresses the same dependencies through the dynamics of their latent causes, attributed to time switches between different market regimes, going from “bull” to “stable” and “bear”.

Alternatively stated, taking into account the multivariate nature of cryptocurrency prices, a Hidden Markov Model is proposed to explain the time evolution of Bitcoin prices, through the evolution of hidden unobserved states, which can be referred to a different equilibrium of the cryptocurrency economy, in accordance with [44]. Doing so, the observed dependencies between prices from different exchange markets, found in section 4.1, may be fully explained. This will be the case when Bitcoin prices from different exchanges become independent (described by a diagonal covariance matrix), conditionally on the latent state, rather than still interdependent (described by a full covariance matrix). This work has been submitted for publishing in [64].

Accordingly, this section addresses the implementation of the proposed model in section 3.2, specifically in 3.2.1. The contributions of this work are two-fold: providing a further understanding of Bitcoin price dynamics from an econometric point of view and implementing an easy-to-use likelihood ratio test [69] for comparing differently implemented Hidden Markov Models. Similarly, although the proposed model is mainly descriptive, it has been extended to predict Bitcoin prices, given the available data, to assess its performance from a predictive point of view.

Having illustrated that, section 4.2.1 presents the collected data, section 4.2.2 explains the obtained descriptive results, and finally, section 4.2.3 evaluates the predictive performance of the proposed model.

4.2.1 Data Collection

Given the purpose of this work, and without loss of generality, the chosen cryptocurrency to be addressed is Bitcoin. Specifically, daily closing prices (USD) are considered.

Accordingly, to better understand the endogenous variations in Bitcoin prices within different exchange markets, seven representative exchanges have been chosen. Namely, Bitfinex, Bitstamp, Bittrex, Coinbase, Gemini, ItBit and Kraken, representing 40% of the total daily volume trades, illustrated in table 4.5.

Table 4.5: Considered Exchange Markets by Daily Trading Volume.

Exchange Market	*Market Share*
Bitfinex	13%
Bitstamp	1%
Bittrex	1%
Coinbase	13%
Gemini	0.31%
ItBit	1%
Kraken	10%

Accordingly, table 4.5 demonstrates the considered exchanges along with their respective market shares, retrieved on November 21st, 2018 from [47]. For each exchange market, Bitcoin closing daily prices data have been collected for a period of time from December 12th, 2015 to October 25th, 2018. The collected data were obtained from CryptoCompare API [37] by implementing a Python script. Thus, the prepared dataset is composed of 7 variables and 1029 data points.

With that said, fig. 4.5 illustrates the evolution of Bitcoin prices within the considered period of time.

As previously explained in 4.1.1, it is obvious that Bitcoin prices in different exchange markets are highly correlated, but not perfectly aligned. Moreover, at a first glance, fig. 4.5 shows that the considered prices went through different equilibrium states: prices were almost “stable” at the start of the considered period of time, when the well-known rise took place in 2017. This stability then was followed by a fluctuation in prices in 2018, a period of high volatility.

Table 4.6 confirms the divergence of Bitcoin prices within the different exchanges considered. Not only in volatility and the maximum value statistics, but also in their means, on the contrary of the economic law “one asset, one price”.

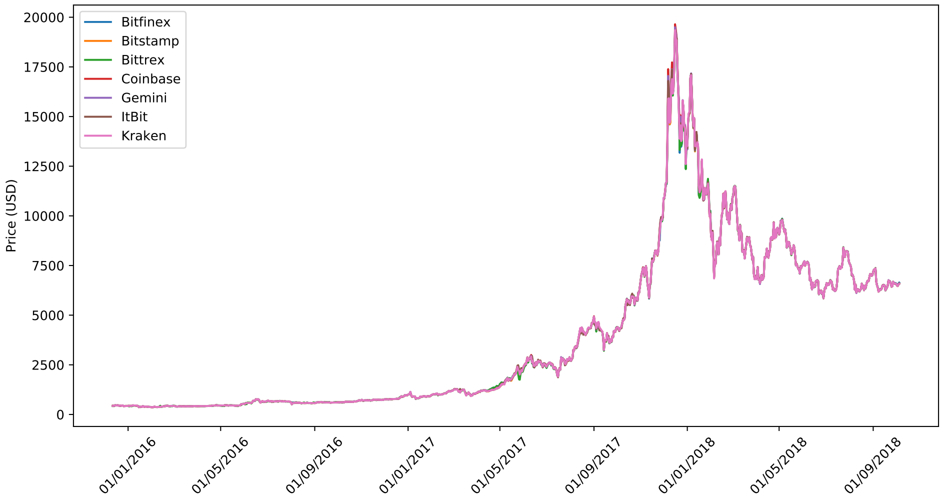


Figure 4.5: Time Series Plot of Bitcoin Prices for the Considered Period of Time.

Table 4.6: Summary Statistics for Closing Prices for the Considered Markets.

Market	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
Bitfinex	3857.53	4025.72	367.01	19210.00
Bitstamp	3859.71	4035.09	367.64	19187.78
Bittrex	3853.65	4023.48	365.00	19261.10
Coinbase	3871.30	4059.16	367.00	19650.00
Gemini	3866.98	4050.82	368.70	19499.99
ItBit	3863.09	4045.71	360.40	19357.97
Kraken	3859.56	4031.61	368.00	19356.90

Thus, this work is based on the hypothesis that these variations in Bitcoin prices can be explained by the endogenous relationships between different exchange markets, which are in turn explained by different latent states of the cryptocurrency economy.

4.2.2 Descriptive Results

Given the prepared dataset, the model proposed in section 3.2, specifically in 3.2.1, is implemented using Python programming language [130]. For the implementation of the proposed model, three alternative types of hidden structures are considered, characterized by two, three and four hidden states and conventionally labeled with $(0, 1, 2, 3)$. Moreover, two different types of variance-covariance matrices are taken into account: a full matrix and a more parsimonious diagonal matrix. Thus, producing a total of 3×2 alternative models. Unless otherwise stated, the presented results are related to a three-states model, implemented twice: with a full and a diagonal variance-covariance matrix.

Having illustrated that, fig. 4.6 demonstrates the distribution of Bitcoin prices within the three hidden states, estimated with a full covariance matrix.

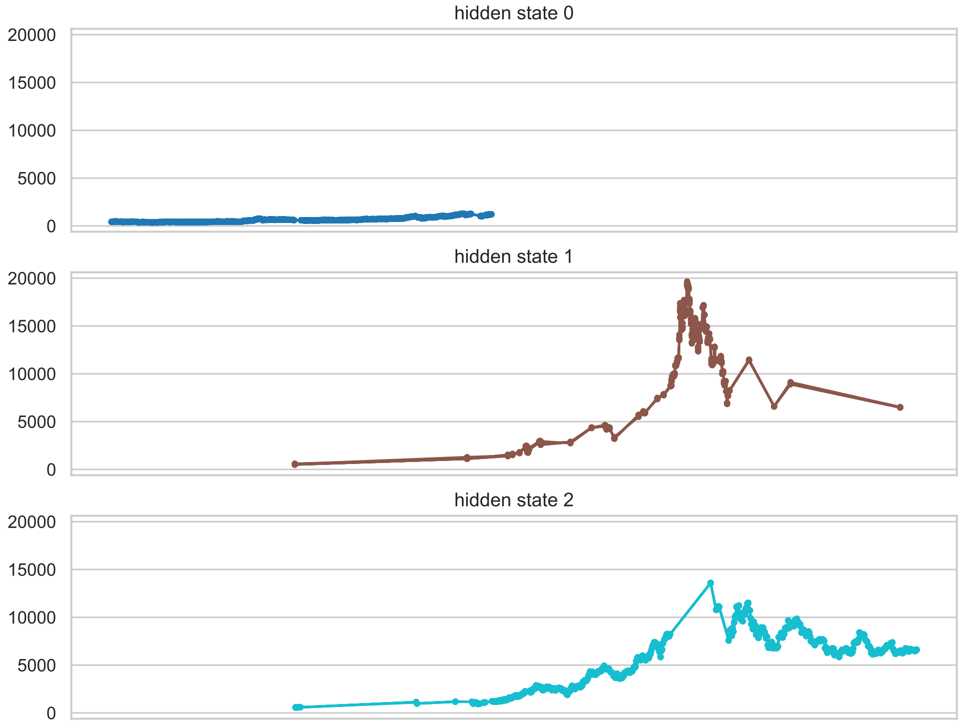


Figure 4.6: Bitcoin Prices from Bitstamp Exchange Market Plotted per Hidden State with a Full Covariance Matrix.

For a clearer understanding, fig. 4.7 presents the time evolution of the three estimated hidden states of Bitcoin prices, considering a full covariance matrix. pink data points correspond to Bitcoin prices with *hiddenstate*0, overlapped and, thus, creating the white line. Moreover, light purple data points correspond to Bitcoin prices with *hiddenstate*1 and dark purple data points correspond to Bitcoin prices with *hiddenstate*2.

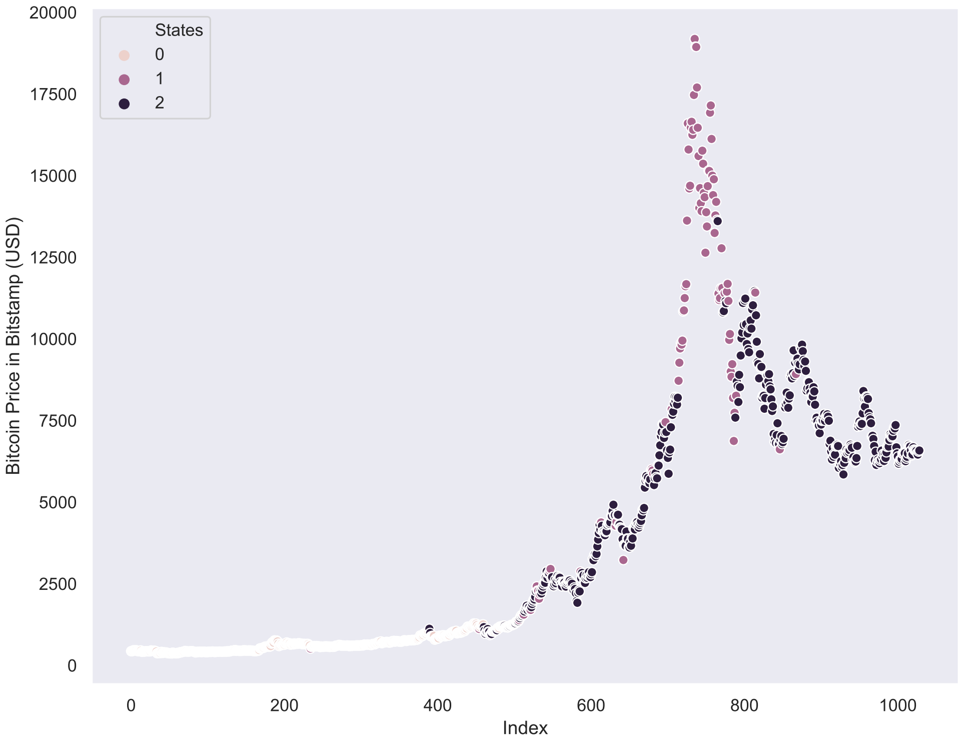


Figure 4.7: Time Series Plot of Bitcoin Prices from Bitstamp Exchange Market Clustered by Hidden States with a Full Covariance Matrix.

Accordingly, it is noted that $hiddenstate0$ is concentrated in the initial time period, when Bitcoin was relatively new and rarely increasing in price, while $hiddenstate1$ and $hiddenstate2$ alternate, between lower and higher prices, in the more recent time period. Moreover, taking a closer look at the first 450 data points, depicting mainly $hiddenstate0$, few data points are marked with $hiddenstate1$ and $hiddenstate2$, as illustrated in fig. 4.8.

