UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

TRADING NATURAL GAS FUTURES THROUGH SIMULATION PREDICTIVE MODELING AND OPTIMIZATION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

AMINE KAMALI
Norman, Oklahoma
2016

TRADING NATURAL GAS FUTURES THROUGH SIMULATION PREDICTIVE MODELING AND OPTIMIZATION

A DISSERTATION APPROVED FOR THE
SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY

Dr. Floyd Grant, Chair

Dr. Robert Dauffenbach

Dr. Kash Barker

Dr. Shivakumar Raman

Dr. Theodore Trafalis

To my family:
My mother Nezha, my father Hassan and my sisters Hasnaa and Fatima Zahraa

Acknowledgements

It is extremely difficult for me to put into words my deepest appreciation for all the amazing people who have helped me and encouraged me during my time in graduate school at the University of Oklahoma.

I am grateful for my family, my father Hassan, my mother Nezha, and my sisters Fatima Zahraa and Hasnaa. Your love and moral support have been tremendous.

I am especially grateful to my advisor, Dr. Hank Grant for his help, encouragement, kindness and responsiveness throughout the duration of my research. Without his guidance and persistent help, the timely completion of this dissertation would not have been possible.

I would like to gratefully acknowledge the members of my committee for their help and encouragement through my research effort. Dr. Kash Barker, Dr. Robert Dauffenbach, Dr. Shivakumar Raman, and Dr. Theodore Trafalis are thanked for their assistance and comments.

I would also like to express my deepest appreciation to Dr. Randa Shehab for her assistance and support throughout my graduate career.

I would also like to thank Onix Solutions Limited, especially Mr. Dan Pritchard, for providing the data used in this research.

Finally I would like to thank the School of Industrial and Systems Engineering at the University of Oklahoma, its former and current faculty and staff, graduate and undergraduate students, my colleagues and classmates whom I consider like a big family who always showed understanding and support.

Table of Contents

ACKNOWLEDGEMENTS IV

TABLE OF CONTENTS V

LIST OF TABLES VII

LIST OF FIGURES VIII

ABSTRACT X

CHAPTER 1: INTRODUCTION 1

- 1.1. Motivation of the Research 3
- 1.2. Objectives of the Research 10
- 1.3. Contribution of the Research 11
- 1.4. Organization of the Dissertation 12

CHAPTER 2: LITERATURE REVIEW 13

- 2.1. Natural Gas Overview 13
- 2.2. Financial Markets Overview 24
- 2.3. Limit Order Book Modeling 37

CHAPTER 3: NATURAL GAS FUTURES TRADING MODEL 50

- 3.1. Overview 50
- 3.2. Simulation 55
- 3.3. Predictive Modeling 63
- 3.4. Optimization 65

CHAPTER 4: SOLUTION APPROACH 70

- 4.1. General Framework 70
- 4.2. Heuristic Approach 72

CHAPTER 5: EXPERIMENTAL RESULTS 75

- 5.1. Simulated Limit Order Book 75
- 5.2. Model Performance 78
- 5.3. Trading Strategies 83

CHAPTER 6: CONCLUSION AND FUTURE WORK 90

REFERENCES 92

APPENDIX A: APPLICATIONS TO THE STOCK MARKET 104

Message file 104

Order book file..105

List of Tables

Table 1: Natural gas ticker symbol for futures delivery dates 20

Table 2: Forward Contracts vs. Futures Contracts 22

Table 3: A sample natural gas futures contract specifications.. 34

Table 4: Order processing algorithm 54

Table 5: Simulation parameters 60

Table 6: Simulation events.. 61

Table 7: A sample of the simulated order book for natural gas 75

Table 8: Description of the predictors 79

Table 9: Training and prediction time performances.. 80

Table 10: SVM model parameters 81

Table 11: Confusion matrix 81

Table 12: Confusion matrix for the SVM model 82

Table 13: Confusion matrix for the DT model 82

Table 14: Input for the optimization model 87

Table 15: Output for the optimization model 87

Table 16: Comparison between three trading strategies 88

List of Figures

Figure 1: Global energy demand by sector and fuel 3

Figure 2: HFT in action when the storage data hit... 7

Figure 3: NG futures prices when the storage data hit in May 7, 2015 8

Figure 4: NG futures volumes when the storage data hit in May 7, 2015 9

Figure 5: Oil and gas value chain 13

Figure 6: Energy consumption in the US by fuel, 1980-2040 14

Figure 7: US total dry natural gas production in four cases, 2005-40 15

Figure 8: US shale gas production in four cases, 2005-40 15

Figure 9: Natural gas transportation network 16

Figure 10: Natural gas storage facilities in the U.S. 17

Figure 11: Natural gas use 18

Figure 12: Natural gas front month futures contract prices 21

Figure 13: Natural gas futures contract 2 prices 22

Figure 14: Factors impacting natural gas spot price 24

Figure 15: A simplified view of buy-side trading... 39

Figure 16: The shape of the LOB 52

Figure 17: Trading System Structure... 53

Figure 18: Natural Gas Futures Traded Volumes.. 55

Figure 19: Simulation of a renewal process with 6 random distributions 58

Figure 20: Time between arrivals for the superposed distributions................................ 59

Figure 21: SVM prediction procedure 63

Figure 22: Overview of the proposed approach.. 74

Figure 23: NG futures bid and ask prices in February 7, 2015 77

Figure 24: NG futures bid and ask volumes in February 7, 2015 78

Figure 25: Profits obtained from different values of k (number of events) 84

Figure 26: Monte Carlo simulation example 86

Abstract

For many years, natural gas prices were strongly correlated with those of crude oil. Recently, natural gas prices started to show an independent trend. Natural gas prices are driven by the law of supply and demand which is reflected by the weather and inventory levels among other factors. In the last decade, electronic trading platforms took over the exchanges. With the advent of algorithm trading (AT) and in particular high-frequency trading (HFT), trading commodities, which include energy trading, became riskier due to their extremely volatile nature.

This dissertation presents a novel framework that provides insight into the use of HFT in natural gas futures markets. Since there are no publicly disclosed data on such practices, the objective is to develop a comprehensive model for natural gas futures trading. A new heuristic simulation, predictive modeling and optimization algorithm that automates trading natural gas futures is proposed and evaluated. Simulation is used to reconstruct the order book using top of the book natural gas futures historical data. Predictive modeling techniques based on multi-class support vector machines are used to predict the occurrence and the amplitude of spread crossings. Finally, an inventory optimization model is used to determine optimal trading volumes for each trading period. Two types of trading strategies are derived: a strategy using Immediate-Or-Cancel orders where an order is totally or partially executed while the remaining is cancelled, and a strategy that limits orders' cancellation.

Both strategies are tested with real and synthetic data. In this setting, both strategies can lead to profit. This could be used by policymakers and market regulators

to implement order cancellation restrictions on commodity futures trading to prevent harmful speculation.

Chapter 1: Introduction

Imagine you are a commodity trader, trading front month natural gas futures contracts on Wednesday, April 27th, 2016. This day happens to be the last trading day, since natural gas futures contracts expire three business days prior to the first calendar day of the delivery month. You are sitting in front of your computer, waiting for the right opportunity to close the deal. A tornado warning has been issued by the weather center, but you are not worried, your office is located in the basement of a very safe building. Then… the power goes off and your internet connection shuts down. You have no choice but to wait. When the storm has passed, you realize that the current trading day is over, and now, you are expected to have a physical delivery of 500,000 million British thermal units (MMbtu) of natural gas. This is a very unlikely but possible scenario that any natural gas futures trader can face. That is why, before focusing on natural gas futures trading, it is important to know the ins and outs of the entire natural gas supply chain.

It is easy to acknowledge that our society is becoming increasingly dependent on energy. Indeed energy became essential in many areas such as transportation, heating, power generation and many other aspects of our lifestyles. Since the beginning of the industrial era, fossil fuels took a very important place in our consumption as they have been the major source of the energy that we use. This energy comes from fossil organic matter that the earth takes billions of years to produce at certain depths, where favorable temperature and pressure conditions are met. This energy mainly exists in three forms which coincide with the three states of matter: oil (liquid), natural gas (gas) and coal (solid). The fossil energy has quickly been exploited on a large scale because of its usability and inexpensiveness compared to other energy sources.

Given the time that takes to naturally produce fossil fuels, it is clear that it is now used at a staggering speed. According to BP Statistical Review 2012, “2011 was an unusually eventful year in global energy. The tumultuous events of the ’Arab Spring‘ shook energy markets and underscored the importance of maintaining spare capacity and strategic stockpiles for dealing with supply disruptions. The earthquake and tsunami in Japan was a humanitarian disaster; and one with immediate implications – in Japan and around the world – for nuclear power and other fuels. Oil prices hit an all-time record high. Yet the revolution in shale gas production drove US natural gas prices lower, reaching record discounts to oil” (Dudley, 2012). World natural gas consumption grew by just 0.4%, well below the 10-year average of 2.4%. Growth was below average in both the OECD and emerging economies, with consumption in the EU (-11.6%) experiencing its largest volumetric and percentage declines on record. Globally, natural gas accounted for 23.7% of primary energy consumption (Dudley, 2015). Thus, in a world facing an alarming situation of exhaustible resources, we turn more and more towards renewable energies such as solar and wind energy. In addition, given the environmental issues, we are beginning to understand the importance urge to limit our impact on the planet, by reducing greenhouse gas emissions. For this purpose, natural gas can provide a transitional solution between our current consumption pattern, based on fossil fuels, and a "carbon-free" way of consumption. Indeed, natural gas is the fossil fuel that releases the least carbon dioxide and pollutants (such as NOx and SOx) and is therefore often regarded as comparatively less polluting (even if it has high methane emission rates). In addition, the proven reserves of natural gas are more important than those of oil. The objective is not to rely on natural gas for our energy supply as it remains an exhaustible resource, but rather, use it as a temporary source for a transition towards a wide use of renewable energies.

Figure 1 below shows the forecasted evolution of the world energy consumption until 2035.

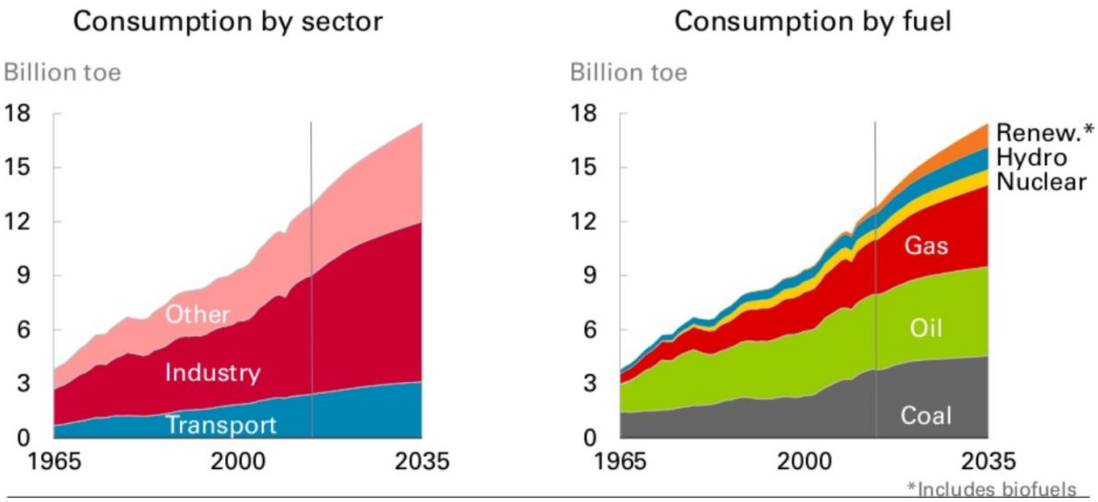


Figure 1: Global energy demand by sector and fuel
(Source: BP Energy Outlook 2035 report (February 2015))

1.1. Motivation of the Research

Natural gas is one of the most volatile commodities that are being traded in the market (Hecht, 2016). Natural gas is traded in two types of markets. The spot market and the futures market. Natural gas is traded for immediate delivery in the spot market. The futures market is where natural gas contracts can be traded in advance (between 1 month and 36 months).

One of the aspects that triggered my curiosity for natural gas trading is that it has an intrinsic value compared to the regular stock trading. Actually it has two interrelated trading components: physical and financial. The former involves an actual delivery of the commodity while the latter involves financial products in which the physical delivery does not take place. It is estimated that the volumes traded in the financial market are about 10 times greater than the ones traded in the physical market. Only a very small portion (about 2%) of the futures contracts end up in deliveries.

The interest for this research comes from the fact that regular traders use the same trading platforms that high frequency traders use. So when it comes down to executing their buy/sell orders, regular traders will be at the mercy of high frequency traders who will scratch every penny out of their trade. Thus the term “banging the beehive” (Dicolo, Rogow, 2012).

Actually, high frequency trading firms act as intermediaries who thanks to their sophisticated algorithms and powerful computers located as close as possible to the exchanges, have the monopole of making profit with almost no risk.

Following this idea, it appeared that commodities are the most sensitive to high frequency trading. In 2008, speculation and high frequency trading practices were accused to have caused the 2008 food bubble. In the world summit of Food Security (November 2009), the Food and Agriculture Organization of the United Nations (FAO) proclaimed the need “to address the issue of speculation in agricultural markets given the serious implications it can have for world food security”. It asked for “in-depth and comprehensive studies to analyze the causal link between speculation and agricultural commodity price movements, with a view of fostering a coherent and effective policy response in the context of food security (Cafiero, 2009).

In the nineties, the US Securities and Exchange Commission (SEC) started allowing electronic markets to compete with traditional market places. Then in the early 2000s, big trading firms started creating black-box algorithms that would implement automated intelligent trading strategies. This would allow algorithms to send millions of orders within seconds. In May 6, 2010, these practices caught the attention of the media when E-mini S&P 500 stock index futures contracts trades made the market disappear, leading to a complete shutdown of the market for five minutes. It took the regulatory commissions FERC and SEC over a year to come up with a report explaining what happened that day. That report concluded that a large order created a domino effect and led to market participants to withdraw from the market. The reaction time of the regulation agencies reflects their limited resources. While big trading firms spend millions if not billions of dollars improving their speed and trading strategies, regulators who are financed by tax payers’ money can only sit and watch those trades being executed at the speed of light. On April 21, 2015, nearly five years after the incident, the U.S. Department of Justice laid "22 criminal counts, including fraud and market manipulation" (Brush, Schoenberg and Suzi, 2015) against Navinder Singh Sarao, a trader. Among the charges included was the use of spoofing algorithms; just prior to the Flash Crash, he placed thousands of E-mini S&P 500 stock index futures contracts which he planned on canceling later (Brush, Schoenberg and Suzi, 2015). These orders amounting to about "$200 million worth of bets that the market would fall" were "replaced or modified 19,000 times" before they were canceled (Brush, Schoenberg and Suzi, 2015). Spoofing, layering and front-running are now banned (Bates, 2015).

In 2009, HFT accounted for 60-73% of all US equity trading volume, with that number falling to approximately 50% in 2012 (Lati, 2009) (Popper, 2012). One of the first reported market manipulation cases took place in October 2014. The US regulators fined Athena Capital Research LLC $1 million for price manipulation. The HFT industry was then “besieged by accusations that it cheats slower traders” (Geiger and Mamudi, 2014).

Trading became completely automated and stripped from any direct human decision making. In the blink of a human eye, millions of trades could be executed. Nowadays, in an ultra-low latency environment, trading strategies based on machine learning are extensively used by traders and brokers. Those strategies are proprietary and hidden to the public. The figure below shows an example of how HFT is used for trading natural gas futures.

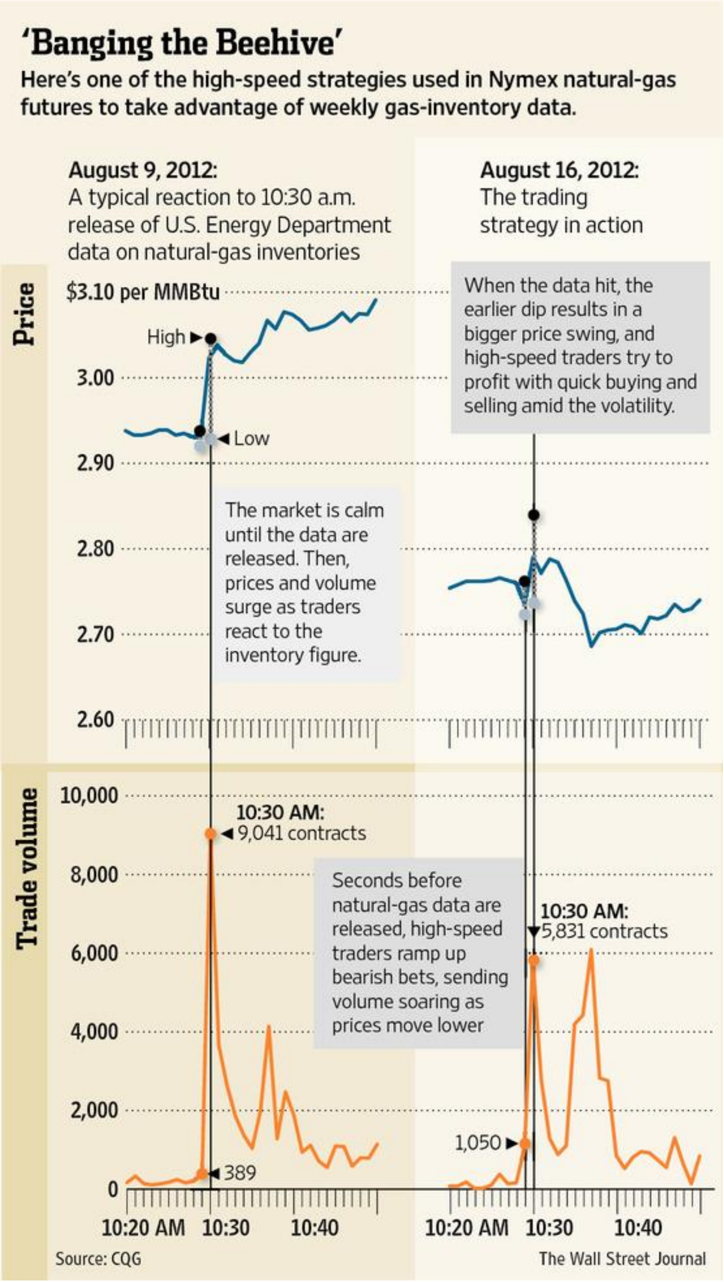


Figure 2: HFT in action when the storage data hit

(Source: The Wall Street Journal (August 2012))

On August 9, 2012 at 10:30 a.m. EST, the U.S. Energy Information Administration (EIA) reported an inventory increase of 25 billion cubic feet, slightly below analysts' expectations, which normally would translate into a modest price rise. In the first seconds after the release, natural-gas futures surged more than 10 cents to

$2.84, but then immediately ricocheted lower. The fastest traders had the chance to profit from rapid buying and selling over a much wider price range as the market slowly found some equilibrium. Over the next seven minutes, futures dropped to a low of $2.685. Prices didn't hit those highs or lows for the rest of the session (Dicolo, Rogow, 2012).

The HFT algorithms can consistently be seen in action every Thursday around 10:30 a.m. EST (9:30 a.m. CST). Figure 3 shows another instance where HFT is used when natural gas inventory levels are released.

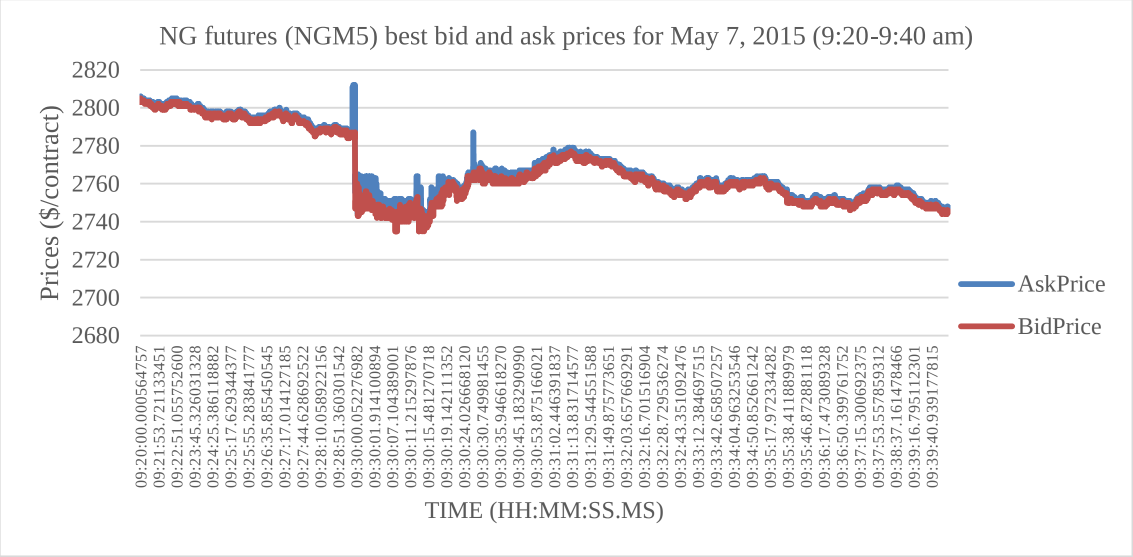


Figure 3: NG futures prices when the storage data hit in May 7, 2015

Right when the data hit, the best ask price goes up to $2812/contract, before immediately dropping down to $2741/contract, which represents a difference of $71/contract within few seconds. Faster traders can profit from the large price swings and make the spread. Figure 4 below shows how the best bid and ask volumes fluctuate when the inventory data hit in May 7, 2015.

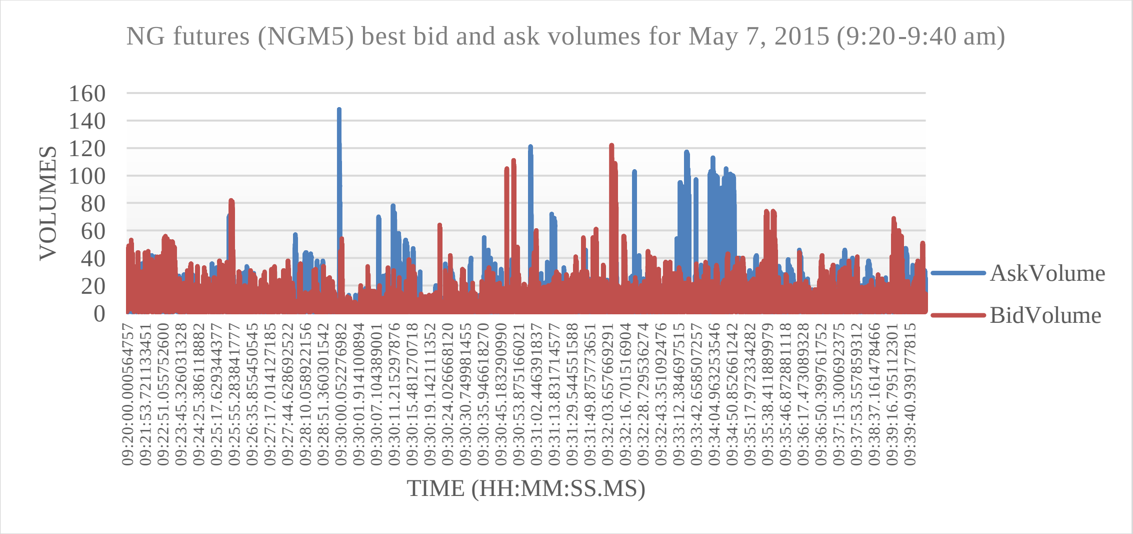


Figure 4: NG futures volumes when the storage data hit in May 7, 2015

Using machine learning in HFT poses some important challenges. For instance it is critical to use high quality datasets with a great level of detail, also referred as granular data. In the context of HFT, the “tick-to-trade” latency, which is the time from events arriving as inputs to an application to that application output, shrunk considerably within the last few years. It went from seconds, to a fraction of seconds, to a fraction of milliseconds. Trading firms actually keep their speed secret but invest a lot of money in computational power and co-location in order to reduce their latency.

HFT is viewed as good and bad for the markets. On the one hand it creates liquidity, but on the other hand, an algorithm can go wrong and plunge the market similar to what happened in the flash crash of May 6, 2010.

The limit order books were first introduced in the stock exchanges, then spread gradually to include commodities. This change affected the market players. In the past, a NYMEX floor trader would have an educated decision about the price movement, before implementing trading strategies. Now, with the advent of high frequency trading, market makers use algorithms that have a tick to trade time in the order of micro-seconds, and take advantage of the high price volatility during specific events (such as the release of the weekly inventory levels of natural gas) to make the spread, without using any prior knowledge of the market conditions.

1.2. Objectives of the Research

There are many factors that affect the dynamics of the limit order book. These factors can be random such as the arrival of market orders that will be executed immediately and will change the state of the top of the book (best ask and best bid prices and volumes). Other factors, such as market depth, will also play a role in updating the state of the top of the book. Pending orders will eventually become active at some point in time if they are not cancelled. Thus the importance of studying the effect of pending orders on the future state of the limit order book.

Currently, in-depth market trading data are proprietary. Un-fulfilled buy/sell orders are partially reflected in the data, but are essential to evaluate performance. Naturally, trading companies do not expose their trading algorithms. Thus the need of developing a set of tools to deal with missing/limited historical data and to represent the use of proprietary algorithms is clear.

In this research, we use historical data of the top of the book to recreate the shape of the limit order book for pending orders. We use simulation to recreate up to level 10 bid and ask prices and volumes. We then use multi-class support vector machines to predict the direction and the amplitude of the spread crossing over time. We train the models on past data, and then test the models on upcoming data. Two types of trading strategies are derived: a strategy using Immediate-Or-Cancel orders where an order is totally or partially executed while the remaining is cancelled, and a strategy that limits orders cancellation by keeping track of the total inventory on hand.

Both strategies are tested with real and hybrid (simulated and historical) data. In this setting, both strategies can lead to profit. A linear programming inventory optimization model is used to determine the volumes to buy and sell natural gas futures, while keeping the inventory under a given level.

1.3. Contribution of the Research

This research focuses on developing a set of tools for evaluating the use of high frequency trading for natural gas futures in the absence of detailed historical data and developing modeling techniques to satisfy this need.