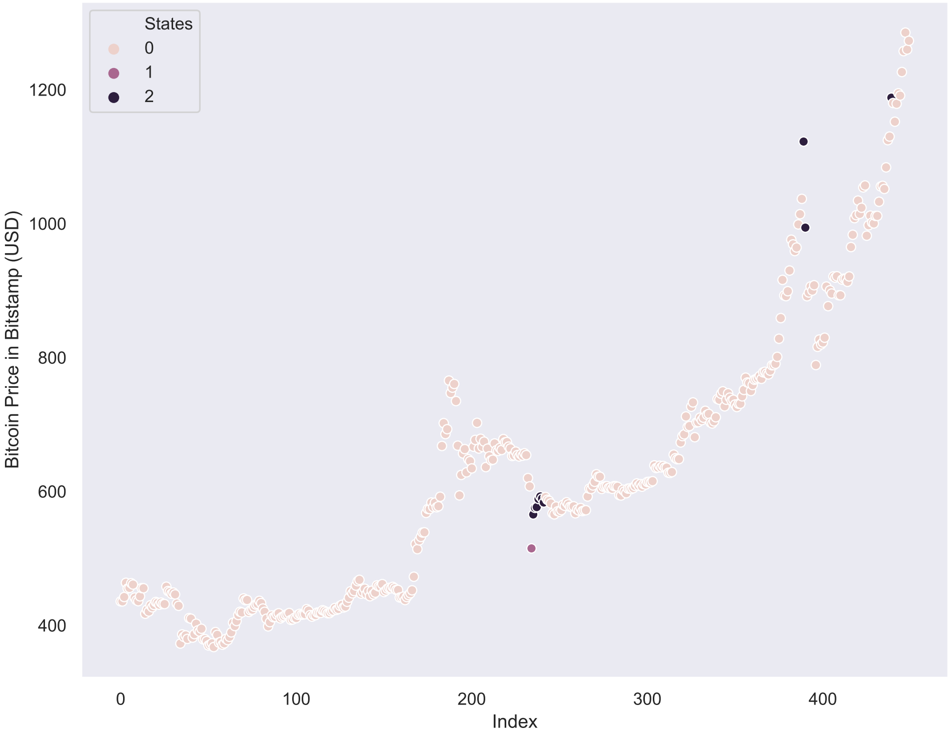


Figure 4.8: First 450 Bitcoin Prices from Bitstamp Exchange Market Clustered by Hidden States with a Full Covariance Matrix.

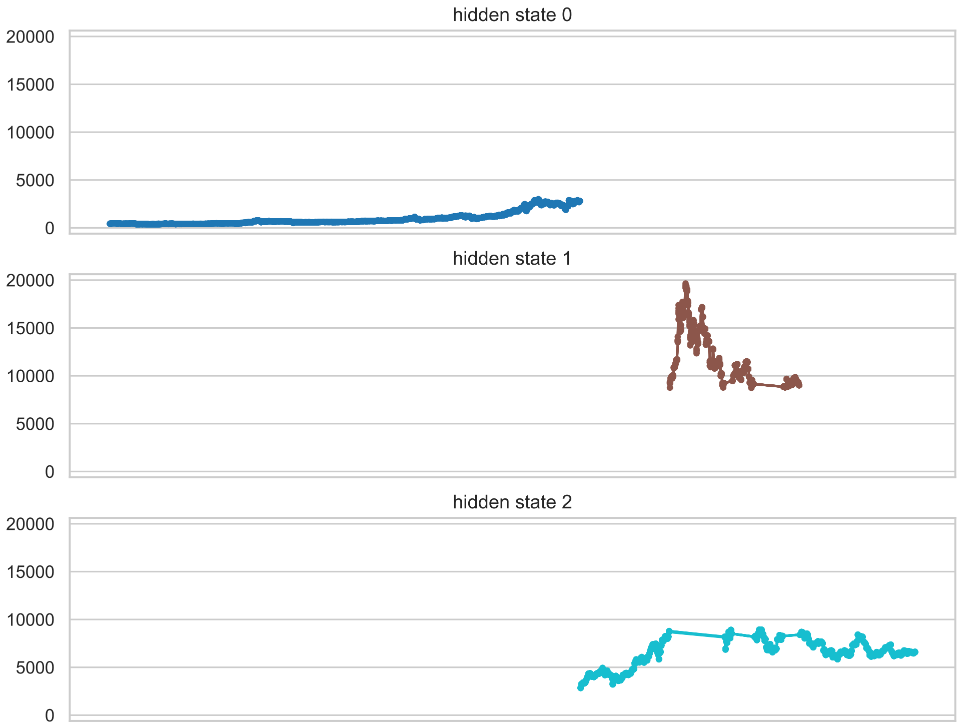


Figure 4.9: Bitcoin Prices from Bitstamp Exchange Market Plotted per Hidden State with a Diagonal Covariance Matrix.

Pointing out that a full covariance matrix involves a model specification that may be too complex to fit the data in a good way, a more parsimonious model, characterized by a diagonal covariance matrix is considered. Such a model implies that Bitcoin prices of any exchange market, conditionally on the hidden state, are independent from prices related to other exchange markets, at any time point. Simply put, this model assumes that the dynamics of Bitcoin prices is fully explained by the dynamics of the hidden states, whereas the variations in price coming from different exchanges are insignificant.

Similarly, fig. 4.9 demonstrates the distribution of Bitcoin prices within the three hidden states, estimated with a full diagonal matrix, while fig. 4.10 presents the time evolution of the three estimated hidden states, considering a diagonal covariance matrix.

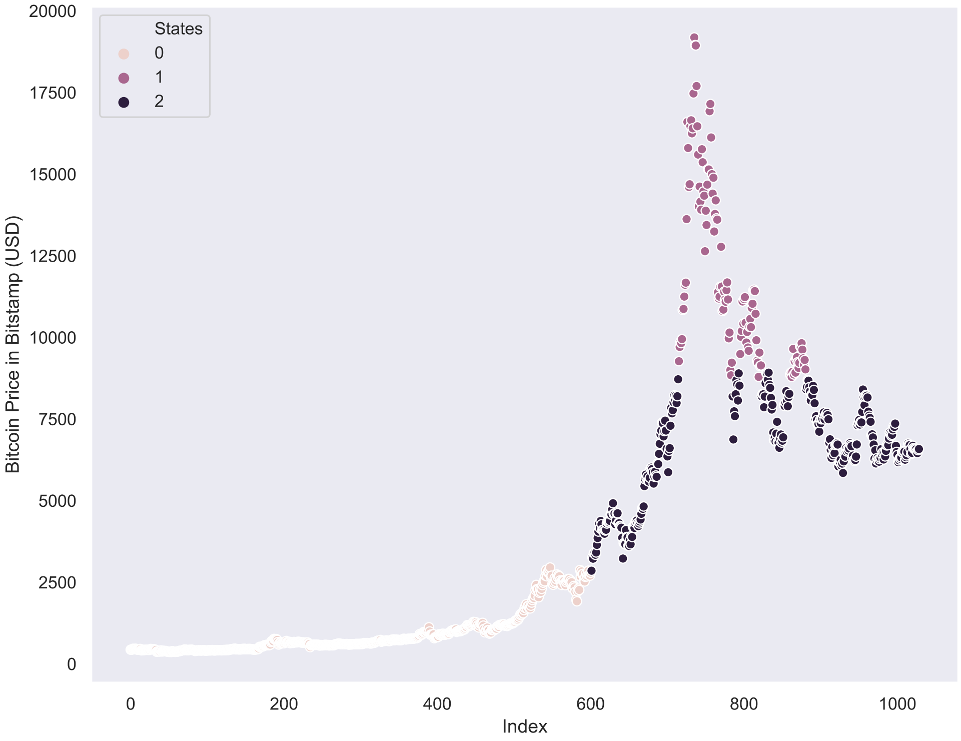


Figure 4.10: Time Series Plot of Bitcoin Prices from Bitstamp Exchange Market Clustered by Hidden States with a Diagonal Covariance Matrix.

From fig. 4.10, it is noted that $hiddenstate0$ is still concentrated in the initial time period. However, differently from what happens with the full covariance matrix, $hiddenstate1$ seems to be mostly concentrated in the second period of time, a period in which Bitcoin prices were steadily rising, while $hiddenstate2$ is spreading but mainly concentrated in the latest period of time. Moreover, neither of $hiddenstate1$ nor $hiddenstate2$ are present in the initial period of time, illustrated in fig. 4.11.

Comparing the previously illustrated figures, it seems that a model with a diagonal covariance matrix, conditionally on three hidden states, provides a better description of "regime switches" implied by the data. To confirm this result, from a more statistical point of view, the likelihood ratio test statistics are computed and tested as in [69], to compare a full covariance matrix model with a diagonal covariance matrix model, given a predefined number of hidden states.

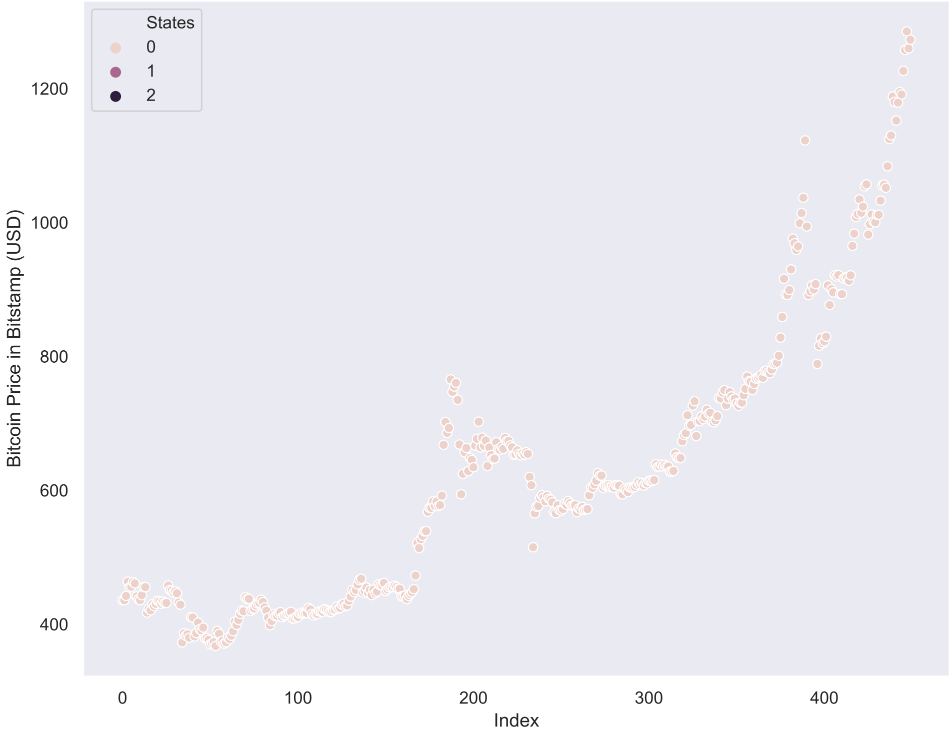


Figure 4.11: First 450 Bitcoin Prices from Bitstamp Exchange Market Clustered by Hidden States with a Diagonal Covariance Matrix.

Given the implemented models within this work, likelihood ratio test statistics are computed given 2, 3 and 4 hidden states. Accordingly, table 4.7 reports the likelihood ratio test for the two considered covariance structures, namely, full and diagonal. It clearly shows that a more parsimonious matrix model is always preferred compared to a full covariance matrix model, for any given number of hidden states.

Table 4.7: Likelihood Ratio Test - Full vs Diagonal Covariance Matrix.

Number of States	*Likelihood Ratio*	*p-value*
2	51493.81844	$3.12e-51$
3	56931.07517	$2.34e-54$
4	50226.37607	$7.12e-49$

These results strongly support the hypothesis of this work, i.e., that market price and correlation dynamics are fully explained by the dynamics of the unobserved, latent states. Furthermore, to assess which state configuration is more supported by the available data, the likelihood ratio test is computed for a diagonal covariance matrix model, given a different number of hidden states, as reported in table 4.8.

Table 4.8: Likelihood Ratio Test - Diagonal Covariance Matrix.

Number of States	*Likelihood Ratio*	*p-value*
3 vs 2	154.38165	$4.87e - 30$
3 vs 4	8440.61863	$1.13e - 41$

From table 4.8, it is noted that the model constructed with three hidden states has a likelihood that is significantly higher than that of a simpler model with two states, and than that of a more complex model with four states. Thus, the model which receives the highest empirical likelihood from the observed data, and should therefore be selected, is the model constructed with a diagonal covariance matrix and three hidden states.

From an economic viewpoint, this implies that the relatively young history of Bitcoin prices can be explained using three alternative states, namely, “bull”, “stable” and “bear” markets.

4.2.3 Predictive Results

While the proposed model is mainly descriptive, it is further extended to predict Bitcoin daily closing prices in order to evaluate its predictive performance, given the prepared dataset.

Accordingly, the proposed model is implemented using 80% of the data, excluding the remaining 20% to further use it as a test dataset, employing one day ahead rolling predictions with the actual prices. Consequently, table 4.9 reports the RMSE of the predictions of two models: a three hidden states model with a full covariance matrix and a three hidden states model with a diagonal covariance matrix.

At a first glance, table 4.9 shows that the overall predictive performance of the proposed model is similar for the considered exchange markets. However, a deeper look shows that differences emerge for different exchanges.

Table 4.9: Comparison between the RMSE of a 3-State HMM with a Full Covariance Matrix and a 3-State HMM with a Diagonal Covariance Matrix.

Market	*3-State Full HMM*	*3-State Diagonal HMM*
Bitfinex	4034.640	1738.244
Bitstamp	4026.765	1741.533
Bittrex	4143.270	1739.667
Coinbase	3999.867	1738.681
Gemini	4009.074	1738.767
itBit	3950.937	1737.396
Kraken	4039.409	1736.894

The results suggest that the trading volume per exchange market, expressed in market share and previously reported in table 4.5, may affect the prediction process using the proposed model. On one hand, considering a full covariance matrix, Bitcoin prices coming from Bittrex are the most difficult to predict, presumably due to the fact that it has a lower market share of 1%, while prices coming from Coinbase are predicted more accurately given its higher market share of 13% at the time. On the other hand, while considering a diagonal covariance matrix, the suggestion is more consistent. The highest RMSE was reported for Bitstamp, given its lower market share of 1%, and the most accurate predictions were reported for Kraken exchange market with a market share of 10%.

Furthermore, and as expected, the RMSEs of all the exchanges using a diagonal covariance matrix are always lower than those obtained with a full covariance matrix. Accordingly, a diagonal variance-covariance structure is preferable, in line with the results obtained using the likelihood ratio test statistics.

Lastly, in order to compare the predictive performance of the model, given a different number of hidden states, the same approach is followed to implement a 2-state full HMM, a 2-state diagonal HMM, a 4-state full HMM and a 4-state diagonal HMM. Accordingly, the results show that the predictive performance for the 4-state diagonal HMM is significantly higher compared to the 4-state full HMM. However, the 2-state full HMM slightly outperforms the 2-state diagonal HMM. This is consistent with the fact that a model with only two hidden states is very parsimonious and, therefore, the need to have a full covariance matrix emerges, as the number of hidden states is too low to explain the price dynamics.

4.3 A Hybrid HMM and GA-Optimized LSTM Networks

While the previously proposed and implemented models were mainly descriptive, following the research area of Bitcoin price dynamics and aiming at modeling such prices and understanding their evolution from different points of view, the implemented model in this section is mainly predictive and falling within the research area of Bitcoin price prediction. The main objective of this work is to develop an innovative and effective model that is able to predict Bitcoin prices, while achieving more accurate prediction results than those available in the literature.

Given the purpose of this work, a hybrid Hidden Markov Model and Genetic Algorithm-optimized LSTM Networks is proposed in section 3.3. Particularly, the proposed model addresses Bitcoin prices from both a descriptive point of view, by extending Hidden Markov Models as previously explained in 3.2.2, and a predictive point of view through Genetic Algorithm-optimized LSTM Networks. This work has been published in [5].

Consequently, this section presents the implementation of the proposed model. Following the four phases of the proposed models illustrated in fig. 3.3, section 4.3.1 introduces the collected data corresponding to the first phase, Data Collection. Followed by section 4.3.2 explaining the extracted features that are believed to be beneficial to the prediction process corresponding to the second phase, Features Extraction. Furthermore, sections 4.3.3 and 4.3.4 illustrate both the descriptive and predictive modeling and results, respectively, corresponding to the third phase, Data Modeling. Finally, the fourth and last phase, Performance Evaluation, is presented along the predictive results in section 4.3.4.

4.3.1 Data Collection

To prepare the dataset, Bitcoin data has been collected from Coinbase exchange market, one of the biggest platforms with a trading volume of 63 million USD per daily trading [33]. Coinbase Pro Public API [32], previously named GDAX, was used to collect real-time updates. Accordingly, all the available public data has been collected since January 2018. However, for the scope of this work, the analysis will be carried out on a subset of the period, namely, from August 20th, 2018 to September 20th, 2018 with a data frequency of 2 minutes, in USD.

Although all the available public data have been collected, and given the purpose of this work, the main focus for preparing the dataset is collecting:

- Market Orders Data: requests to buy or sell a specified amount of an assets, Bitcoin, at the best possible price, which includes ask/bid price and ask/bid amount.

- Market Trades Data: which includes buy/sell price and buy/sell amount. The i^{th} sample includes the price p_i and the amount of the traded asset, also known as volume, v_i. A trade occurs when two orders at the opposite side, buy and sell, match. Moreover, a trade can be either a perfect fill, meaning that both the price and the volume coincide, or a partial fill, meaning that only the prices are matching.

4.3.2 Features Extraction

The raw data that have been collected provides valuable information about Bitcoin orders and trades within the considered market. However, a further step is needed to extract information that cannot be directly revealed by modeling such data alone. Thus, the raw collected data is further investigated, and several features have been extracted that are believed to be beneficial to the prediction process.

The set of extracted features is divided into two categories: Orders and Trades and Technical Indicators, which are explained as follows.

Orders and Trades Features

Orders and Trades are features that were calculated based on the raw data. Let's define the best ask price (p^*_{ask}), i.e., the highest price that a buyer is willing to pay for a Bitcoin bid order, and the best bid price (p^*_{bid}), the lowest price that a seller is willing to accept for a Bitcoin ask order. Consequently, the set of the extracted features, within this category, can be described as follows:

- Mid-Market Price (mmp): indicates the average market price and can be calculated as follows:

$$mmp = \frac{p^*_{bid} + p^*_{ask}}{2} \tag{4.1}$$

  This feature represents the target variable to be predicted, where mmp at time t represents an accurate estimate of the true price of Bitcoin at that time instant. Fig. 4.12 illustrates the Bitcoin Mid-Market Price for the considered time period.

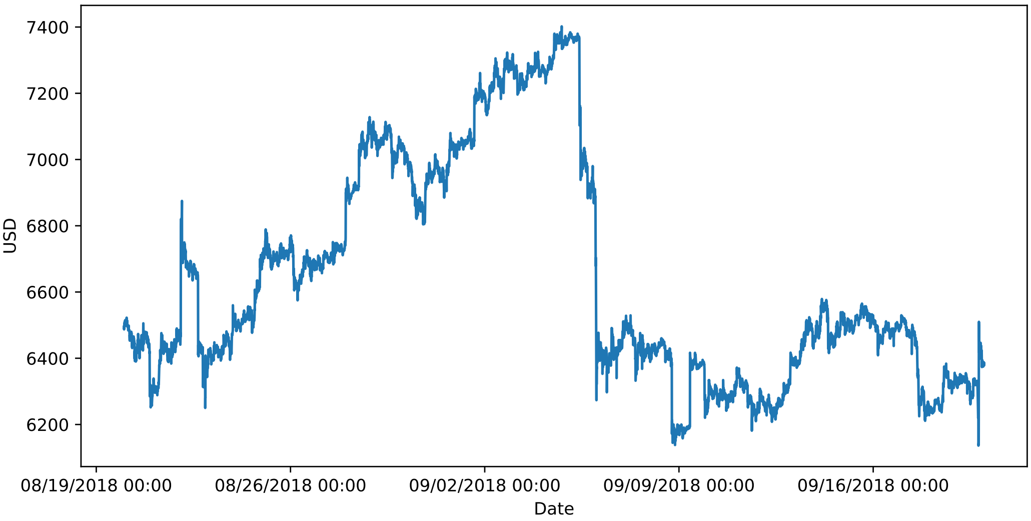


Figure 4.12: Bitcoin Mid-Market Price in the Time Period from August 20th, 2018 to September 20th, 2018.

- Market Spread ($mspread$): indicates the difference between the best ask price and the best bid price, and can be calculated by:

$$mspread = p^*_{ask} - p^*_{bid} \tag{4.2}$$

Smaller values of *mspread* indicate a lower volatility, which result in an insignificant movement of the price.

- Ask/Bid Depth ($D_{\{ask,bid\}}(t)$): indicate the number of available orders per ask side ($D_{ask}(t)$) and bid side ($D_{bid}(t)$), respectively, at time t [2]. Ask/Bid depth represents the liquidity of the market. A market is said to be deep when it is able to fulfill larger buy and sell orders before an order moves the price of Bitcoin. Fig. 4.13 shows the ask/bid depth for the considered time interval.

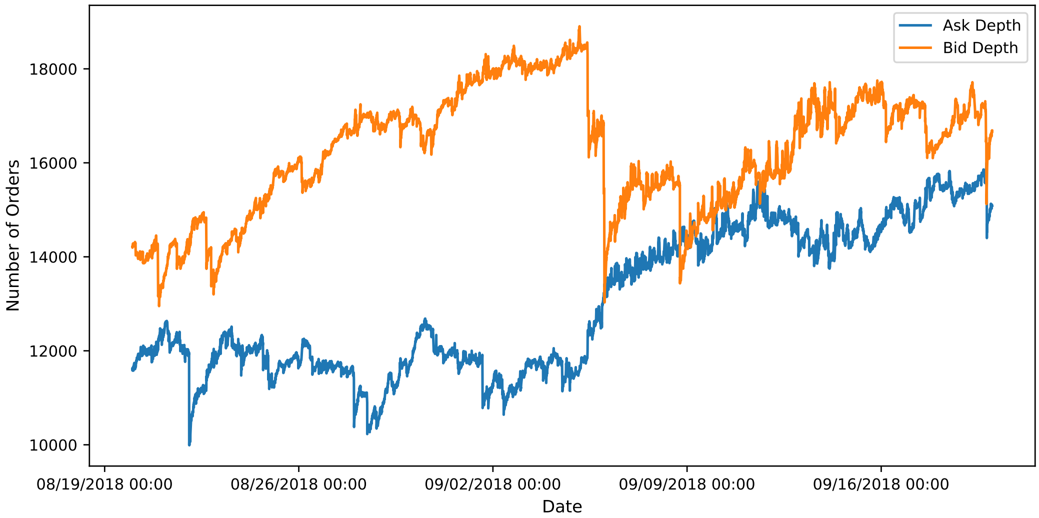


Figure 4.13: Ask/Bid Depth in the Considered Time Frame.

- Ask/Bid Volume ($volume_{\{ask,bid\}}$): indicate the total volume of ask and bid orders. Given a sample data, the ask/bid volume can be computed by:

$$volume_{ask} = \sum_{i=-0}^{-D_{bid}} v_i \tag{4.3}$$

$$volume_{bid} = \sum_{i=0}^{D_{ask}} v_i \tag{4.4}$$

[2] The reference to time t will be dropped when not necessary.