The research contributions are five-folds:

1. Help the community understand natural gas futures limit order book dynamics in a high-frequency environment by developing new and unique modeling tools to address the lack of access to proprietary information.
2. The simulation framework suggested in this dissertation generates limit order book data using top of the book historical data. These hybrid data mimic real world data and can be modified to create multiple scenarios reflecting less common events. These data are used to build robust learning models.
3. A set of novel labels such as spread crossing amplitude are introduced and used to design optimal trading strategies.
4. This dissertation implements an automated trading framework that can predict effectively and efficiently metrics of natural gas futures limit order book dynamics by employing multi-class SVM techniques on a set of novel labels and

predictors such as spread crossing amplitude. This automated framework yields to profit.

5. A linear programming inventory optimization model is efficiently used to derive novel optimal trading strategies limiting orders' cancellations while leading to profitable trades.

The proposed framework can easily be extended to incorporate other machine learning techniques such as logistic regression, decision trees, genetic algorithms and artificial neural networks. The proposed models could help enhance the understanding of the limit order book for financial products other than natural gas futures.

1.4. Organization of the Dissertation

Chapter 2 will be devoted to reviewing the literature related to natural gas futures' trading and limit order book modeling. It serves to delineate the research and development discussed in Chapter 3 from previous related work. In Chapter 3, we will discuss the procedures which have been developed for natural gas futures trading and the related literature in the area. Chapter 4 presents the solution approach, Chapter 5 presents the experimental results of the models built in the proposed framework and finally Chapter 6 provides conclusions concerning this research as well as recommendations for further work.

Chapter 2: Literature Review

In this chapter, an overview of natural gas, financial markets, and limit order book modeling is discussed.

2.1. Natural Gas Overview

The process of getting natural gas from the ground and into everyday life is actually an extensive and complex process. This section provides an overview of the process, from exploration to marketing of the natural gas that will be sold for home and industry usage. Figure 5 below gives an overview of the oil and gas value chain.

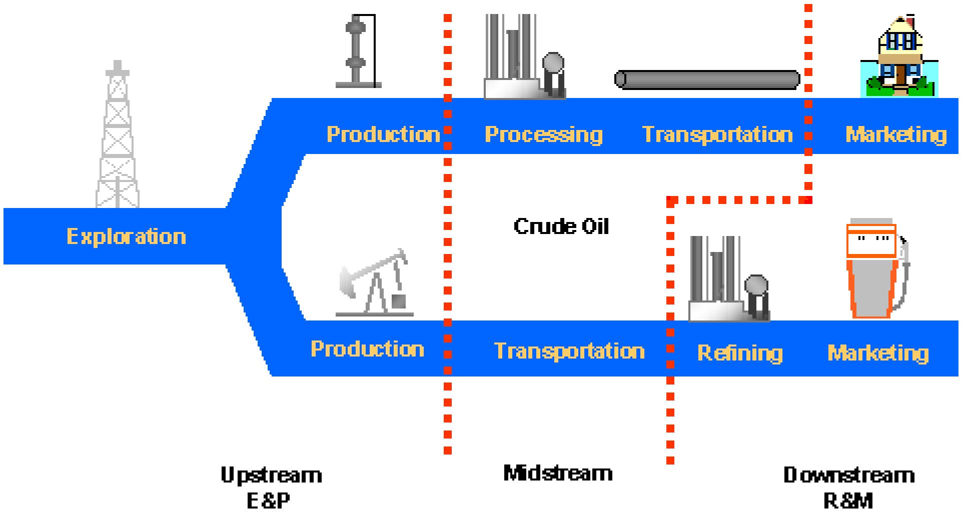


Figure 5: Oil and gas value chain
(Source: PetroStrategies, Inc., 2013)

Exploration and extraction

Natural gas exploration and extraction processes are often associated with crude oil. Both use similar techniques and are usually discovered simultaneously. Decision analysis techniques are used to determine which wells are worth drilling (Schuyler and Newendorp, 2013).

Natural gas has traditionally been extracted from natural gas wells. Recently, with the development of shale gas resources, the current U.S. production and the projections for the years to come have reached all-time highs, while the demand is not expected to rise drastically as shown in figure 6.

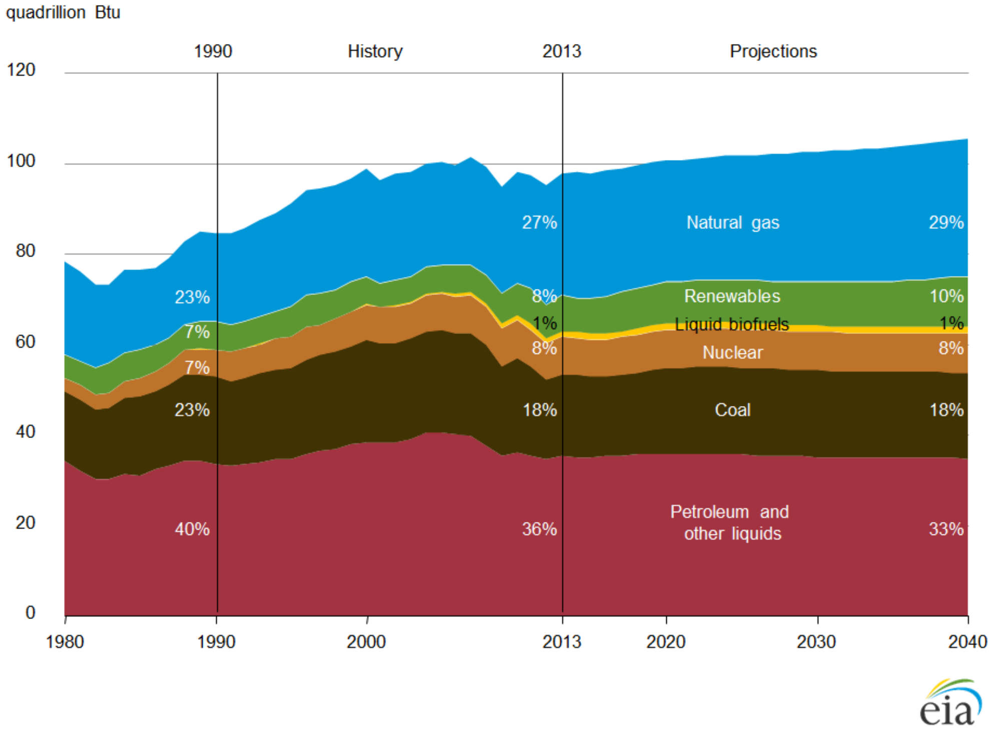


Figure 6: Energy consumption in the US by fuel, 1980-2040

(Source: Energy Information Administration 2015)

While U.S. crude oil production is expected to stabilize or decrease, natural gas production would still be increasing. Figures 7 and 8 below show how U.S. dry gas and shale gas production are respectively projected to increase in the coming years regardless of the oil price movements.

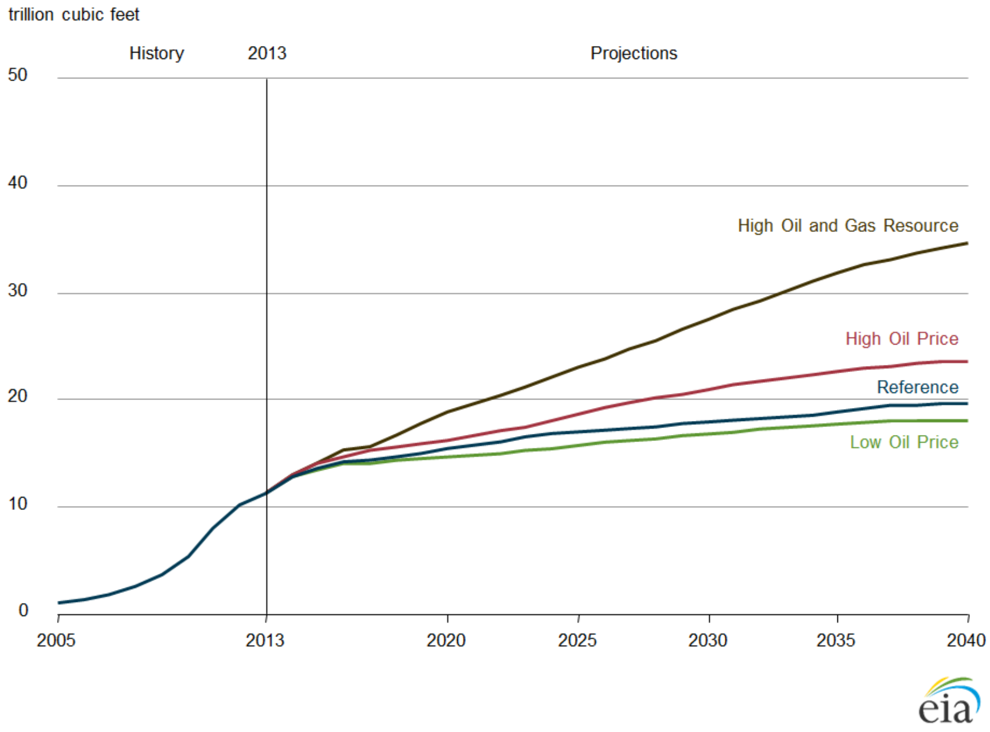


Figure 7: US total dry natural gas production in four cases, 2005-40
(Source: Energy Information Administration 2015)

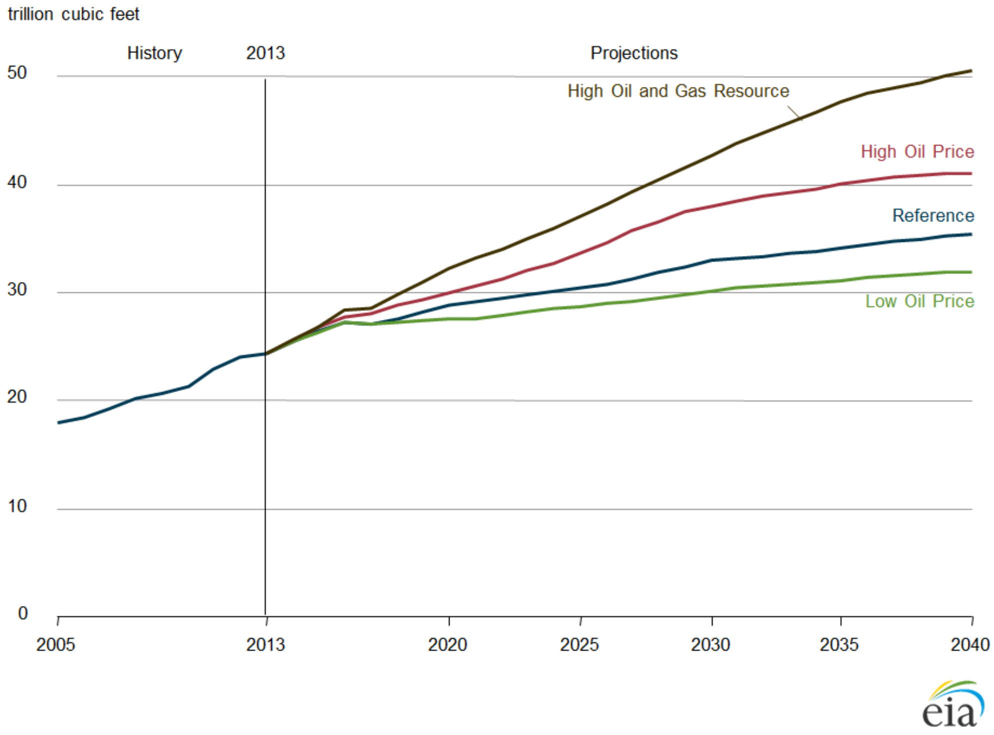


Figure 8: US shale gas production in four cases, 2005-40
(Source: Energy Information Administration 2015)

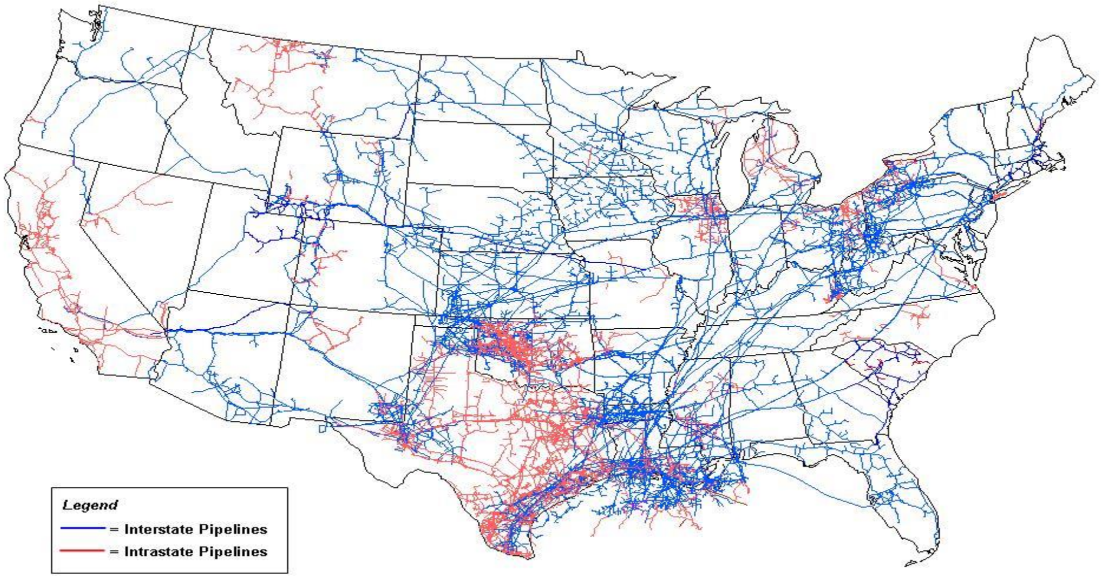
The supply and demand outlook for natural gas will contribute to greater competitiveness of U.S. manufacturing, while the use of more efficient technologies could offset increases in demand and provide cost effective compliance with emerging environmental requirements (Moniz et al., 2011).