- Weighted Ask/Bid Volume ($weighted\_volume_{\{ask,bid\}}$): indicate the weighted volume of ask and bid orders to better capture the relevance of orders to the movement of the price, and is calculated by:

$$weighted\_volume_{ask} = \sum_{i=-0}^{-D_{bid}} v_i . \frac{1}{mmp - p_i} \tag{4.5}$$

$$weighted\_volume_{bid} = \sum_{i=0}^{D_{ask}} v_i . \frac{1}{p_i - mmp} \tag{4.6}$$

- Depth Chart Quantization: generally, depth charts are bin charts that hold information about the cumulative supply and demand of an asset, at different prices. Thus, valuable information is encapsulated inside such charts that may not be tackled by the previously extracted features. A depth chart is composed of a horizontal axis that depicts Bitcoin prices $pc(i)$ and a vertical axis that depicts the corresponding tradable amount of Bitcoin $vc(p)$ at a specific price p, for both sides, namely, ask and bid. The quantization of the depth chart depends on which side of the chart is considered. For the ask side of the chart, the depth is quantized by collecting the tradable amounts of Bitcoin that are at or below a specific price and taking their sum. As for the bid side, the depth is quantized by collecting the tradable amounts of Bitcoin that are at or above a specific price and taking their sum. Prices and volumes in each chart are distinguished with the notations $pc_{\{ask,bid\}}(i)$ and $vc_{\{ask,bid\}}(p)$, respectively.

  To represent the quantization of the depth chart, a number of bins equal to N_{bins} for each chart is considered. Thus, the width of the bins in each chart can be calculated by:

$$w_{ask} = \frac{pc(D_{ask}) - p^*_{ask}}{N_{bins}} \tag{4.7}$$

$$w_{bid} = \frac{p^*_{bid} - pc(-D_{bid})}{N_{bins}} \tag{4.8}$$

Thus:

$$bin_{ask}(i) = \frac{1}{w_{ask}} \sum_{p=p^*_{ask}+(i-1)\cdot w_{ask}}^{p^*_{ask}+i\cdot w_{ask}} vc_{ask}(p) \tag{4.9}$$

$$bin_{bid}(i) = \frac{1}{w_{bid}} \sum_{p=p^*_{bid}-(i-1)\cdot w_{bid}}^{p^*_{bid}-i\cdot w_{bid}} vc_{bid}(p) \tag{4.10}$$

As a trade-off between resolution and dimensionality, N_{bins} was set to 10. Fig. 4.14 shows the relevant quantities used for the definition of the chart and for the quantization of the bins.

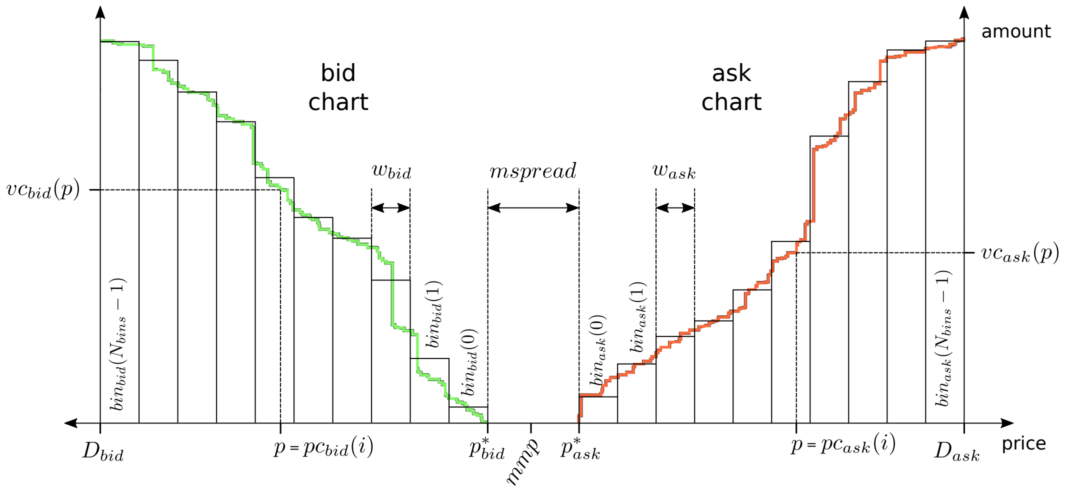


Figure 4.14: Ask/Bid Depth Chart Showing Relevant Quantities.

- Sell/Buy Count ($count_{\{sell,buy\}}$): generally, a trade is always considered to be aggressive. Meaning that a specific trade would affect the movement of the price. Accordingly, and as the name implies, these features indicate the number of trades that have been generated by an aggressive sell or buy.

- Sell/Buy Traded Volume ($traded\_volume_{\{sell,buy\}}$): given that the traded amount has an effect on liquidity, thus, the price movement, these features indicate the amount of Bitcoin that has been transacted due to an aggressive sell or buy.

Table 4.10 illustrates a number of selected features in terms of descriptive statistics in the considered time period, from August 20th, 2018 to September 20th, 2018.

Table 4.10: Summary Statistics for Selected Features. Values are Rounded to Two Decimal Places.

Feature Name	*Minimum*	*Maximum*	*Mean*	*Standard Deviation*
Ask Depth	9993.00	16000.00	13019.42	1527.40
Bid Depth	12951.00	18906.00	16249.11	1289.03
Ask Weighted Volume	3.14	33807.82	2853.07	3820.77
Bid Weighted Volume	21.60	36863.72	2263.84	2936.62
Sell Traded Volume	0.00	927.03	5.20	17.44
Buy Traded Volume	0.00	787.97	5.20	19.03
Market Spread	0.01	94.97	0.06	0.79
Mid-Market Price	6136.25	7402.01	6639.98	335.49

Technical Indicators

Technical indicators can be identified as mathematical calculations, computed based on historical data of an asset, to predict the price movement. Accordingly, the following technical indicators were considered, and computed based on *mmp*:

- Simple Moving Average (*sma*): it indicates the arithmetic moving average of the price of an asset, that can be calculated in a specific time period. It is normally used to smooth out price fluctuations. Taking into account the frequency of the collected data, two *sma*'s were calculated, namely, sma_6 and sma_{16}, indicating the moving averages of *mmp* at time periods 6 (every 12 minutes) and 16 (every 32 minuted), respectively. Given n, the time period considered, *sma* can be calculated by:

$$sma = \frac{mmp_1 + mmp_2 + ... + mmp_n}{n} \tag{4.11}$$

- Exponential Moving Average (*ema*): it indicates the exponential weighted moving average of the price of an asset, by placing exponentially decreasing weights to the prices. Such weights represent higher weights to most recent prices, in the sense that recent prices have a higher significance in predicting future ones. Given n, the time period considered, *ema* can be calculated as follows:

$$ema_t = \alpha[mmp_1 + (1-\alpha)mmp_2 + ... + (1-\alpha)^{n-1}mmp_n] \tag{4.12}$$

where α represents the degree of weighting decrease at each point. Similarly, taking into account the frequency of the collected data, *ema*, ema_{12} and ema_{26} were calculated, indicating the exponential moving average at time periods 1 (every 2 minutes), 12 (every 24 minutes) and 26 (every 52 minutes), respectively.

- Moving Average Convergence Divergence (*macd*): indicates the relationship between the previously calculated *ema*'s, where it can be calculated simply by finding the difference between ema_{26} and ema_{12}. This indicator helps in understanding bearish (prices expected to fall) and bullish (prices expected to rise) movements.

- Bollinger Bands ($band_{\{upper,lower\}}$): Bollinger Bands are lines that are associated with two standard deviations (*sd*) plotted away from the *sma* of the price of an asset. Usually, almost 90% of the original prices are contained within these bands. Thus, breakouts that fall either under or above these bands indicate a major event that may have affected the price movement. Bollinger Bands are typically calculated by:

$$band_{upper} = sma_{21} + (sd_{20} * 2) \tag{4.13}$$

$$band_{lower} = sma_{21} - (sd_{20} * 2) \tag{4.14}$$

- Momentum (*momentum*): it indicates the speed at which a price is changing. It is simply computed by subtracting 1 from the target variable *mmp*.

Fig. 4.15 shows some of the previously mentioned technical indicators for the last 500 data points, from September 19th, 2018 at 07:22:00 a.m. to September 20th,

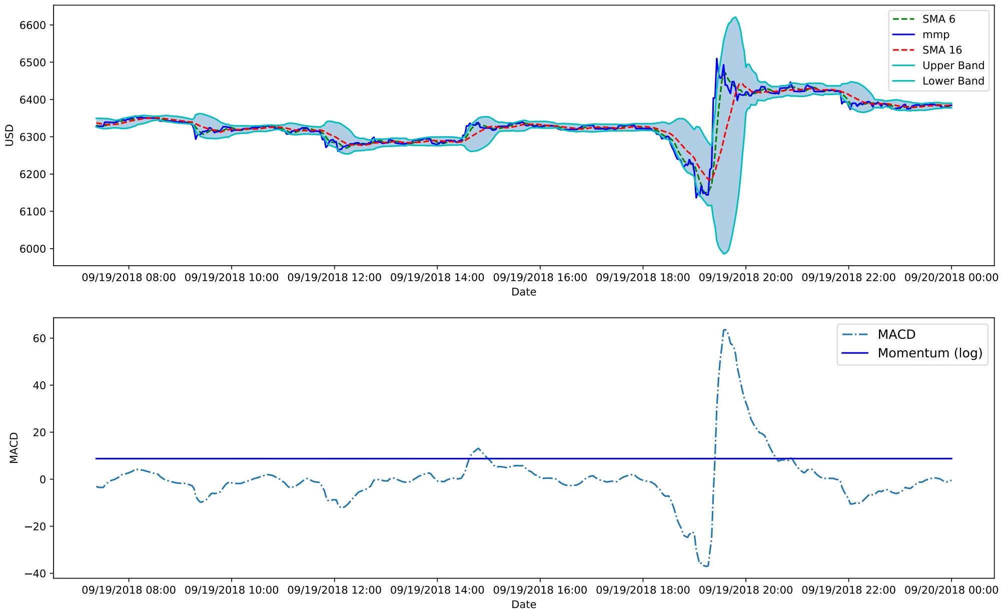


Figure 4.15: Mid-Market Price Technical Indicators.

2018 at 12:00:00 a.m.

4.3.3 Descriptive Results

To better describe the movement of the prices and assess their evolution over time, the target variable *mmp* is further explored by modeling the two variables used to compute it, namely, best ask price (p_0) and best bid price (p_{-0}). For this purpose, HMM are implemented following the approach previously explained in section 3.2.2, through which a new feature is created by clustering the prices, with respect to a predefined number of hidden states.

Consequently, to construct the descriptive model, Problem 3, as explained in section 3.2, is addressed using Baum-Welch algorithm, and given Y^i, the maximum likelihood estimates of the HMM parameters were computed using the Expectation Maximization algorithm, assuming a predefined number of hidden states M.

Now that the HMM parameters are estimated, the likelihood of Y^i can be calculated using the Forward Backward algorithm, followed by predicting the optimal sequence of the hidden states, which represents a new feature (*state*) added to the set of extracted features that were previously defined.

Given the nature of the data, the predefined number of hidden states is set to 2, which in fact, proved to have the highest likelihood compared to other presumed number of hidden states. Table 4.11 shows the means and variances of the prices per each of the hidden states. Moreover, fig. 4.16 illustrates the set of prices plotted with respect to the hidden state assigned by the proposed HMM, using a diagonal variance-covariance matrix.

Table 4.11: Descriptive Statistics of Hidden States. Values are Rounded to Three Decimal Places.

Hidden	*Best Ask Price*		*Best Bid Price*	
State	*Mean*	*Variance*	*Mean*	*Variance*
0	6395.035	9595.902	6394.976	9598.216
1	6985.890	53543.020	6985.834	53547.423

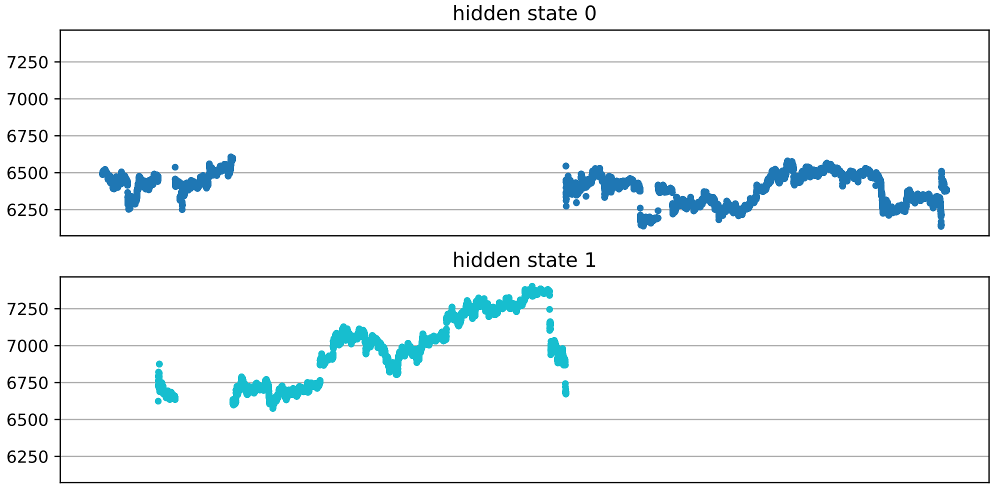


Figure 4.16: Prices Plotted per Hidden State.