Production and Transport

An intricate system of intra-state and inter-state pipelines connects natural gas markets with the production areas.

First, raw natural gas is gathered and processed to remove impurities and extract the liquefied natural gas (Kidnay, Parrish and McCartney, 2011). Then it is transported via a large network of pipelines to the market areas (Mokhatab and Poe, 2012).

Interstate and intrastate pipelines are used to meet the customer's demand. Figure 9 below shows the natural gas transportation network in the U.S.



Source: Energy Information Administration, Office of Oil & Gas, Natural Gas Division, Gas Transportation Information System

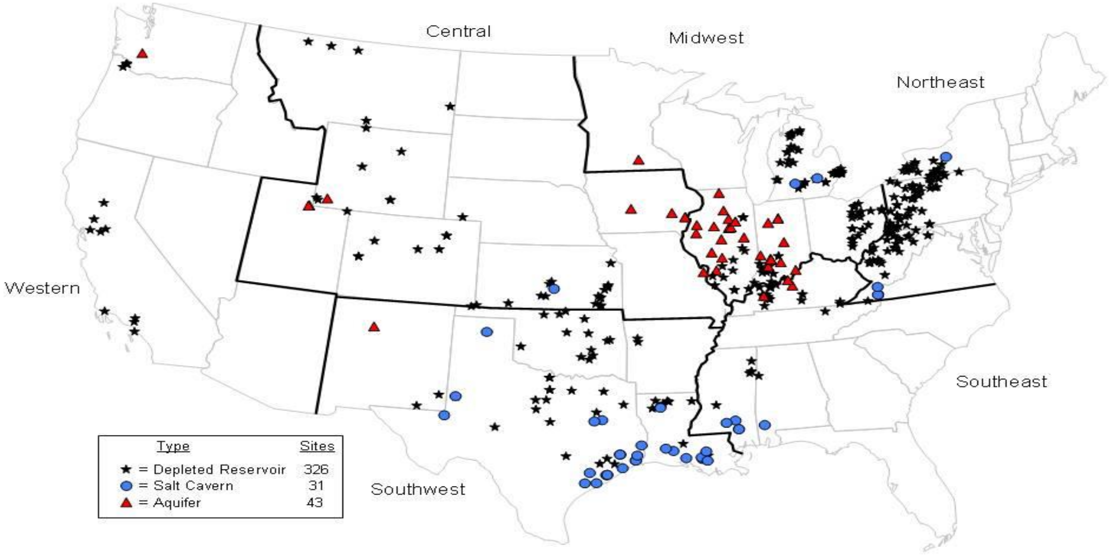
Figure 9: Natural gas transportation network
(Source: Energy Information Administration 2011)

Storage

After it has been refined, natural gas is delivered and stored (Fennell, 2011). Underground storage facilities, fashioned from depleted oil, natural gas, or aquifer reservoirs or salt caverns, are used to store natural gas as a seasonal backup supply (EIA, 2008).

Natural gas storage valuation is an interesting topic on many levels. It requires both appropriate price models and optimization models (Li, 2009). Marketers can move gas in and out of storage in order to capitalize on the changes in price levels as well as in combination with financial instruments such as futures and other option contracts. Thus, natural gas storage has focused on more flexible operations with high deliverability and rapid cycling in their inventories (EIA 2011).

Figure 10 below shows the distribution of different types of storage facilities in the U.S.



Source: Energy Information Administration, Office of Oil & Gas, Natural Gas Division Gas, Gas Transportation Information System, December 2008.

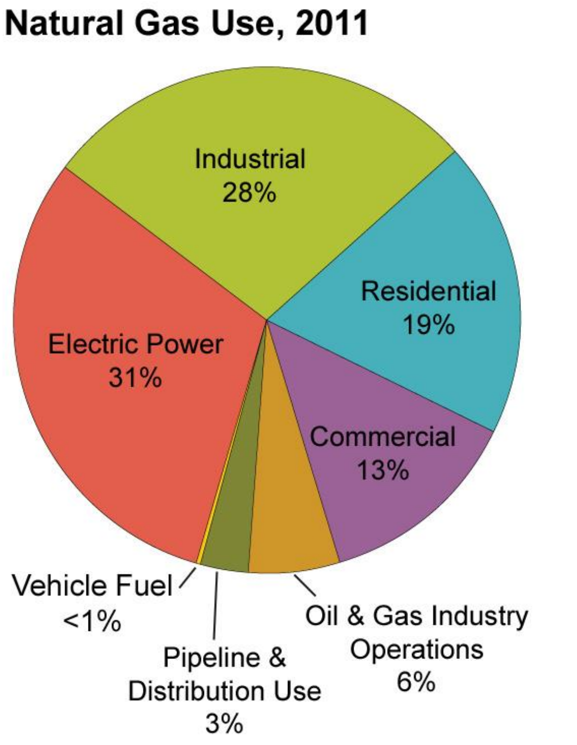
Figure 10: Natural gas storage facilities in the U.S.
(Source: Energy Information Administration, 2008)

Distribution and Marketing

Natural gas is distributed to the end user via local distribution companies. The major consumers of natural gas in the United States in 2011 included:

- Electric power sector — 7.6 trillion cubic feet (Tcf)
- Industrial sector — 6.9 Tcf
- Residential sector — 4.7 Tcf
- Commercial sector — 3.2 Tcf

The pie chart below shows the distribution of the use of natural gas in the U.S by sectors (EIA 2011).



Source: U.S. Energy Information Administration, *Natural Gas Monthly* (April 2012).

Figure 11: Natural gas use
(Source: Energy Information Administration, 2012)

Natural Gas Markets

In the United States, natural gas is traded in the spot and futures markets. Natural gas futures prices are set by the New York Mercantile Exchange (NYMEX) in the daily transactions of the commodity and greatly impacts the activities of marketers. Natural gas prices are also crucial to the problem of natural gas storage valuation where prices dictate the injection and withdrawal of natural gas from storage as speculators attempt to capitalize on market movements.

Natural gas prices are set in market hubs, which are generally found at the intersection of the major pipelines. The largest hub for natural gas in the U.S. is the Henry Hub and it is located in Louisiana.

Natural gas spot market

In the spot market, natural gas is traded for immediate delivery. The traded volumes are set to be paid ($/MMBtu) and delivered on the next business day following the trading day. For example, if the transaction date was Wednesday, February 3, 2016 the settlement date would be Thursday, February 4, 2016.

Natural gas futures market

Natural gas is traded using forward and futures contracts which are widely traded on the New York Mercantile Exchange (NYMEX). One natural gas contract has an energy value of 10,000 MMBtus. Natural gas futures contracts expire three business days prior to the first day of the next month. For example, the February 2016 contract will expire on January 27, 2016. After a contract expiration date has passed, it will have to be delivered following its specifications.

Natural gas futures are delivered every month of the year unlike other commodities such as corn or wheat which are delivered only during specific months. The following is a table with NG futures delivery months and the corresponding ticker symbol.

Delivery Month	Full Ticker Symbol
January, 2015	NGF5
February, 2015	NGG5
March, 2015	NGH5
April, 2015	NGJ5
May, 2015	NGK5
June, 2015	NGM5
July, 2015	NGN5
August, 2015	NGQ5
September, 2015	NGU5
October, 2015	NGV5
November, 2015	NGX5
December, 2015	NGZ5

Table 1: Natural gas ticker symbol for futures delivery dates

The month with the closest delivery is called the front month or the spot month. Figure 12 shows the evolution of front month natural gas futures contract prices over the last 20 years.

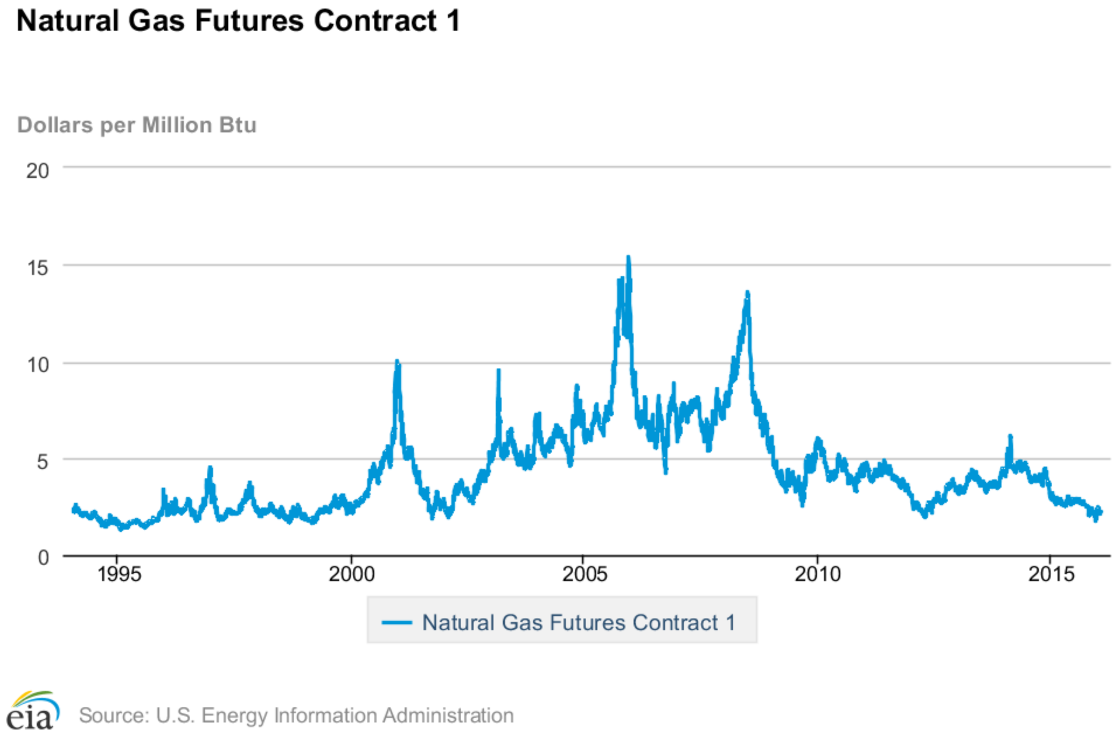


Figure 12: Natural gas front month futures contract prices
(Source: Energy Information Administration 2016)

The term "Contract 1" refers to a futures contract specifying the earliest delivery date. Natural gas contracts expire three business days prior to the first calendar day of the delivery month. Thus, the delivery month for Contract 1 is the calendar month following the trade date (Energy Information Administration, 2016). Similarly, Contract 2-4 represent the successive delivery months following Contract 1.

Figure 13 shows the evolution of natural gas futures contract prices for the month following Contract 1 over the last 20 years.

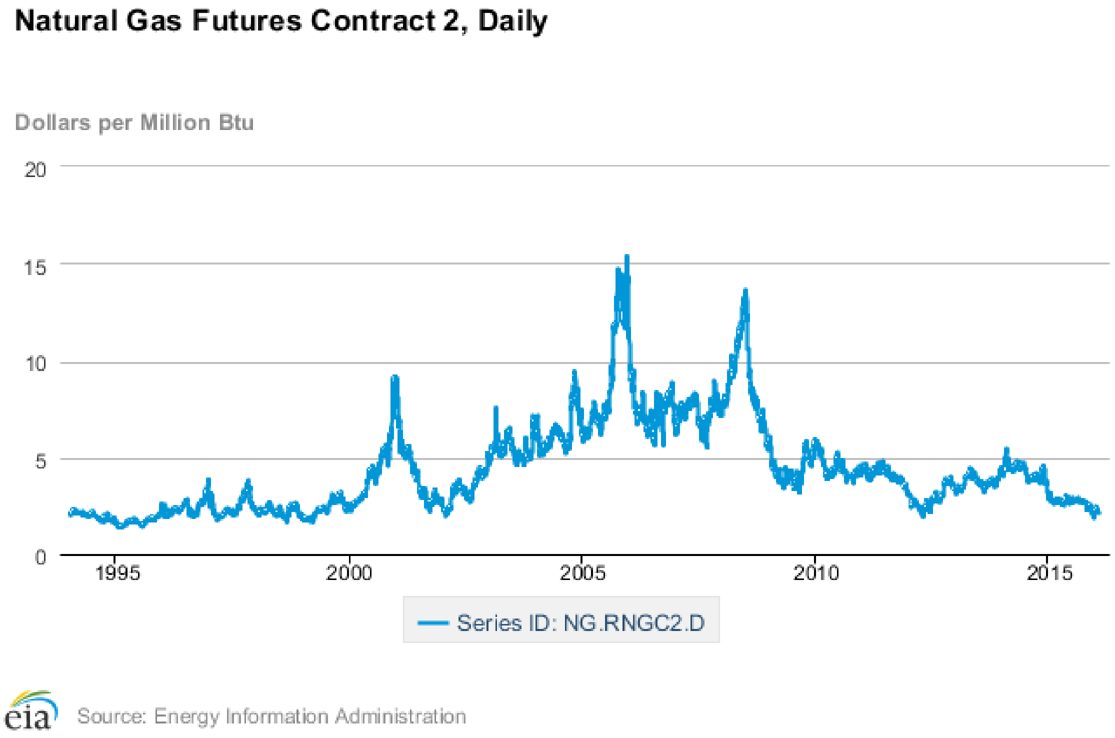


Figure 13: Natural gas futures contract 2 prices
(Source: Energy Information Administration 2016)

The curves for Contract 1 and for Contract 2 follow a similar trend.

The table below describes the differences between futures and forward contracts.

Forward	Futures
Private contract between two parties	Traded on an exchange
Not standardized	Standardized
Usually one specified delivery date	Range of delivery dates
Settled at end of contract	Settled daily
Delivery or final settlement usual	Usually closed out prior to maturity
Some credit risk	Virtually no credit risk

Table 2: Forward Contracts vs. Futures Contracts
(Source: Hull, 2008)

Speculative market services

After 1992, the Federal Energy Regulatory Commission's (FERC) introduced Order 636, requiring pipelines to separate transportation and sales as part of the deregulation of natural gas markets. This has expanded the use of storage from its role as a back-up supply source in times of excess demand to an agent of the financial markets.

Factors Affecting Natural Gas Prices

(Adapted from EIA 2012)

Natural gas prices are a function of market supply and demand. Because of limited alternatives for natural gas consumption or production in the short run, even small changes in supply or demand over a short period can result in large price movements to bring supply and demand back into balance. Factors on the supply side that may affect prices include:

- Variations in the amount of natural gas being produced
- The volume of gas being imported and/or exported
- The amount of gas in storage facilities
- Increases in supply result in lower prices, and vice-versa

Factors on the demand side that may affect prices include:

- The level of economic growth
- Variations in winter and summer weather
- Oil prices
- Higher demand tends to lead to higher prices, while lower demand tends to lead to lower prices.

Most of the natural gas consumed in the U.S. is produced in the U.S. Production increased drastically since the 1970's with the enhancement of drilling techniques and the more recent use of shale gas.

Figure 14 below displays some of the factors that influenced natural gas spot prices in the past.

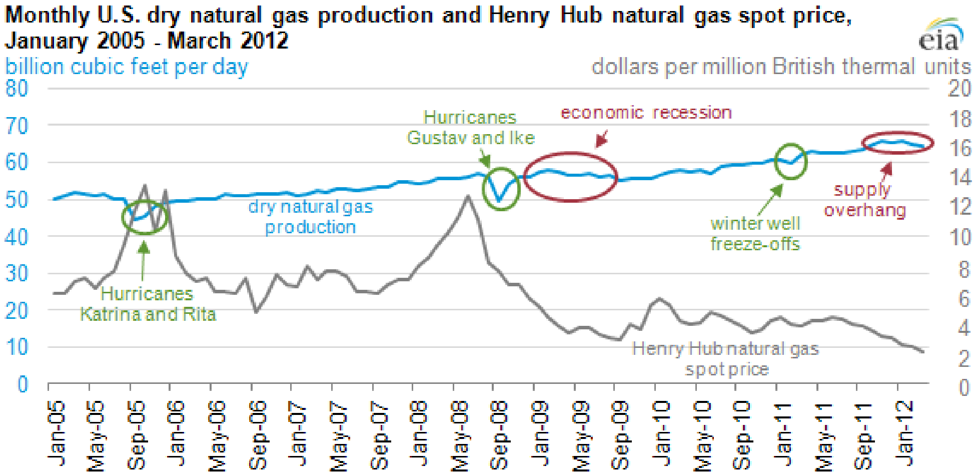


Figure 14: Factors impacting natural gas spot price
(Source: Energy Information Administration, 2012)

Next is an overview of financial markets.

2.2. Financial Markets Overview

Almost all modern financial theory and models are based on assumptions that are not accurate in the real world. From the pricing model of Black-Scholes, to the optimal portfolio model of Harry Markowitz, in order to make the calculations doable, very often, financial modeling is based on what is commonly known as "the bell curve". Indeed, stock prices are often considered to be normally distributed. In his book "The Black Swan", Taleb (2007) claims that "financial markets are from "Extremistan", while the academics think that markets are from "Mediocristan". "Mediocristan is the terrain of the ordinary, the part of the world that conforms to the bell curve. It answers to statistics and knowable probabilities”. Extremistan represents all the outliers or black swans that no one could have predicted. For example, Black Monday which refers to Monday October 19, 1987 when stock market crashed and the Dow Jones index fell by more than 20 percent (Browning, E.S 2007), is considered a black swan that does not fit in any model. This crash was so improbable considering financial models based on standard statistics, that it led financial authorities to reconsider the entire basis of quantitative finance.

Recent decades have witnessed several financial crises that have caused serious economic problems worldwide. The last known example is the recent subprime crisis (subprime mortgage meltdown), which was triggered by a crisis in the market for subprime mortgages in the U.S. It led to a national economic recession and then proliferated to become a global financial crisis from the summer 2007 whose effects continued to be felt until today.

Following such events, the authorities have intervened several times to introduce stricter regulations for financial institutions. They required them to develop internal and most effective risk measures in order to detect failures and adjust them to avoid large losses, bankruptcies and even global crises. In this context, methods for measuring and controlling the risk in financial markets became critical.

Financial markets are structures through which funds flow. They can be distinguished along two dimensions: primary versus secondary markets and money versus capital markets (Bodie, Kane, Marcus, 2009).

Primary markets are the markets in which users of funds (most likely businesses and governments) raise funds by issuing financial instruments (stocks and bonds). While in secondary markets, financial instruments are traded among investors through organized exchanges. Money markets are the markets that trade debt securities with maturities of one year or less (e.g. Treasury bills), while capital markets trade debts (bonds) and equity (stock) with maturities of more than one year (Bodie, Kane, Marcus, 2009). In this context, interest rate is essential. The interest rate is the rate charged for the use of money (borrowing) – a reward to the lender for delaying consumption, but may also include rewards for bearing risk and for inflation. We often talk about nominal interest rates which are the rates actually observed in financial markets. According to Rose P. (2007), “the fundamental level of the interest rate is determined by the interplay of two forces:

- The demand for credit, by domestic businesses, consumers, governments as well as foreign borrowers.
- The supply of loanable funds from domestic savings, money creation by the banking system as well as foreign lending”.

For the subprime crisis, the Federal Reserve (Fed) took actions to reduce the Fed Funds rate sharply after the bursting of the stock market bubble in March 2000. In the eyes of many, the Fed cut rates too far and held them down too long, fueling a housing bubble (Gerald P. O'Driscoll, Jr, 2007). In their article about Subprime Monetary Policy, Gerald P. and O'Driscoll, Jr (2007) stated that “it was clearly enunciated by Greenspan in his December 19, 2002, speech, in which he made an asymmetric argument leading to an asymmetric monetary policy. He argued that asset bubbles cannot be detected and monetary policy ought not in any case to be used to offset them.

The collapse of bubbles can be detected, however, and monetary policy ought to be used to offset the fallout".

After 2007, several "Quantitative Easing" policies have been used worldwide to face the financial crisis. The central bank targets a particular interest rate it wants to reduce, credits its own bank account with money it creates electronically then buys government bonds (including long-term government bonds) or other financial assets, from commercial banks or other financial institutions, largely with the newly created (electronic) money (Bernanke, B. ,2009). The effect is to change the market prices of these bonds and effective interest rates. Ben Bernanke, the current Chairman of the Reserve said: "Our approach—which could be described as "credit easing"—resembles quantitative easing in one respect: It involves an expansion of the central bank's balance sheet. However, in a pure QE regime, the focus of policy is the quantity of bank reserves, which are liabilities of the central bank; the composition of loans and securities on the asset side of the central bank's balance sheet is incidental…. the Federal Reserve's credit easing approach focuses on the mix of loans and securities that it holds and on how this composition of assets affects credit conditions for households and businesses".

The first round of "quantitative easing" by the Fed reduced the short term rate to nearly zero. The second round, popularly known as QE2 was intended to reduce long term rates. A third round of an open-ended quantitative easing, QE3, was announced on September 13, 2012. On 19 June 2013, Ben Bernanke announced a "tapering" of some of the Fed's QE policies contingent upon continued positive economic data. The Federal Reserve ended its monthly asset purchases program (QE3) in October 2014, ten months after it began the tapering process (Sharf, 2014). On 16 December 2015 the Fed increased its key interest rate, the Federal Funds Rate, for first time after June 2006; the hike was from the range [0%, 0.25%] to the range [0.25%, 0.5%] (Gillespie, 2015)

Financial derivatives played an important role in the crisis. One of the oldest derivatives is rice futures, which have been traded on the Dojima Rice Exchange since the eighteenth century (Suzuki and Turner, 2005). Primary financial derivatives are common stocks, bonds, options and futures. A stock is an equity ownership of a company. A bond is a debt security, in which the authorized issuer owes the holders a debt and, depending on the terms of the bond, is obliged to pay interest (the coupon) to use and/or to repay the principal at a later date, termed maturity (O'Sullivan and Sheffrin, 2003). A futures price of a commodity is the price at which one can agree to buy or sell it at a given time in the future. The futures prices adjust to balance demand to buy the commodity in the future with demand to sell the commodity in the future. Whenever a contract is opened, there is someone on each side. The person who agrees to buy is long the commodity, and the person who agrees to sell is short (Fischer, 1975). This means that the long and short positions should cancel out.

Financial markets are supposed to be auto regulated. “In developing pricing models, the notion of arbitrage is very important. For futures contracts, the price must be related to the spot plus some discounting, as the futures price becomes the spot price at maturity” (Hull, 2008).

Prior studies have found that using passive investable commodity futures indices as proxies for direct commodity investments shows good stand-alone performance. Adding commodity futures to a portfolio of stocks and bonds can substantially reduce portfolio risk (Jensen et al., 2000, 2002).

The modern portfolio theory can explain how to optimize the performance of a portfolio by optimizing diversification. To achieve this goal, it is necessary to develop tools for risk management. The optimal portfolio theoretical foundations started in the17th century, but thanks to Harry M. Markowitz who published his first article “Portfolio Selection” in the “Journal of Finance” in 1952, the active risk management was able to develop to become an essential element of portfolio management. Harry Markowitz formalized what investors already knew, when seeking to match the investment returns with the level of risk taken. But he was the first to establish mathematically that the total risk of a portfolio is less than the sum of individual risks of each component of the portfolio. Taking the periodic returns on investments as random variables, it then became possible to calculate the expected return, standard deviations and correlations. In seeking the minimum risk for each level of return, you get what Markowitz called: “the efficient frontier”.

Next is a review of the literature of natural pricing models.

Natural gas pricing models

Various methods and approaches have been developed by the researchers for natural gas pricing; some can be identified as deterministic or stochastic, dynamic or static and linear or nonlinear models (Gutiérrez, A., Nafidi, A., Gutiérrez Sánchez, R., 2005).

Natural gas price prediction methods can be distinguished as traditional (time series) and computational intelligence (CI)-based approaches.

Some authors focused on the spot prices such as Thompson et al. (2003), de Jong et al. (2002) or Boogert et al. (2008). Some others used forward prices. See Eydeland et al. (2003), Blanco et al. (2002) and Gray et al. (2004).

An industry standard for modeling price evolution is the use of geometric Brownian motion (GBM). GBM has been extensively studied, and its properties are well known, including its ability to generate only positive random numbers which is important for financial applications when the numbers represent prices (Eydeland and Wolyniec 2003). Though excellent for a variety of financial applications, geometric Brownian motion in its standard form is ill suited to describe the evolution of energy prices. Particularly, the natural gas market requires a model which describes the presence of mean reversion within the spot prices. This attribute is present due to the pressures of supply and demand on the price caused by the seasonal use of natural gas. High demand during the winter months will put upward pressure on the price of natural gas, and similarly the limited demand and thus increased supply of natural gas during the spring will place downward pressure on the price. Amongst these high and low swings, the price is centered on an average that the market wants to regress to as it moves further away.

Stochastic models are also in the literature. For instance Cartea et al. (2005) developed mean-reversion jump-diffusion models to describe spot prices of energy commodities.

Recently, data mining techniques spread to uncover hidden patterns in the data (Witten and Frank, 2005). For instance Kimoto, Asakawa and Takeoka (1990) used artificial neural networks (ANNs) to predict the Tokyo stock exchange index. Ince and

Trafalis, Ince and Mishina (2003) used support vector machines and focused on support vector regression (SVR) for option pricing. Trafalis and Ince (2008) compared SVR, ANN and radial based function (RBF) networks for pricing stocks. Holland (2008) also used SVR for natural gas pricing.

There are many uncertain factors influencing natural gas consumption which make gas consumption series highly complex and nonlinear (Sanchez-Ubeda and Berzosa, 2007). Therefore, traditional linear models and statistical approaches such as linear regression or the method one proposed by (Gutierrez, Nafidi and Gutierrez Sanchez, 2005), are not suitable for gas consumption prediction. Computational intelligence (CI) based models, including fuzzy logic, neural networks (NN) and support vector machines (SVM) are elaborate models which are effective in dealing with highly nonlinear and complex processes (Jang, Sun and Mizutani, 1997). The CI-based models have been used for energy demand predictions to a great extent (Hippert, Pedreira, and Souza, 2001). Prediction of daily natural gas consumption by combination of artificial neural-network forecasters has been also carried out (Khotanzad, Elragal and Lu, 2000). In this study, Khotanzad et al. proposed a two-stage system with the first stage containing two NN forecasters. The second stage consisted of a combination module to mix the two individual forecasts produced in the first stage. They implemented their approach on real data from six different gas utilities. Support vector machines, established based on the statistical learning theory, exhibit distinctive advantages to solve complex problems (Suykens, Van Gestel, De Brabanter, De Moor and Vandewalle, 2002.). Another methodology based on the work of Boogert and de Jong (2008), uses Monte Carlo simulation with ordinary least square regression.