Specifically, fig. 4.17 illustrates thes first 4500 data points of best ask price, clustered by the hidden states.

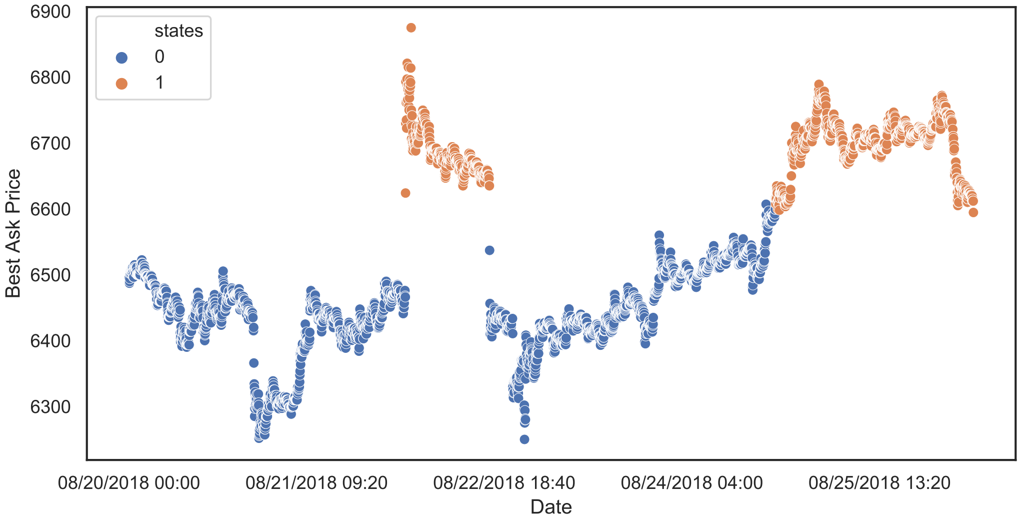


Figure 4.17: Best Ask Price Clustered by HMM States.

4.3.4 Predictive Results

Prior to presenting the predictive results based on the implementation of the proposed model, two more steps are taken into account: the optimization of several parameters of LSTM using the Genetic Algorithms and preparing the optimized LSTM for training the final prepared dataset.

Genetic Algorithms for LSTM Optimization

Following the methodology explained in section 3.3.1 and adopting the implementations of [57, 58], a number of LSTM parameters is optimized before training the proposed model with the final set of features.

To do so, the process starts with the initialization of the population by defining the parameters to be optimized and, thus, creating a number of LSTM networks corresponding to the combinations of the defined parameters.

Having done that, the prepared dataset is split into a training dataset and a testing dataset. Given the different combinations defined for the LSTM, the training process starts using the training dataset, followed by a performance evaluation using the testing dataset. The evaluation is done given a fitness function corresponding to the loss function defined in the LSTM, in this case MSE. Accordingly, the different LSTM networks are ranked with respect to the fitness function defined, and a percentage of the population from the highly ranked LSTM networks is kept, a few percentage from the non-highly ranked ones is kept as well and, finally, a percentage of the population is randomly dropped, completing a generation. Based on the identified number of generations, the algorithm iterates over these steps, and the LSTM with the lowest MSE is selected.

Accordingly, the considered optimized parameters include: number of epochs (100), batch size (10), number of layers (3), number of neurons (100), dropout rate (0.2), optimizer (*adam*), loss MSE and evaluation metrics MAE.

Predictive Modeling with LSTM

Before training the optimized LSTM with the final prepared set of features, a final step is needed: the introduction of the log returns ($log\_returns$). Given the nature of the collected data, as well as the target variable, the log returns of the target variable *mmp* are computed by:

$$log\_return_{H(t)} = log\left(\frac{mmp(t)}{mmp(t-H)}\right) \tag{4.15}$$

where H is the prediction horizon, which defines how far ahead the model predicts in the future.

Accordingly, the final set of features is composed of $22,321$ data points and 52 features. 70% of the data has been used for training the LSTM network (15610 data points), 20% has been used for validation (4683 data points), while the final 10% (2008 data points) has been used for testing as an out of sample dataset. Additionally, dropout blocks were used between the hidden layers to avoid over-fitting, paired with an early stopping mechanism.

Fig. 4.18 shows a comparison between the actual prices and 1-step ahead predictions where $H = 1$. As illustrated, the predictions of the proposed model are close to the actual ones and the movement of the prices is somewhat consistent. Moreover,

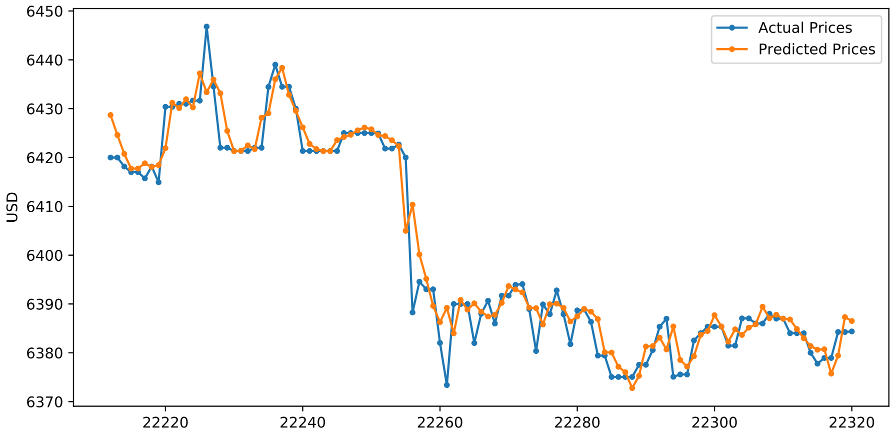


Figure 4.18: Results for 1-Step Ahead Prices Prediction.

to compare the performance of the proposed model to more traditional time-series forecasting models, an ARIMA model has been implemented and using the final set of features for training and predicting based on an out of sample dataset. Similarly, a Genetic Algorithm-optimized conventional LSTM has been implemented to evaluate the importance of the proposed model.

Accordingly, table 4.12 shows the performance of the three implemented models. As illustrated, the proposed model decreased the error rate significantly compared to ARIMA and the conventional LSTM, which proves the impact of HMM on enhancing the performance of a conventional LSTM. Additionally, in order to provide an unbiased sense of models' performance, these measurements are computed based on an out of sample (test) dataset that was not used to neither train nor fine-tune the proposed model.

Table 4.12: Performance Evaluation for Implemented Models. Values are rounded to three decimal places.

Model Name	*MSE*	*RMSE*	*MAE*
ARIMA	20153.722	141.964	112.060
LSTM	49.089	7.006	2.652
HMM-LSTM	33.888	5.821	2.510

As a further step, the performance of the proposed model is tested for multi-step prediction where $H = 2$. The results are illustrated in fig. 4.19. As expected, the performance was affected where MSE increased to 63.574. The rise is due to the iterative structure of LSTM where the prediction of one layer is passed to the next, thus, the error is accumulated to a larger number after two time steps ahead compared to only one step. However, looking back at the predicted results compared to the actual ones in fig. 4.19, it is safe to say that the predictions are realistic even with a relatively higher error rate.

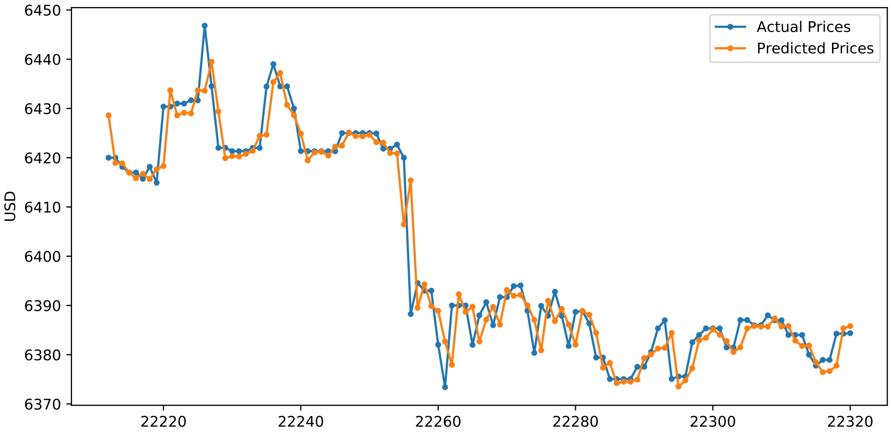


Figure 4.19: Results for 2-Step Ahead Prices Prediction.

Chapter 5

Conclusions

Living in a data-intensive environment is a natural consequence of the continuous innovations and technological advancements. The availability of such data created countless opportunities for the research community in addressing possible domain-specific challenges following the Data Science approach, aiming at learning, modeling, and mining complex domain-related data, focusing on discovering the knowledge hidden inside.

Theoretically speaking, Data Science approaches are applicable to any given domain. A domain that is particularly interesting is the Financial Technology, or what is currently known as FinTech, reflecting the “marriage” of technology and finance. Applying Data Science to such a domain contributes in the possibility of improving already-existing financial services as well as creating new ones. Within this field, a genuine financial innovation is represented by cryptocurrencies.

The first cryptocurrency, the most popular, and widely used, i.e., Bitcoin, has caught the eye of many researchers due its decentralized structure and low-cost transfer of value anytime and anywhere in the world. Consequently, implementing Data Science on Bitcoin-related data, opens many opportunities for perceiving such a newly presented, non-traditional asset, through analyzing related pricing data to understand its respective market that has massively grown in popularity, prices and volatility. Therefore, the overall objective of this work is to present applied Data

Science approaches in FinTech, focusing on proposing innovative descriptive and predictive models for studying and exploring Bitcoin Price Dynamics and Bitcoin Price Prediction.

A first step for addressing this objective, is conducting a thorough literature review in the domains of Data Science, FinTech and Data Science for FinTech as presented in Chapter 2. Given that neither Data Science nor FinTech are novel, the evolution of each domain is discussed, along with their respective challenges and opportunities. Thus, the overall objective can be further considered to address the challenges in the previously mentioned domains. Firstly, the objective within the field of Data Science is represented by developing domain-specific innovative models and algorithms that aims at modeling, learning and mining related data and discovering insightful information. Secondly, taking the FinTech domain into account, the objective is to address the emergence of Bitcoin by providing empirical evidences and developing related theories. Lastly, the third objective within Data Science for FinTech is to propose innovative descriptive and predictive models aiming at studying the research areas of Bitcoin Price Dynamics and Bitcoin Price Prediction.

From a finer perspective, in Chapter 3 two different models were proposed addressing the research area of Bitcoin Price Dynamics, along with their respective implementations in Chapter 4.

The first proposed model is a Network VAR model that explains the dynamics of Bitcoin prices, based on a correlation network VAR process that models the interconnections between Bitcoin prices from different exchange markets and classical assets prices. The methodological contribution lies in the introduction of partial correlations and correlations networks into VAR models. In turn, this allows to describe the correlation patterns between Bitcoin prices to disentangle the autoregressive component of prices from its contemporaneous part, explained by the co-movement with other market prices. Although the proposed model is mainly descriptive, the introduction of VAR correlation networks enables the development of a predictive model, which leverages the information contained in the correlation patterns.