Bringedal (2003) developed a model using stochastic dual dynamic programming. These methods can incorporate bid and ask prices.

Murry and Zhu (2004) found evidence that the introduction and demise of EnronOnline coincided with the improvement and worsening in the degree of the market informational efficiency.

High-frequency traders (HFT) are playing a major role in price discovery and price efficiency (Brogaard, Hendershott and Riordan, 2012). Hendershott, Jones, and Menkveld (2011) show that algorithmic trading (AT) improves liquidity and makes quotes more informative. Boehmer, Fong, and Wu (2012) provide international evidence on algorithmic trading in equity markets. Chaboud, Chiquoine, Hjalmarsson, and Vega (2009) relate AT to volatility and find little relation. Hendershott and Riordan (2012) focus on the monitoring capabilities of AT and study the relationship between AT and liquidity supply and demand dynamics. Hasbrouck and Saar (2010) studied low-latency trading and found that increased low-latency trading is associated with improved market quality. Biais, Foucault, and Moinas (2011) and Pagnotta and Philippon (2011) provide models where investors and markets compete on speed. Foucault, Hombert, and Rosu (2015) examine a model where a trader receives information one period ahead of the rest of the market.

According to Ross (2012), HFT using a range of strategies target natural gas because of the wide gaps between the prices offered to buy and sell the futures contracts, which can lead to easier profits. HFT firms are fighting to fend off regulation as scrutiny of their practice of unleashing blizzards of orders coincides with repeated technical glitches in the markets. As the firms work to convince policy makers their practices are benign or even beneficial, one of their primary tools has been research seeded by the industry itself, promoted by lobbying that has increased in recent years (Strasburg and Patterson, 2012).

Natural gas futures trading

The following is an overview of the mechanics of the natural gas futures markets.

Futures trading including natural gas trading is not suitable for all investors, and involves the risk of loss. Futures are a leveraged investment, and because only a percentage of a contract’s value is required to trade, it is possible to lose more than the amount of money deposited for a futures position. Therefore, traders should only use funds that they can afford to lose without affecting their lifestyles. And only a portion of those funds should be devoted to any one trade because they cannot expect to profit on every trade (CME group, 2015).

In futures markets, traders deal in standardized futures contracts only. Below is an example of a natural gas futures contract:

Natural Gas Contract Specifications	
Ticker Symbol	Open Outcry: NG (NYMEX) Electronic: ENG (eCBOT)
Contract Size	10,000 million British thermal units
Deliverable Grades	Pipeline specifications in effect at time of delivery
Contract Months	All months
Trading Hours	NYMEX Open Outcry: Monday-Friday9am-2:30pm EST eCBOT Electronic: Sunday-Friday 6pm-5:15pm CST
Last Trading Day	Trading terminates three business days prior to the first calendar day of the delivery month.
Last Delivery Day	Last business day of the contract month
Price Quote	Cents per million Btu (MMBtu)
Tick Size	NYMEX: .1 cents per MMBtu ($10/per contract)

Table 3: A sample natural gas futures contract specifications
(Source: Investopedia.com)

Traders can buy and sell natural gas futures contracts without owing them. They do not need to care about making or receiving a delivery as long as they do not trade during the delivery month. Traders may at any time cancel out a previous sale by an equal offsetting purchase and vice versa. If done prior to the delivery month the trades cancel out and thus there is no receipt or delivery of natural gas. Only a small percentage of natural futures contracts are settled through deliveries.

NYMEX as any other futures exchange has its own clearing house. All members of an exchange are required to clear their trades through the clearing house at the end of each trading session, and to deposit with the clearing house a sum of money (based on clearinghouse margin requirements) sufficient to cover the member’s debit balance. For example, if a member broker reports to the clearing house at the end of the day total purchases of 100,000 mmBtu of May natural gas and total sales of 50,000 mmBtu of

May natural gas, he would be net long 50,000 mmBtu of May natural gas. Assuming that this is the broker's only position in futures and that the clearing house margin is six cents per mmBtu, this would mean that the broker would be required to have $3,000 on deposit with the clearing house (Futures Market Definitions, Investopedia).

The following are some definitions and concepts related to commodity futures trading (Covel, 2009).

Considering the huge volume of individual transactions that are made, it would be virtually impossible to do business if each party of a trade were obligated to settle directly with each other in completing their transactions.

The justification for futures trading is that it provides the means for those who produce or deal in cash commodities to hedge, or insure, against unpredictable price changes.

The primary function of the commodity trader, or speculator, is to assume the risks that are hedged in the futures market. To a certain extent these hedges offset one another, but for the most part speculative traders carry the hedging load. To sell a commodity future short, a trader sells first and then closes out (or covers) this sale with an offsetting purchase at a later date. A trader does not need to have, or own, the particular commodity involved. The practice of selling short is a common one in futures markets. Those who sell short (with the exception of those placing hedges to protect a cash commodity position) do so in the expectation that prices will decline and that they will be able to buy later at a profit. A short position in the market is of course just the opposite of a long position, which involves buying first and closing out (or liquidating) later with an offsetting sale (Covel, 2009). Selling short is actually possible and legal for futures markets.

One of the conditions is that one agrees to deliver what he sells at a later date. Another condition is that, if one does not deliver, he will stand any loss that the buyer may suffer as a result of an advance in price between the time one makes the sale and the time he cancels out his delivery obligation by means of an offsetting purchase. Let us explore the case where a trader sells 10,000 MM Btu's of NYMEX May natural gas short at $4.10 per MM Btu and then later covers this short sale with an offsetting purchase at $4 per MMBtu, or $1000 on the 10,000 MM Btu contract, less the broker's commission. In the event natural gas prices advance and the trader is forced to cover his short sale at $4.20 and would have a loss of $500, plus commission.

Short sales in commodities are much simpler than in stocks. When a trader sells a stock short he must borrow the stock for immediate delivery against his short sale. This involves a substantial loan deposit and costs that are not involved when one goes long on a stock. Also, stock exchange rules prohibit a stock from being sold short in a declining market unless the short sale is made at a price above the last sale price of the stock, or in other words on an "uptick." The short seller in commodities is faced with none of these restrictions.

Another important aspect of futures trading is margin requirements. They are comparable to the "earnest money" in real estate transactions. Full payment is made at the delivery but prior to that all that is needed is a deposit sufficient to bind the contract. When one establishes a position in a commodity future, either long or short, it is necessary to deposit with the broker a sufficient amount of money to protect the position – actually to protect the broker against loss in the event the trade entered into is unprofitable. This deposit is referred to as the margin. It should not be confused with the clearinghouse margin required of an exchange member. The margin required of a customer by a broker is a different margin than that required of the broker by the clearinghouse. Both margins serve the same purpose, however – they insure that obligations arising from commitments in commodity futures are fulfilled.

The amount of margin that one is required to deposit with the broker in order to trade in commodities is usually 10 percent or less of the market price of the commodity. Exchange regulations prescribe the minimum margins that brokers require of customers. These minimums are changed from time to time, depending on market conditions. The broker is limited only with respect to minimum requirements (Covel, 2009).

Next is an overview of Limit order book modeling.

2.3. Limit Order Book Modeling

Before defining the natural gas trading model based on limit order books, here is some necessary background about the market microstructure.

Algorithm Trading

Algorithmic trading, also called automated trading, black-box trading, or algo trading, is the use of electronic platforms for entering trading orders with an algorithm which executes pre-programmed trading instructions whose variables may include timing, price, or quantity of the order, or in many cases initiating the order by a "robot", without human intervention (Lin and Tom 2013).

An order can generally be decomposed into the three following steps:

Step 1: Trade scheduling: The parent order is split into time “slices”.

Step 2: Optimal execution of a slice: Divide time slices into child orders.

Step3: Order routing: Decide when to send each child order (Moallemi C., 2013).

There are many types of orders that are widely used by traders. Common orders include:

- Good ’Til Canceled (GTC): orders to buy/sell that will stay active in the order book until the trader decides to cancel them. They are usually cancelled within a month. For example if a trader decides to sell 10 contracts of natural gas futures at $5/MMBTU while the current price is $4/MMBTU. So this order will be executed if the price of natural gas futures reaches the $5 mark.
- Fill or kill (FOK): orders to buy/sell that will be either entirely executed immediately otherwise cancelled immediately.
- Immediate or Cancel (IOC): orders to buy/sell that will be either entirely or partially executed immediately otherwise cancelled immediately.
- Good-Til-Date/Time (GTD): orders to buy/sell that will stay active in the order book until executed or cancel at a set date/time.

Market-Making

A market maker or liquidity provider is a company, or an individual, that quotes both a buy and a sell price in a financial instrument or commodity held in inventory, hoping to make a profit on the bid-offer spread, or turn (Radcliffe, 1997).

The figure below gives an overview of the actors in play in electronic financial trading.

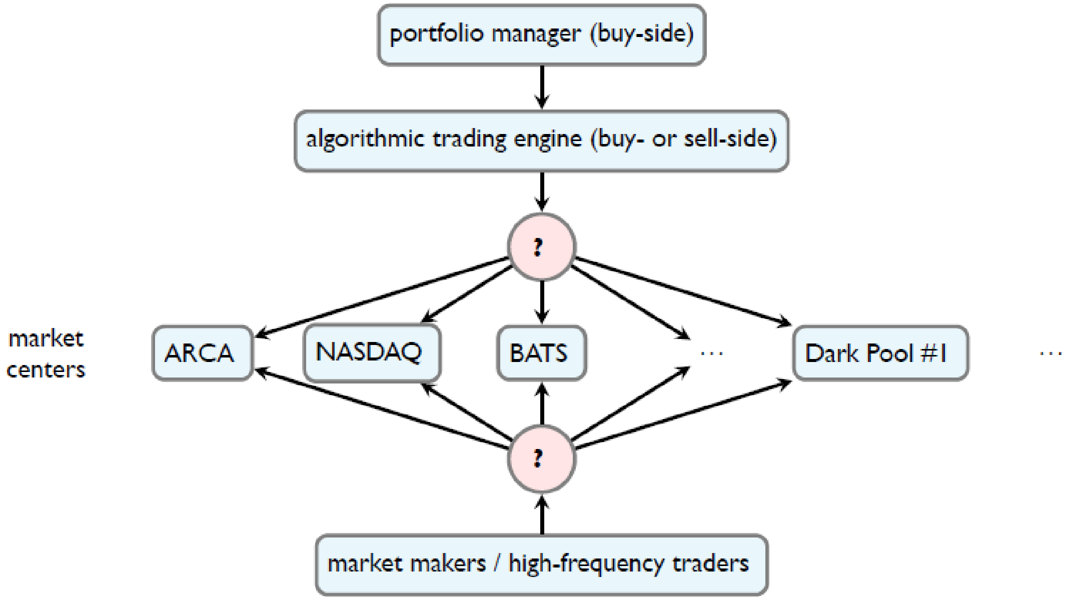

portfolio manager (buy-side)
algorithmic trading engine (buy- or sell-side)
?
market centers
ARCA
NASDAQ
BATS
...
Dark Pool #1
...
?
market makers / high-frequency traders

Figure 15: A simplified view of buy-side trading
(Source: Moallemi, 2013)

The U.S. Securities and Exchange Commission (SEC) lists several characteristics commonly attributed to HFTs including “(1) the use of extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders; (2) use of co-location services and individual data feeds offered by exchanges and others to minimize network and other types of latencies; (3) very short time-frames for establishing and liquidating positions; (4) the submission of numerous orders that are cancelled shortly after submission; and (5) ending the trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions over-night).”

There are two types of traders: informed traders (hedge funds, fundamental traders) and uninformed traders (market makers, high frequency traders).

High frequency traders contribute in the volume but do not provide much liquidity since they are not willing to accumulate large positions. Usually fundamental traders mistake large trading volumes for liquidity.

In (Baron, Brogaard, and Kirilenko, 2014) the authors show that “HFT firms who specialize in liquidity-taking (aggressive) strategies generate substantially more revenue than those who specialize in liquidity-providing (passive) strategies”.

“Given that there is no publicly available data set on HFT firms, several papers study HFT activity despite being unable to directly observe individual high frequency traders (e.g., Hasbrouck and Saar, 2013). Other papers make use of limited or aggregated proprietary data sets. For example, Jovanovic and Menkveld (2012) and Menkveld (2013) study the July 2007 entry into Dutch stocks of a single high-frequency market maker, and Brogaard, Hendershott, and Riordan (2013) study aggregated HFT activity on NASDAQ. It is similar to the work of Kirilenko, Kyle, Samadi, and Tuzun (2014), which studies whether HFTs caused the Flash Crash of May 6, 2010”

Regulations of Trading

Regulation aims to protect the public interest. CFTC, FERC and SEC try to detect frauds such as front running, market cornering (control the long positions to drain the supply thus forcing prices to go up). They then enforce policies by imposing sanctions on the traders who commit fraudulent transactions.

Latency is crucial. Only financial institutions that use a direct link to the exchanges receive the information without the delay that was observed by private providers of financial data. The delay reached up to 36 seconds in the May 6th flash crash.

When it comes to making money out of the malfunctions of the system, the SEC and CFTC are subject to disclosure rules. On top of the prices and trading volumes they have access to the actual brokers and financial institutions that make the trades.

VPIN Flow Toxicity metric was introduced after May 6th flash crash. It delivers a real time estimate of the conditions under which liquidity is provided. This feedback mechanism triggers market makers to be forced out of the market. Since then, regulators have been introducing safeguard mechanisms to ensure an equal market for all its participants even if there is a culture of secrecy in Wall Street. The investigations are taking years. The trading firms are so sophisticated and complex that SEC and CFTC cannot keep up with the speed, sophistication and intelligence.

During the flash crash, the market disappeared: temporarily, $1 trillion in market value disappeared (Senators Seek Regulators' Report on Causes of Market Volatility, Wall Street Journal, May 7, 2010). It appears that the regulation agencies technical capabilities are limited to monitor the current markets.

May 6th Crash

According to Schapiro (2010):

“The absurd result of valuable stocks being executed for a penny likely was attributable to the use of a practice called “stub quoting.” When a market order is submitted for a stock, if available liquidity has already been taken out, the market order will seek the next available liquidity, regardless of price. When a market maker’s liquidity has been exhausted, or if it is unwilling to provide liquidity, it may at that time submit what is called a stub quote – for example, an offer to buy a given stock at a penny. A stub quote is essentially a place holder quote because that quote would never – it is thought – be reached. When a market order is seeking liquidity and the only liquidity available is a penny-priced stub quote, the market order, by its terms, will execute against the stub quote. In this respect, automated trading systems will follow their coded logic regardless of outcome, while human involvement likely would have prevented these orders from executing at absurd prices. As noted below, we are reviewing the practice of displaying stub quotes that are never intended to be executed.”

The recent volumes of trading activity, to some degree, are regarded as more natural levels than during the financial crisis and its aftermath. Some described those lofty levels of trading activity were never an accurate picture of demand among investors. It was a reflection of computer-driven traders passing securities back and forth between day-trading hedge funds. The flash crash exposed this phantom liquidity. High-frequency trading firms are increasingly active in markets like futures and currencies, where volatility remains high (Lauricella, 2011).

Regulators are desperately trying to keep up with the fleet-footed traders that always seem one step ahead, but so far haven not indicated any major plans to restrict their practices. Short of a major change in stance from regulators and exchanges, traditional traders may just need to learn to adapt.

The practice of high-frequency trading has come into question after high-profile allegations of market unfairness, while regulators have insisted the markets remain fair for the average investors (Shroeder, 2014).

Market Infraction Examples

For nearly three months in 2011, Panther would place a small order to sell a futures contract, according to the CFTC. It would then place large orders to buy similar futures contracts at higher prices, giving the impression to the market at large that there was big demand. But the firm would then quickly cancel its buy orders as soon as it sold the contracts it wanted to sell. “The sequence would quickly repeat, but in reverse,” the CFTC said. The company and Coscia would sometimes use the spoofing method hundreds of times in an individual futures contract in a single day, the government said in its complaint. “Spoofing sends false signals to markets in order to lure prey and game the system," said Chilton, who coined the term "cheetah" to describe high-frequency traders because of their speed. "The good news is that regulators around the world are starting to catch up with the cheetah traders and we are shutting them down when they violate the law" (ElBoghdady, 2013). His computer-driven strategy, as described by the government, was to enter large orders with the intent to cancel them before they could be executed. These orders would cause other traders to think the price of a particular contract was moving higher, in the case of a buy order, or lower, in the case of a sell order. Mr. Coscia would take advantage of the market reaction by entering orders to sell at the higher price or buy at the lower price, while canceling the bogus orders, according to the government (Alden, 2014).

Other market manipulation practices include quote stuffing where HFT try to flood the market by quickly entering and canceling large orders.

Limit Order Book Models

In (Chen, Zhou, 2010) the authors built an order book simulator and tested different trading strategies for single stocks. They applied "four variate Hawkes processes to simulate arrival of orders, and measured the market impact caused by different sizes of orders". They found that "the bigger the liquidation size, the larger the market impact is". They designed three "optimal liquidation strategies": naïve split, valley and relative volume. They compared these strategies to a benchmark model based on the market volume weighted average (VWAP). They found that "the relative volume strategy outperformed the VWAP benchmark.

For modeling the arrival of orders, they used an I-variate Hawkes-E(K) process as in (Hawkes, Oakes, 1974).

The idea behind this is that the arrival of specific types of orders would trigger the arrival of other types of orders. For example, when traders notice a trend of market orders coming in, some of them will follow the observed trend and place market orders as well, while some others would place buy orders thinking that the price would go up. This is how the multivariate Hawkes process works (Chen, Zhou, 2010):

N_t^1, N_t^2, N_t^3 and N_t^4 are the counting processes for the 4 major orders (Market Buy Order, Market Sell Order, Limit Buy Order and Limit Sell Order).

$\lambda_t^1, \lambda_t^2, \lambda_t^3, \lambda_t^4$ are their corresponding intensities

The Hawkes process is determined as follows:

$$\lambda_t^i = \mu^i + \sum_{j=1}^{4} \int_0^t \alpha_{ij} e^{-\beta_{ij}(t-u)} \, dN_u^j \qquad , i = 1,2,3,4$$

$$\lambda_t^{cancel} = \mu_t^{cancel}$$

μ^i is the initial intensity of N_t^i.

α_{ij} describes how big the jump of the intensity is.

α_{ij} is the impact of the jth type of order to the ith type of order.

β_{ij} describes how fast the temporary impact vanishes. When β_{ij} is big, it means that the jump in the intensity is only momentary. It will fade away in very short time period Descriptive statistics from real trading data are used to tune the Hawkes process parameters. While simulating their trading strategies, they found that “the larger size of ‘First Order’, the longer effect is”.

Guo (2013) discusses the differences between optimal placement and optimal execution models. Several models from the literature are presented.

The optimal execution models focus on the price impact. The idea behind this is that trading large volumes over a short period of time will impact the price. The objective is to minimize the price impact and/or to maximize expected utility functions. This is done by optimally slicing big orders into smaller ones while keeping in mind that: “too large orders may depress the price and reduce the potential profit and too many small transactions may be costly and may take too long to complete” (Guo, 2014). The optimal placement models focus on “how to optimally place the small-sized orders in the LOB” (Guo, 2014).

Below are some optimal placement models from the literature.

In (Hult and Kiessling 2010), the author presented a Markov chain with N=1 to model the LOB:

There are d price levels in the order book: $\pi^1 < \cdots < \pi^d$

The Markov chain X_t = (X_t^1, · · · , X_t^d) represents the volume at time t of buy orders with negative values and of sell orders with positive values at each price level.

The generator matrix of X is Q = (Q_{xy})

Q_{xy} is the transition matrix from state x = (x^1 , . . ., x^d) to y = (y^1 , . . ., y^d)

For example, a limit sell order of size k at level j corresponds to a transition from state x to state x + ke^j . For the convenience of analysis, they also assume that the highest bid level is always lower than the lowest ask level and that there is always someone to sell at the highest possible price and buy at the lowest possible price.

There is no specific constraint on the spread between the best bid and ask levels. Within this framework, they consider a special version of the optimal placement problem with N = 1. That is, an agent has to buy one unit, and has to decide between place a market order at the best ask level or place the limit buy at a lower level. The key is then to compute the probability that a limit buy order is executed before the price moves up as well as the expected buy price resulting from a limit buy order. Using the potential theory for Markov chains, they derive conditions for the existence of an optimal strategy and provide a value-iteration algorithm to numerically find the optimal strategy. They also calibrated the model using real market data.

In (Guo, Ruan, 2013) the authors proposed a correlated random walk model, focused on modeling the ask price. In (Guo, Larrard, Ruan, 2012), the authors proposed a continuous-time model. Since "the probability of a particular limit order being executed depends on the queue length, its position in the LOB, the frequency of price changes, the arrival rate of market orders and the cancellation of orders" (Guo, Larrard and Ruan 2013). They showed that the limit dynamics of the LOB may be approximated by a Markov process $Q = (Q_t^a, Q_t^b)$.

Ait-Sahalia and Saglam (2013) propose a model of dynamic trading where a strategic high frequency trader receives an imperfect signal about the future order flow, and exploits his speed advantage to act as a market maker.

Foucault et al. (2015) extended Kyle's model by incorporating heterogeneity in the speed of information processing. Foucault et al. (2015) developed a model in which HFTs choose the speed at which to react to news, based on a trade-off between the advantages of trading first compared to the attention costs of following the news. Biais et al. (2011) analyze the arms race and equilibria arising in a model where traders choose whether to invest in fast trading technologies. Jovanovic and Menkveld (2010) study the effect of high frequency trading activity on welfare and adverse selection costs. Cvitani´c and Kirilenko (2010) study the distribution of prices in a market before and after the introduction of HFTs. Following Kirilenko, Kyle, Samadi, and Tuzun (2014), they define different trader types based on two selection criteria: inventory and trading volume. HFTs are identified as those firms with extremely high volume, low intraday inventory and low overnight inventory

Baron, Brogaard and Kirilenko (2014) analyzed competition in the HFT industry to show and understand why profits are concentrated among a small number of incumbents who realize high and persistent returns. According to the winner-takes-all idea, they expect to see a high concentration of profits that persist over time. Recent theoretical papers have highlighted concerns of faster traders adversely selecting slower traders and competition on speed leading to socially inefficient arms races for speed. Baron, Brogaard and Kirilenko (2014) suggest that HFTs have strong incentives to take liquidity and compete over small increases in speed in an industry dominated by a small number of incumbents earning high and persistent returns. Understanding the industrial organization of HFTs allows researchers to think more comprehensively about the role and implications of HFTs in financial markets.

Other relevant papers include Pagnotta and Philippon (2011), Jarrow and Protter (2012), Moallemi and Saglam (2012), Pagnotta (2010) and Cespa and Foucault (2008).

The dynamic optimization aspect is in the line of the classical inventory control problem of Amihud and Mendelson (1980) and Ho and Stoll (1981) (see also Hendershott and Menkveld (2014)) for “traditional” market makers.

Inventory models consider the inventory problem of a dealer who is facing buyers and sellers arriving asynchronously. The string of the literature originates from

Garman (1976) who models the arrival processes of buyers and sellers as Poisson processes. He characterized the inventory problem of a market maker who has to ensure that its holdings on security and cash do not drop below a given level. Amihud and Mendelson (1980) present a similar framework where the market maker's inventory is constrained to lie between upper and lower bounds. Hendershott and Seasholes (2006) examined daily inventory/asset price dynamics using 11 years of NYSE specialist data. Fodra and Labadie (2012) extended the market making models with inventory constraints of Avellaneda and Stoikov (2008) and Gueant, Lehalle and Fernandez-Tapia (2011), to the case of a general class of mid-price processes, under either exponential or linear P&L utility functions.

Order cancellations are widely observed in empirical high frequency data. Hasbrouck and Saar (2009) note that over one third of limit orders is cancelled within two seconds and term those "fleeting orders". Baruch and Glosten (2013) show that quote cancellations can emerge as an equilibrium strategy in a trading game.

Ait-Sahalia and Saglam (2013) compared 3 different HFT regulations. The first one is a Tobin tax taxing each trade executed by HFT. The argument in favor of the tax is that HFTs are under-taxed relative to the economy which encourages excessive HFT. But then HFTs may decide to move to other financial instruments or exchanges that are not subject to the tax, which could harm the financial sector. The second policy is based on imposing a minimum rest time before a quote can be cancelled by HFT. Finally, the third policy consists of taxing limit order cancellations. The authors found that none of these policies result in an improvement compared to doing nothing. Both minimum rest times and a cancellation tax result in more liquidity in good (low volatility) environments but less in bad (high volatility) environments.

Next is an overview of the natural gas futures trading model.

Chapter 3: Natural Gas Futures Trading Model

3.1. Overview

Natural gas futures data are critical for modeling the limit order book. In the context of high frequency trading, the full depth of the natural gas futures limit order book may not be available. Thus the need for simulation as a mean to recover the events that shape the limit order book but may not be fully recorded. Simulation is actually widely used in financial markets. For instance, the Stock Market Crash of 1929, also known as the Black Tuesday was mainly analyzed using simulation tools due to the limited availability of the data. Similarly, the domino effects in stock market - local emerging market crashes evolve rapidly into more severe regional endemics or even global epidemics (Markwat et al., 2008) - are usually studied and examined through simulation.

The simulation methods applied in financial analysis can be broadly categorized into two groups: continuous-time and discrete-time simulations. The limit order book dynamics are complicate and intricate, demanding careful scrutiny on the micro-structure and interactions of various elements in the system (Zhang, 2013).

Simulation also helps generating scenarios that rarely happen in the real world. This allows building robust and sustainable models. Nevertheless, event-based and asynchronous discrete-time simulations, are still rarely utilized on financial analysis in general and evaluation of limit order book dynamics in particular (Jacobs et al., 2004).

In this research, simulation is used to generate hybrid data where the best ask and bid prices and volumes are derived from historical data while the prices and volumes for the remaining levels in the limit order book are recreated using simulation.

Zhang (2013) suggests the use of an asynchronous discrete-time simulation model to constructively create all transactional events so that financial market to be simulated can be derived. He employs a process-based simulation model to generate artificial data for limit order books. This allows to directly model each element within the limit order book via a mathematical process (Shek, 2011), more specifically, price and volume dynamics can be characterized stochastic processes such as Poisson processes or Brownian motions (Lorenz et al., 2008), and similarly, the correlation among different processes may be represented accordingly (Zhang, 2013). For instance, the simulation model in (Luckock et al., 2001) generates a flow of orders with a specified distribution in price instead of mimicking the behavior of each individual trader (Zhang, 2013).