The empirical findings show that Bitcoin prices from different exchange markets are highly interrelated, as in an efficiently integrated market, with prices from larger and/or more connected exchange markets driving other prices. The results also confirm that Bitcoin prices are typically unrelated with classical market prices, thus, further supporting the diversification benefit property of cryptocurrencies. Additionally, the proposed model is able to predict Bitcoin prices with an error that can be approximated to about 11% of the average price. However, this error varies considerably among different exchange markets; prices from central Bitcoin exchange markets are easier to predict. For almost all markets, the inclusion of a contemporaneous component in the predictive model leads to a higher predictive accuracy than that obtained with a simpler, pure autoregressive model.

Further research directions within this area, given the proposed Network VAR model, may include, collecting more data on traded volumes and possibly the electronic identities of the traders, to investigate the reasons behind “local” behaviors of different exchanges. From a methodological perspective, it may be worth considering extending correlation network models to be time-dependent, although this requires getting data with a higher frequency.

The second proposed model is a Hidden Markov Model that explains the observed time dynamics of Bitcoin prices from different exchange markets, by means of the latent time dynamics of a predefined number of latent states, to model regime switches between different price vectors, going from “bull” to “stable” and “bear” times. The contributions of this work are two-fold: providing a further understanding of Bitcoin price dynamics from an econometric point of view and implementing an easy-to-use likelihood ratio test for comparing differently implemented Hidden Markov Models. Accordingly, three alternative types of hidden structure has been considered, using two, three, and four predefined number of hidden states, along with two different variance-covariance matrices; a full variance-covariance matrix and a more parsimonious diagonal variance-covariance matrix, thus, a total of 3×2 alternative models have been implemented.

A first look at the collected data showed a divergence of Bitcoin prices from different exchange markets, opposing the economic law of “one asset, one price”. The hypothesis constructed is that such differences can be explained by the endogenous relationships between exchange markets prices, which are in turn explained by different latent states of the cryptoasset economy.

Although several models were implemented, the empirical findings are mainly based on a 3-state model that has been built twice using a full and a diagonal variance-covariance matrix, as it proved to have the highest likelihood compared to those models implemented using two and four hidden states. Accordingly, considering a 3-state model implemented using a full variance-covariance matrix, the results show that one hidden state is concentrated in the initial considered time period where Bitcoin prices were relatively new and barely increasing, while the other two hidden states alternate between lower and higher prices in more recent times.

Given that a full variance-covariance matrix may be too complex to fit the data well, a 3-state diagonal variance-covariance matrix model has been implemented. Such a model implies that Bitcoin prices from any exchange market is independent on the price of other markets, conditionally on the hidden states, at any time point. Simply put, this suggests that the dynamics of Bitcoin prices from different exchange markets is fully explained by the dynamics of the hidden states rather than the differences in prices between different exchanges. Indeed, the empirical findings show that the first hidden states in concentrated in the initial time period, the second hidden state is mostly concentrated in a period of time where Bitcoin prices were steadily increasing, while the third hidden state is mostly concentrated in the last period.

As a consequence of the above-mentioned results, it is safe to say that a 3-state model with a diagonal variance-covariance matrix provides a better modeling for regime switches implied by the data. To confirm this conclusion, the likelihood ratio testing statistics is adopted. With a $p - value$ of $2.34e - 54$, it is shown that a more parsimonious 3-states diagonal matrix model is better. Moreover, the test has been implemented assuming two and four hidden states, showing that the use of a diagonal variance-covariance matrix is always preferred. On the other hand, the same test has been implemented, assuming a diagonal variance-covariance matrix and a different number of predefined hidden states. Accordingly, it is also proved that a 3-state model has a likelihood that is significantly higher than that of a 2-state and 4-state model. Thus, implying that Bitcoin prices, given the young history, can be explained by using three alternative states of “bear”, “stable” and “bull” markets.

Finally, the proposed model, although mainly descriptive, has been extended to predict Bitcoin prices, showing a good predictive performance when implemented on an out-of-sample dataset. Given the considered exchange markets, the predictive power of a 3-state diagonal model is always higher than a 3-state full model, in line with the likelihood ratio test.

Further research directions within the same area, given the proposed Hidden Markov Model may include:

- extending the comparison to any number of exchange markets and any number of hidden states, thanks to the employed likelihood ration test statistics;
- extending the analysis to include different cryptocurrencies;
- implementing the model using intra-daily data.

The third, and final, proposed model in the research area of Bitcoin Price Prediction was investigated after conducting a thorough literature review. The main research challenge was the need to provide an innovative model that predicts Bitcoin prices accurately, by introducing new features that are not usually considered in the literature. Accordingly, an innovative hybrid model is proposed using Hidden Markov Models and Genetic Algorithm-optimized LSTM networks. The details of the proposed model are explained in Chapter 3 along with the implementation in Chapter 4.

The essence of the proposed model falls, from one hand, in its ability to address Bitcoin prices from a descriptive point of view by implementing Hidden Markov

Models and creating a new feature based on encapsulated hidden information that cannot be directly seen nor extracted, and on implementing Genetic Algorithms on LSTM network to fine-tune its parameters, from the other hand.

Composed of four phases, the proposed model starts from Data Collection, where a raw dataset is created based on raw Bitcoin-related collected data, which is meant to be exploited in the following phases. Followed by a Feature Extraction phase, the raw data are used and new features, which are beneficial to the prediction process, are extracted and saved in a features' dataset. The third phase is Data Modeling, which can be divided into descriptive modeling using Hidden Markov Model and predictive modeling using Genetic Algorithm-optimized LSTM network. The fourth and final phase is the performance evaluation including three metrics, namely, Mean Squared Error, Root Mean Squared Error and Mean Absolute Error, calculated based on an out-of-sample dataset that has not been used neither to train the model nor to fine-tune the LSTM network.

To compare the performance of the proposed model to other models, a more traditional ARIMA model has been implemented, as well as a conventional Genetic Algorithm-optimized LSTM. With an MSE of 33.888, an $RMSE$ of 5.821 and an MAE of 2.510, the proposed model achieved the lowest errors among all of the implemented models, which proves the effectiveness of the proposed model in predicting Bitcoin prices.

Further research directions, given the proposed model, within the research area of Bitcoin price prediction, may include; extending the proposed model by considering additional features that can be extracted from the Blockchain to provide information about the internal details of Bitcoin transactions and adapting the model to study the dynamics of different cryptocurrencies.

Chapter 6

List of Submissions and Publications

The following sections 6.1 and 6.2 list the submitted abstracts and published full papers, carried out through the past three years of PhD.

6.1 Abstracts

- **Iman Abu Hashish**, Gianmario Motta and Michela Meazza, *"NavApp: An Indoor Navigation Application on Smartphones for Libraries and alike Environments,"* 3$^{\text{rd}}$ Italian Conference on ICT for Smart Cities & Communities (I-CiTies), Bari, Italy, 2017.

- Kaixu Liu, Gianmario Motta, Tianyi Ma and **Iman Abu Hashish**, *"SMARTIN: A Smart Indoor Mobility Platform,"* 3$^{\text{rd}}$ Italian Conference on ICT for Smart Cities & Communities (I-CiTies), Bari, Italy, 2017.

- Linlin You, Gianmario Motta, Kaixu Liu, Tianyi Ma and **Iman Abu Hashish**, *"CITY FEED: A Pilot System of Citizen-Sourcing for City Issue Management,"* 3$^{\text{rd}}$ Italian Conference on ICT for Smart Cities & Communities (I-CiTies), Bari, Italy, 2017.

- Paolo Giudici, **Iman Abu Hashish**, Paolo Pagnottoni and Kamonchai Rujirarangsan, *"Price Discovery in Crypto-Currency Markets,"* Crypto-Currencies in a Digital Economy, Berlin, Germany, 2017.

6.2 Full Papers

- **Iman Abu Hashish**, Gianmario Motta, Tianyi Ma and Kaixu Liu, *"An Analysis of Social Data Credibility for Services Systems in Smart Cities – Credibility Assessment and Classification of Tweets,"* in Cloud Infrastructures, Services, and IoT Systems for Smart Cities, pp. 119-130, Springer.

- Gianmario Motta, Kaixu Liu, **Iman Abu Hashish** and Michela Meazza, *"Integration of Services Systems,"* China Europe Symposium on Software Engineering Education (CEISEE), Athens, Greece, 2017.

- Kaixu Liu, Gianmario Motta, Bige Tunçer and **Iman Abu Hashish**, *"A 2D and 3D Indoor Mapping Approach for Virtual Navigation Services,"* 2017 IEEE Symposium on Service-Oriented System Engineering (SOSE), San Francisco, CA, 2017, pp. 102-107.

- Gianmario Motta, Antonella Longo, Kaixu Liu and **Iman Abu Hashish**, *"Services Systems for Digital Services: A Framework,"* International Conference on Business and Information (BAI), Hiroshima, Japan, 2017.

- **Iman Abu Hashish**, Gianmario Motta, Michela Meazza, Guoqing Bu, Kaixu Liu, Lorenzo Duico and Antonella Longo, *"NavApp: An Indoor Navigation Application – A Smartphone Application for Libraries,"* WPNC'17 IEEE 14th Workshop on Positioning, Navigation and Communications, Bremen, Germany, 2017.

- Paolo Giudici and **Iman Abu Hashish**, *"What determines bitcoin exchange prices? A network VAR approach,"* Finance Research Letters, 28, pp. 309-318.

- Gianmario Motta, **Iman Abu Hashish** and Antonella Longo, *"NavApp: A mobile App as a Master's Thesis,"* China Europe Symposium on Software Engineering Education (CEISEE), Shenzhen, China, 2018.

- Gianmario Motta, **Iman Abu Hashish**, Michela Meazza, Guoqing Bu and Antonella Longo, *"Indoor Mobility as a Project Work: a case study,"* 14th Conference of the Italian Chapter of AIS, Milan, Italy, 2018.

- Paolo Giudici and **Iman Abu Hashish**, *"A Hidden Markov Model to Detect Regime Changes in Cryptoasset Markets,"*, 2018, Submitted.

- **Iman Abu Hashish**, Shiva Darjani, Fabio Forni, Gianluca Andreotti, and Tullio Facchinetti, *"A Hybrid Model for Bitcoin Prices Prediction Using Hidden Markov Models and Optimized LSTM Networks,"* 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2019. To Appear.

Bibliography

[1] American banker. https://en.wikipedia.org/wiki/American_Banker. Accessed: 2019-08-05.

[2] Did citi coin the term 'fintech'? https://www.americanbanker.com/opinion/friday-flashback-did-citi-coin-the-term-fintech. Accessed: 2019-08-05.

[3] Bankthink fintech (the word, that is) evolves. https://www.americanbanker.com/opinion/fintech-the-word-that-is-evolves. Accessed: 2019-08-05.

[4] Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra. Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, 1(3):1, 2018.

[5] Iman Abu Hashish, Shiva Darjani, Fabio Forni, Gianluca Andreotti, and Tullio Facchinetti. A hybrid model for bitcoin prices prediction using hidden markov models and optimized LSTM networks. In *2019 24^{th} IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2019. To Appear.

[6] Khamis Hamed Al-Yahyaee, Mobeen Ur Rehman, Walid Mensi, and Idries Mohammad Wanas Al-Jarrah. Can uncertainty indices predict bitcoin prices? a revisited analysis using partial and multivariate wavelet approaches. *The North American Journal of Economics and Finance*, 49:47–56, 2019.

[7] Douglas W Arner, Janos Barberis, and Ross P Buckley. The evolution of fintech: A new post-crisis paradigm. *Geo. J. Int'l L.*, 47:1271, 2015.

[8] Douglas W Arner, Jànos Barberis, and Ross P Buckley. *FinTech and RegTech in a Nutshell, and the Future in a Sandbox.* CFA Institute Research Foundation, 2017.

[9] Douglas W Arner, Janos Barberis, Ross P Buckley, et al. 150 years of fintech: An evolutionary analysis. *Jassa*, page 22, 2016.

[10] Henri Arslanian and Fabrice Fischer. The rise of fintech. In *The Future of Finance*, pages 25–56. Springer, 2019.

[11] George S Atsalakis, Ioanna G Atsalaki, Fotios Pasiouras, and Constantin Zopounidis. Bitcoin price forecasting with neuro-fuzzy techniques. *European Journal of Operational Research*, 276(2):770–780, 2019.

[12] Amin Azari. Bitcoin price prediction: An ARIMA approach. *arXiv preprint arXiv:1904.05315*, 2019.

[13] Prateek Bedi and Tripti Nashier. On the investment credentials of bitcoin: A cross-currency perspective. *Research in International Business and Finance*, page 101087, 2019.

[14] Benjamin M Blau. Price dynamics and speculative trading in bitcoin. *Research in International Business and Finance*, 41:493–499, 2017.

[15] Bloomberg terminal. `https://www.bloomberg.com/professional/solution/bloomberg-terminal/`. Accessed: 2020-01-14.

[16] Jamal Bouoiyour, Refk Selmi, Aviral Kumar Tiwari, Olaolu Richard Olayeni, et al. What drives bitcoin price. *Economics Bulletin*, 36(2):843–850, 2016.

[17] Elie Bouri, Georges Azzi, and Anne Haubo Dyhrberg. On the return-volatility relationship in the bitcoin market around the price crash of 2013. *Available at SSRN 2869855*, 2016.

[18] Morten Brandvold, Peter Molnár, Kristian Vagstad, and Ole Christian Andreas Valstad. Price discovery on bitcoin exchanges. *Journal of International Financial Markets, Institutions and Money*, 36:18–35, 2015.

[19] Longbing Cao. Domain-driven data mining: Challenges and prospects. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):755–769, 2010.

[20] Longbing Cao. Data science: Nature and pitfalls. *IEEE Intelligent Systems*, 31(5):66–75, 2016.

[21] Longbing Cao. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3):43, 2017.

[22] Longbing Cao. Data science: challenges and directions. *Communications of the ACM*, 60(8):59–68, 2017.

[23] Longbing Cao. *Data Science Thinking: The Next Scientific, Technological and Economic Revolution*. Springer, 2018.

[24] Jenna Carr. An introduction to genetic algorithms. *Senior Project*, 1(40):7, 2014.

[25] Jaturon Chattratichat, John Darlington, Moustafa Ghanem, Yike Guo, Harald Frank Hüning, Martin Köhler, Janjao Sutiwaraphun, Hing Wing To, and Dan Yang. Large scale data mining: Challenges and responses. In *KDD*, pages 143–146, 1997.

[26] Dunren Che, Mejdl Safran, and Zhiyong Peng. From big data to big data mining: challenges, issues, and opportunities. In *International conference on database systems for advanced applications*, pages 1–15. Springer, 2013.

[27] Germán Cheuque Cerda and Juan L Reutter. Bitcoin price prediction through opinion mining. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 755–762. ACM, 2019.

[28] Thanaset Chevapatrakul and Danilo V Mascia. Detecting overreaction in the bitcoin market: A quantile autoregression approach. *Finance Research Letters*, 2018.

[29] Hyejung Chung and Kyung-shik Shin. Genetic algorithm-optimized long short-term memory network for stock market prediction. *Sustainability*, 10(10):3765, 2018.

[30] Pavel Ciaian, Miroslava Rajcaniova, et al. Virtual relationships: Short-and long-run evidence from bitcoin and altcoin markets. *Journal of International Financial Markets, Institutions and Money*, 52:173–195, 2018.

[31] Global investment bank and financial services - citi. `https://www.citigroup.com/citi/`. Accessed: 2019-08-06.

[32] Coinbase pro public api. `https://docs.pro.coinbase.com/`. Accessed: 2019-01-10.

[33] Top 100 cryptocurrency exchanges by trade volume. `https://coinmarketcap.com/rankings/exchanges/`. Accessed: 2019-03-20.

[34] Shaen Corbet, Veysel Eraslan, Brian Lucey, and Ahmet Sensoy. The effectiveness of technical trading rules in cryptocurrency markets. *Finance Research Letters*, 31:32–37, 2019.

[35] Shaen Corbet, Brian Lucey, Andrew Urquhart, and Larisa Yarovaya. Cryptocurrencies as a financial asset: A systematic analysis. *International Review of Financial Analysis*, 62:182–199, 2019.

[36] Shaen Corbet, Andrew Meegan, Charles Larkin, Brian Lucey, and Larisa Yarovaya. Exploring the dynamic relationships between cryptocurrencies and other financial assets. *Economics Letters*, 165:28–34, 2018.

[37] Cryptocompare api: The ultimate API solution. `https://min-api.cryptocompare.com`. Accessed: 2019-09-10.

[38] Ben Kei Daniel. Big data and data science: A critical review of issues for educational research. *British Journal of Educational Technology*, 50(1):101–113, 2019.

[39] Data analytics: Crunching the future. `https://www.bloomberg.com/news/articles/2011-09-08/data-analytics-crunching-the-future`. Accessed: 2019-08-05.

[40] André Henrique de Oliveira Monteiro, Adler Diniz de Souza, Bruno Guazzelli Batista, and Mauricio Zaparoli. Market prediction in criptocurrency: A systematic literature mapping. In *16th International Conference on Information Technology-New Generations (ITNG 2019)*, pages 601–604. Springer, 2019.

[41] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[42] Dot-com bubble. `https://en.wikipedia.org/wiki/Dot-com_bubble`. Accessed: 2019-08-06.

[43] International conference on data science and advanced analytics. `http://www.dsaa.co/`. Accessed: 2019-06-30.

[44] Gerald P Dwyer. The economics of bitcoin and similar private digital currencies. *Journal of Financial Stability*, 17:81–91, 2015.

[45] Anne Haubo Dyhrberg. Bitcoin, gold and the dollar–a GARCH volatility analysis. *Finance Research Letters*, 16:85–92, 2016.

[46] Mehmet Levent Erdas and Abdullah Emre Caglar. Analysis of the relationships between bitcoin and exchange rate, commodities and global indexes by asymmetric causality test. *Eastern Journal of European Studies*, 9(2):27, 2018.

[47] Top cryptocurrency exchange list. `https://coin.market/markets/info`. Accessed: 2018-11-21.

[48] Top cryptocurrency exchange list. https://cryptocoincharts.info/markets/info. Accessed: 2018-01-01.

[49] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.

[50] Usama Fayyad and Paul Stolorz. Data mining and KDD: Promise and challenges. *Future generation computer systems*, 13(2-3):99–115, 1997.

[51] Andrea Flori. News and subjective beliefs: A bayesian approach to bitcoin investments. *Research in International Business and Finance*, 2019.

[52] Andrea Flori et al. Cryptocurrencies in finance: Review and applications. *International Journal of Theoretical and Applied Finance (IJTAF)*, 22(05):1–22, 2019.

[53] Sean Foley, Jonathan R Karlsen, and Tālis J Putniņš. Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *The Review of Financial Studies*, 32(5):1798–1853, 2019.

[54] Jerome H Friedman. Data mining and statistics: What's the connection? *Computing Science and Statistics*, 29(1):3–9, 1998.

[55] The fed should seize blockchain's potential central banks must embrace new financial technologies to boost market stability. https://www.ft.com/content/9df7562c-093a-11e9-a242-6043097d0789. Accessed: 2019-08-07.

[56] Keke Gai, Meikang Qiu, and Xiaotong Sun. A survey on fintech. *Journal of Network and Computer Applications*, 103:262–273, 2018.

[57] Evolve a neural network with a genetic algorithm. https://github.com/harvitronix/neural-network-genetic-algorithm. Accessed: 2019-09-16.

[58] Neuro-evelution for neural network hyper parameter tuning. https://github.com/subpath/neuro-evolution. Accessed: 2019-09-16.

[59] Global financial crisis. `https://en.wikipedia.org/wiki/Financial_crisis_of_2007-2008`. Accessed: 2019-08-05.

[60] Zoubin Ghahramani. An introduction to hidden markov models and bayesian networks. In *Hidden Markov models: applications in computer vision*, pages 9–41. World Scientific, 2001.

[61] Paolo Giudici. *Applied data mining: statistical methods for business and industry*. John Wiley & Sons, 2005.

[62] Paolo Giudici. Financial data science. *Statistics & Probability Letters*, 136:160–164, 2018.

[63] Paolo Giudici. Fintech risk management: A research challenge for artificial intelligence in finance. *Frontiers in Artificial Intelligence*, 1:1–6, 2018.

[64] Paolo Giudici and Iman Abu Hashish. A hidden markov model to detect regime changes in cryptoasset markets. Unpublished, 2018.

[65] Paolo Giudici and Iman Abu-Hashish. What determines bitcoin exchange prices? a network VAR approach. *Finance Research Letters*, 28:309–318, 2019.

[66] Paolo Giudici and Paolo Pagnottoni. Vector error correction models to measure connectedness of bitcoin exchange markets. *Applied Stochastic Models in Business and Industry*, 2019.

[67] Paolo Giudici and Laura Parisi. Corisk: Credit risk contagion with correlation network models. *Risks*, 6(3):95, 2018.

[68] Paolo Giudici and Gloria Polinesi. Crypto price discovery through correlation networks. *Annals of Operations Research*, pages 1–15, 2019.

[69] Paolo Giudici, Tobias Ryden, and Pierre Vandekerkhove. Likelihood-ratio tests for hidden markov models. *Biometrics*, 56(3):742–747, 2000.

[70] Itay Goldstein, Wei Jiang, and G Andrew Karolyi. To fintech and beyond. *The Review of Financial Studies*, 32(5):1647–1661, 04 2019.

[71] Yakup Görür. Bitcoin price detection with pyspark using random forest. *Ozyegin University*, 2018.

[72] Matthew J Graham. The art of data science. In *Astrostatistics and Data Mining*, pages 47–59. Springer, 2012.

[73] Do-Alex Greaves and Benjamin Au. Using the bitcoin transaction graph to predict the price of bitcoin. *No Data*, 2015.

[74] Sneha Gullapalli. Learning to predict cryptocurrency price using artificial neural network models of time series. *Kansas State University*, 2018.

[75] Tian Guo and Nino Antulov-Fantulin. Predicting short-term bitcoin price fluctuations from buy and sell orders. *arXiv preprint arXiv:1802.04065*, 2018.

[76] Tian Guo, Albert Bifet, and Nino Antulov-Fantulin. Bitcoin volatility forecasting with a glimpse into buy and sell orders. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 989–994. IEEE, 2018.

[77] David J Hand. Data mining: statistics and more? *The American Statistician*, 52(2):112–118, 1998.

[78] David J Hand. Data mining: new challenges for statisticians. *Social Science Computer Review*, 18(4):442–449, 2000.

[79] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[80] Mark Holub and Jackie Johnson. Bitcoin research across disciplines. *The information society*, 34(2):114–126, 2018.

[81] Jang Huisu, Jaewook Lee, Hyungjin Ko, and Woojin Lee. Predicting bitcoin prices by using rolling window LSTM model. *ACM, July*, 2018.

[82] International association for statistical computing (IASC). `http://iasc-isi.org/`. Accessed: 2019-06-30.

[83] The international statistical institute (ISI). `https://www.isi-web.org/`. Accessed: 2019-06-30.

[84] Arti Jain, Shashank Tripathi, Harsh DharDwivedi, and Pranav Saxena. Forecasting price of cryptocurrencies using tweets sentiment analysis. In *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–7. IEEE, 2018.

[85] Huisu Jang and Jaewook Lee. An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *Ieee Access*, 6:5427–5437, 2017.

[86] ES Karakoyun and AO Cibikdiken. Comparison of ARIMA time series model and LSTM deep learning algorithm for bitcoin price forecasting. In *The 13th Multidisciplinary Academic Conference in Prague 2018 (The 13th MAC 2018)*, pages 171–180, 2018.

[87] Paraskevi Katsiampa, Shaen Corbet, and Brian Lucey. High frequency volatility co-movements in cryptocurrency markets. *Journal of International Financial Markets, Institutions and Money*, 2019.

[88] Kdd-89: Ijcai-89 workshop on knowledge discovery in databases. `https://www.kdnuggets.com/meetings-past/kdd89/index.html`. Accessed: 2019-06-30.

[89] Kdd conferences. `https://www.kdd.org/conferences`. Accessed: 2019-06-30.

[90] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one*, 11(8):e0161197, 2016.

[91] Marius Kinderis, Marija Bezbradica, and Martin Crane. Bitcoin currency fluctuation. *COMPLEXIS*, 2018.

[92] Kenneth E Kinnear, William B Langdon, Lee Spector, Peter J Angeline, and Una-May O'Reilly. *Advances in genetic programming*, volume 3. MIT press, 1999.

[93] Frode Kjærland, Aras Khazal, Erlend A. Krogstad, Frans B. G. Nordstrøm, and Are Oust. An analysis of bitcoin's price dynamics. *Journal of Risk and Financial Management*, 11(4), 2018.

[94] Osamu Kodama, Lukáš Pichl, and Taisei Kaizoji. Regime change and trend prediction for bitcoin time series data. In *CBU International Conference Proceedings*, volume 5, pages 384–388, 2017.

[95] The pulse of fintech 2018 - biannual global analysis of investment in fintech. `https://home.kpmg/xx/en/home.html`. Accessed: 2019-08-05.

[96] Do-Hyung Kwon, Ju-Bong Kim, Ju-Sung Heo, Chan-Myung Kim, and Youn-Hee Han. Time series classification of cryptocurrency price trend based on a recurrent LSTM neural network. *Journal of Information Processing Systems*, 15(3), 2019.

[97] Nikolaos A Kyriazis and Paraskevi Prassa. Which cryptocurrencies are mostly traded in distressed times? *Journal of Risk and Financial Management*, 12(3):135, 2019.

[98] Salim Lahmiri and Stelios Bekiros. Cryptocurrency forecasting with deep learning chaotic neural networks. *Chaos, Solitons & Fractals*, 118:35–40, 2019.

[99] Connor Lamon, Eric Nielsen, and Eric Redondo. Cryptocurrency price prediction using news and social media sentiment. *SMU Data Sci. Rev*, 1(3):1–22, 2017.

[100] Connor Lamon, Eric Nielsen, and Eric Redondo. Cryptocurrency price change prediction using news and social media sentiment. *Stanford University*, 2018.

[101] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[102] Alexandra L'heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. Machine learning with big data: Challenges and approaches. *IEEE Access*, 5:7776–7797, 2017.

[103] Christopher olah: Understanding LSTM networks. `http://colah.github.io/posts/2015-08-UnderstandingLSTMs/`. Accessed: 2019-08-27.

[104] Wei Lu. Blockchain technology and its applications in fintech. *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: Second International Conference*, 2018.

[105] Isaac Madan, Shaurya Saluja, and Aojia Zhao. Automated bitcoin trading via machine learning algorithms. *URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan*, 20, 2015.

[106] Feng Mai, Qing Bai, Zhe Shan, X Wang, and R Chiang. From bitcoin to big coin: The impacts of social media on bitcoin performance. *SSRN Electronic Journal*, 2015.

[107] Tobias P Mann. Numerically stable hidden markov model implementation. *An HMM scaling tutorial*, pages 1–8, 2006.

[108] Roman Matkovskyy and Akanksha Jalan. From financial markets to bitcoin markets: A fresh look at the contagion effect. *Finance Research Letters*, 31:93–97, 2019.

[109] Sean McNally, Jason Roche, and Simon Caton. Predicting the price of bitcoin using machine learning. In *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 339–343. IEEE, 2018.

[110] Ralph C Merkle. Protocols for public key cryptosystems. In *1980 IEEE Symposium on Security and Privacy*, pages 122–122. IEEE, 1980.

[111] Ralph C Merkle. A digital signature based on a conventional encryption function. In *Conference on the theory and application of cryptographic techniques*, pages 369–378. Springer, 1987.

[112] Ruchi Mittal, Rashmi Gehi, and MPS Bhatia. Forecasting the price of cryptocurrencies and validating using ARIMA. *International Journal of Information Systems & Management Science*, 1(2), 2018.

[113] Ziaul Haque Munim, Mohammad Hassan Shakil, and Ilan Alon. Next-day bitcoin price forecast. *Journal of Risk and Financial Management*, 12(2):103, 2019.

[114] Saralees Nadarajah and Jeffrey Chu. On the inefficiency of bitcoin. *Economics Letters*, 150:6–9, 2017.

[115] Satoshi Nakamoto et al. Bitcoin: A peer-to-peer electronic cash system. *Working Paper*, 2008.

[116] Peter Naur. The science of datalogy. *Communications of the ACM*, 9(7):485, 1966.

[117] Peter Naur. "datalogy", the science of data and data processes. In *IFIP Congress (2)*, pages 1383–1387, 1968.

[118] Peter Naur. *Concise survey of computer methods*. Studentlitteratur, 1974.

[119] Peter Naur. Computing versus human thinking. *Communications of the ACM*, 50(1):85–94, 2007.

[120] Peter Naur: Concise survey of computer methods. `http://www.naur.com/Conc.Surv.html`. Accessed: 2019-06-30.

[121] Paolo Pagnottoni and Thomas Dimpfl. Price discovery on bitcoin markets. *Digital Finance*, pages 1–23, 2018.

[122] Sankar K Pal and Paul P Wang. *Genetic algorithms for pattern recognition*. CRC press, 2017.

[123] Theodore Panagiotidis, Thanasis Stengos, and Orestis Vravosinos. The effects of markets, uncertainty and search intensity on bitcoin returns. *International Review of Financial Analysis*, 63:220–242, 2019.

[124] Thearasak Phaladisailoed and Thanisa Numnonda. Machine learning models comparison for bitcoin price prediction. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 506–511. IEEE, 2018.

[125] Giuseppe Antonio Pierro and Henrique Rocha. The influence factors on ethereum transaction fees. In *2019 IEEE/ACM 2nd International Workshop on Emerging Trends in Software Engineering for Blockchain (WETSEB)*, pages 24–31. IEEE, 2019.

[126] Michal Polasik, Anna Iwona Piotrowska, Tomasz Piotr Wisniewski, Radoslaw Kotkowski, and Geoffrey Lightfoot. Price fluctuations and the use of bitcoin: An empirical inquiry. *International Journal of Electronic Commerce*, 20(1):9–49, 2015.

[127] Ian Pollari et al. The rise of fintech opportunities and challenges. *Jassa*, page 15, 2016.

[128] Jack Press. LSTM online training and prediction: Non-stationary real time data stream forecasting. *Wayne State University*, 2018.

[129] Thomas Puschmann. Fintech. *Business & Information Systems Engineering*, 59(1):69–76, 2017.

[130] Python programming language. https://www.python.org. Accessed: 2019-09-10.

[131] The R project for statistical computing. https://www.r-project.org/. Accessed: 2019-09-08.

[132] Shaily Roy, Samiha Nanjiba, and Amitabha Chakrabarty. Bitcoin price forecasting using time series analysis. In *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pages 1–5. IEEE, 2018.

[133] Ed Saiedi, Andres Broström, and Felipe Ruiz Lopez. Global drivers of cryptocurrency infrastructure adoption. *SSRN*, 2019.

[134] John W Schindler. Fintech and financial innovation: Drivers and depth. *Finance and Economics Discussion Series 2017-081. Washington: Board of Governors of the Federal Reserve System*, 2017.

[135] Devavrat Shah and Kang Zhang. Bayesian regression and bitcoin. In *2014 52nd annual Allerton conference on communication, control, and computing (Allerton)*, pages 409–414. IEEE, 2014.

[136] Dehua Shen, Andrew Urquhart, and Pengfei Wang. Does twitter predict bitcoin? *Economics Letters*, 174:118–122, 2019.

[137] Takuya Shintate and Lukáš Pichl. Trend prediction classification for high frequency bitcoin time series with deep learning. *Journal of Risk and Financial Management*, 12(1):17, 2019.

[138] Nico Smuts. What drives cryptocurrency prices?: An investigation of google trends and telegram sentiment. *ACM SIGMETRICS Performance Evaluation Review*, 46(3):131–134, 2019.

[139] Yhlas Sovbetov. Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero. *Journal of Economics and Financial Analysis*, 2(2):1–27, 2018.

[140] Mark Stamp. A revealing introduction to hidden markov models. *Department of Computer Science San Jose State University*, pages 26–56, 2004.

[141] Thomas R Stewart and Claude McMillan. Descriptive and prescriptive models for judgment and decision making: implications for knowledge engineering. In *Expert judgment and expert systems*, pages 305–320. Springer, 1987.

[142] Kejsi Struga and Olti Qirici. Bitcoin price prediction with neural networks. In *RTA-CSIT*, pages 41–49, 2018.

[143] Dian Utami Sutiksno, Ansari Saleh Ahmar, Nuning Kurniasih, Eko Susanto, and Audrey Leiwakabessy. Forecasting historical data of bitcoin using ARIMA and α-sutte indicator. In *Journal of Physics: Conference Series*, volume 1028, page 012194. IOP Publishing, 2018.

[144] Edda Sveinsdottir and Erik Frøkjær. Datalogy — the Copenhagen tradition of computer science. *BIT Numerical Mathematics*, 28(3):450–472, 1988.

[145] The millennial disruption index - viacom media networks. `https://www.bbva.com/wp-content/uploads/2015/08/millenials.pdf`. Accessed: 2019-08-06.

[146] Aviral Kumar Tiwari, Ibrahim Dolapo Raheem, and Sang Hoon Kang. Time-varying dynamic conditional correlation between stock and cryptocurrency markets using the copula-ADCC-EGARCH model. *Physica A: Statistical Mechanics and its Applications*, 535:122295, 2019.

[147] The review of financial studies. `https://academic.oup.com/rfs`. Accessed: 2019-08-06.

[148] John Tukey. *Exploratory Data Analysis.* Addison Wesley, 1977.

[149] John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[150] The global Findex database 2017. `https://globalfindex.worldbank.org/`. Accessed: 2019-08-07.

[151] Andrew Urquhart. The inefficiency of bitcoin. *Economics Letters*, 148:80–82, 2016.

[152] Gulin Vardar and Berna Aydogan. Return and volatility spillovers between bitcoin and other asset classes in Turkey. *EuroMed Journal of Business*, 2019.

[153] Robert Viglione. Does governance have a role in pricing. *Cross-Country Evidence from Bitcoin Markets.[online] Available at: https://papers. ssrn. com/sol3/papers. cfm*, 2015.

[154] Dharminder Singh Virk. Prediction of bitcoin price using data mining. *National College of Ireland*, 2018.

[155] Gang-Jin Wang, Chi Xie, Danyan Wen, and Longfeng Zhao. When bitcoin meets economic policy uncertainty (EPU): Measuring risk spillover effect from EPU to bitcoin. *Finance Research Letters*, 2018.

[156] Joe Whittaker. *Graphical models in applied multivariate statistics.* Wiley Publishing, 2009.

[157] Chih-Hung Wu, Chih-Chiang Lu, Yu-Feng Ma, and Ruei-Shan Lu. A new forecasting framework for bitcoin price with LSTM. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 168–175. IEEE, 2018.

[158] Wu, j.: Statistics = data science? `https://www2.isye.gatech.edu/~jeffwu/`. Accessed: 2019-07-08.

[159] Shuyue Yi, Zishuang Xu, and Gang-Jin Wang. Volatility connectedness in the cryptocurrency market: Is bitcoin a dominant cryptocurrency? *International Review of Financial Analysis*, 60:98–114, 2018.

[160] Liudmila Zavolokina, Mateusz Dolata, and Gerhard Schwabe. Fintech–what's in a name? *Thirty Seventh International Conference on Information Systems*, 2016.

[161] Lina Zhou, Shimei Pan, Jianwu Wang, and Athanasios V Vasilakos. Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237:350–361, 2017.

[162] Ahmed Zouhair, Noah Kasraie, et al. Disrupting fintech: Key factors for adopting bitcoin. *Business and Economic Research*, 9(2):33–44, 2019.