Data Visualization

The shape of the LOB is displayed in figure 16 shows how the order book can be viewed by levels, where level 1 defines the best ask and best bid and level 2 defines the second best ask and bid and so on. The levels following level 1 and their corresponding volumes can give an insight on the future behavior of the order book.

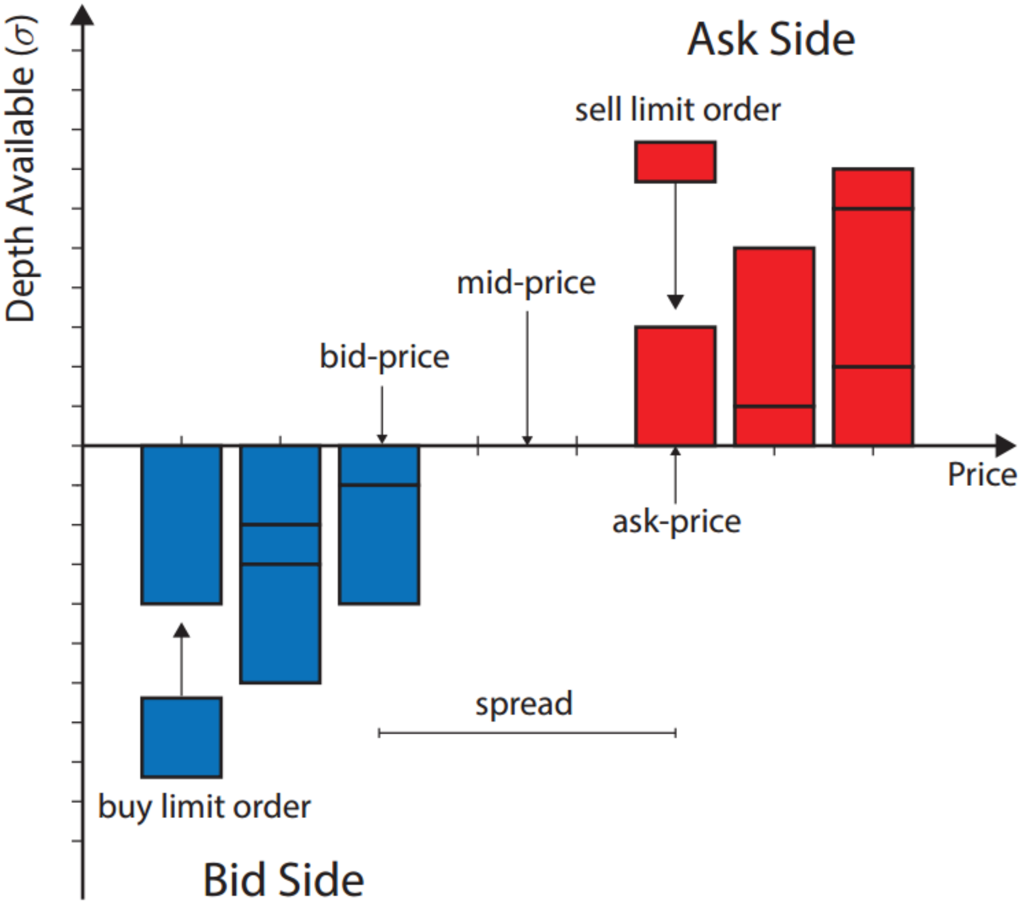


Figure 16: The shape of the LOB
(Source: Gould, Porter, Williams, McDonald, Fenn and Howison, 2013)

The order book of a specific Exchange will keep track of the incoming limit, market and cancellation orders. The structure of this system is described in figure 17.

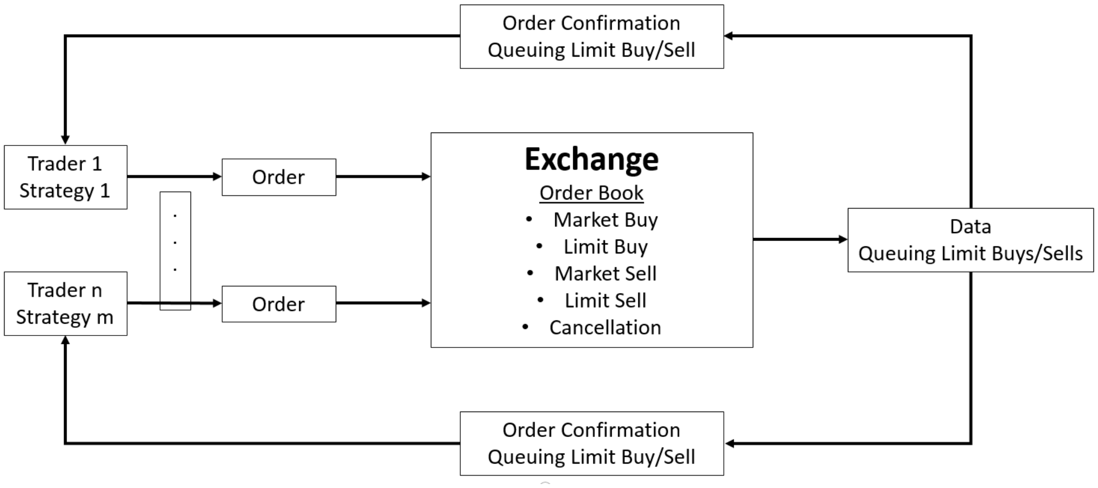


Figure 17: Trading System Structure

The order processing rules that are implemented in the exchange are described in table 4 below.

Order Type	Processing Algorithm
Market Buy	1. Match with orders in the market sell order book; 2. Match with orders in the limit sell order book; 3. The rest is written into the order book for market buy orders, and is to be matched with market/limit sell orders in the future, or to be canceled by cancellation order.
Market Sell	1. Match with orders in the market buy order book; 2. Match with orders in the limit sell order book; 3. The rest is written into the market sell order book, and is to be matched with market/limit buy orders in the future, or to be canceled by cancellation order.
Limit Buy	1. Match with orders in the market sell order book; 2. Match with orders in the limit sell order book if the limit buy price is above the lowest limit sell price; 3. The rest is written into the order book for limit buy orders, and is to be matched with market/limit sell orders in the future, or to be canceled by cancellation order.
Limit Sell	1. Match with orders in the market buy order book; 2. Match with orders in the limit buy order book if the limit sell price is below the highest limit buy price; 3. The rest is written into the order book for limit sell orders, and is to be matched with market/limit buy orders in the future, or to be canceled by cancellation order.
Cancellation	1. Find whether the order the cancellation order is supposed to cancel is still in the order books. 2. If so, remove that order. Otherwise, omit this cancellation order.

Table 4: Order processing algorithm

Standard queuing rules are used in the exchange as orders arrive: First orders are ranked according to the price priority rule. When two or more orders have the same price level, then they are sorted according to the first in first out (FIFO) rule.

Natural gas futures historical data could be displayed annually, monthly, weekly, daily, hourly and even by the minute. The figure below shows the volume fluctuations taken from 1-minute data of natural gas futures prices.

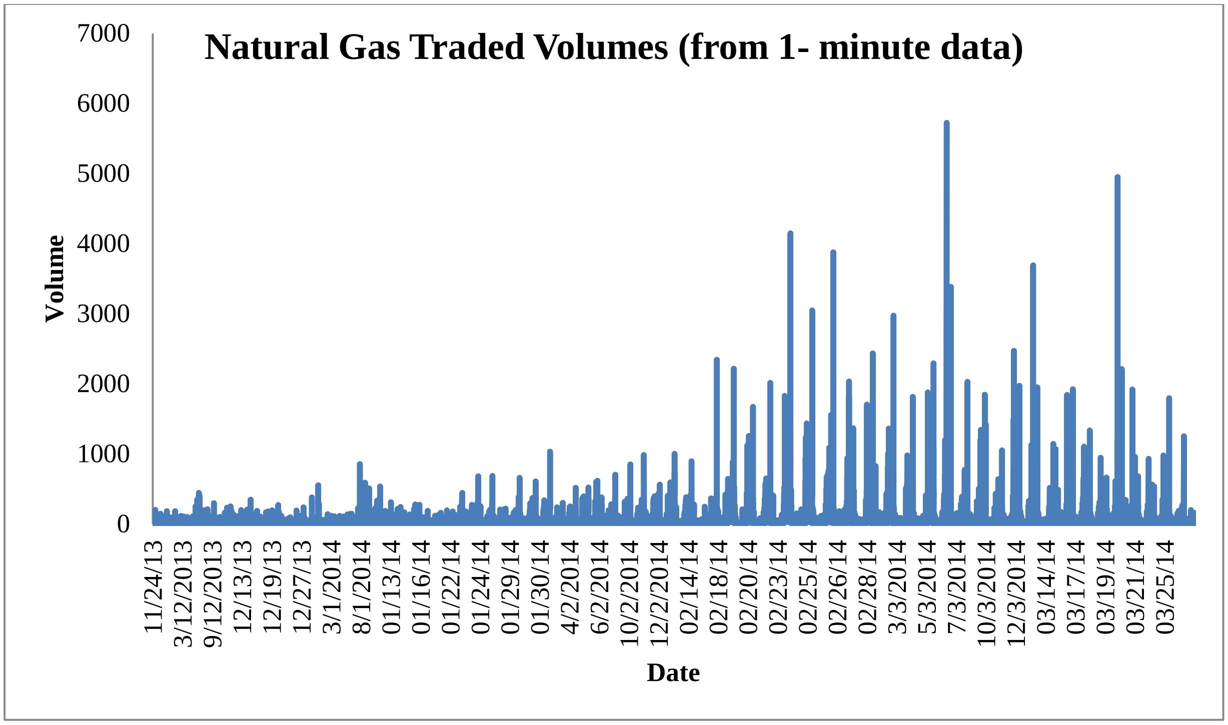


Figure 18: Natural Gas Futures Traded Volumes

The peaks in the traded volumes coincide with the weekly natural gas storage report release dates. The report is released every Thursday at 9:30 am. CST except for few exceptions such as Thanksgiving Day.

3.2. Simulation

Simulation is the imitation of the operation of a real-world process or system over time (Jerry, 1984). Simulation is used before an existing system is altered or a new system is built, to reduce the chances of failure to meet specifications to eliminate unforeseen bottlenecks, to prevent under or over utilization of resources and to optimize systems performance (Maria, 1997). Simulation can be used in lieu of analytical approaches when they cannot be implemented or when the data are not available. An analytical approach based on a mathematical model can provide an exact solution.

However this method could be based on assumptions often not verified in the real world. On the other hand, simulation may not lead to the best solution, nevertheless the precision could be improved by increasing the number of simulation runs (Roberts, Andersen, Deal, Garet, and Shaffer, 1983; Vangheluwe, 2004; Chen and Lee, 2010).

The simulation methods applied in the industry can be categorized into three groups: static or dynamic, deterministic or stochastic, and discrete-time or continuous-time models. A static simulation is one that is not based on time. It often involves drawing samples to generate a statistical outcome. It is sometimes referred to as Monte Carlo simulation. Dynamic simulation looks at state changes that occur over time. Examples of dynamic simulation can be found in manufacturing and service systems such as a conveyor belt system. Stochastic simulation involves the use of random input variables from probabilistic distributions. Deterministic simulation on the other hand, does not use any random input variables. Deterministic simulation will therefore produce the same output regardless of the number of runs (Harrell, Bowden and Ghosh, 2000). In discrete simulation, the state of the system can change only at event times. A discrete simulation model can be formulated by defining the changes in state that occur at each event time, describing the process through which the entities in the system engage, or describing the process through which the entities in the system flow. In a continuous simulation model, the state of the system is represented by dependent variables which change continuously over time. The continuous change variables are referred to as state variables whose dynamic behavior simulates the real system (Pritsker and O’Reilley, 1999). Most systems are modeled based on a combination of the above simulation methods (Law and Kelton, 1991). Simulation modeling can be classified into three phases: data generation, bookkeeping and output analysis (Gross and Harris, 1985).

In the context of the limit order book, the simulated data provide a mean to reproduce scenarios that happen in the real world. Furthermore they allow to build robust models by testing different scenarios that are less likely to happen.

Zhang (2013) proposed a framework to simulate the limit order book data following four main phases:

1. Parameters determination: These are inferred from real world data and include order submission and cancellation rates.
2. Order flow generation: A Poisson process is used to simulate order arrivals of each level in the order book. Cancellations are proportional to the order volume. Real data are used to calibrate the simulation model.
3. Data flow aggregation: All the events are aggregated and merged into the limit order book depending on when they were created.
4. Simulated data validation: The mean squared error is used to compare the simulated data to the real world data by comparing the arrival intensities among other metrics.

The proposed simulation model to reconstruct the limit order book reuses some of the concepts outlined above from Zhang (2013). In this case, historical data for best ask and bid prices and volumes are used to reconstruct the top level in the limit order book. The remaining levels on the bid and ask sides are simulated. The resulting hybrid limit order book data mimic the behavior of the actual limit order book.

Poisson Process

Poisson process is one of the most important models used in queuing theory (Virtamo, 2005). It is also used in renewal theory. The arrival of orders to the limit order book can be modeled as a Poisson process.

The figure below shows the simulation of different time between arrival rates.

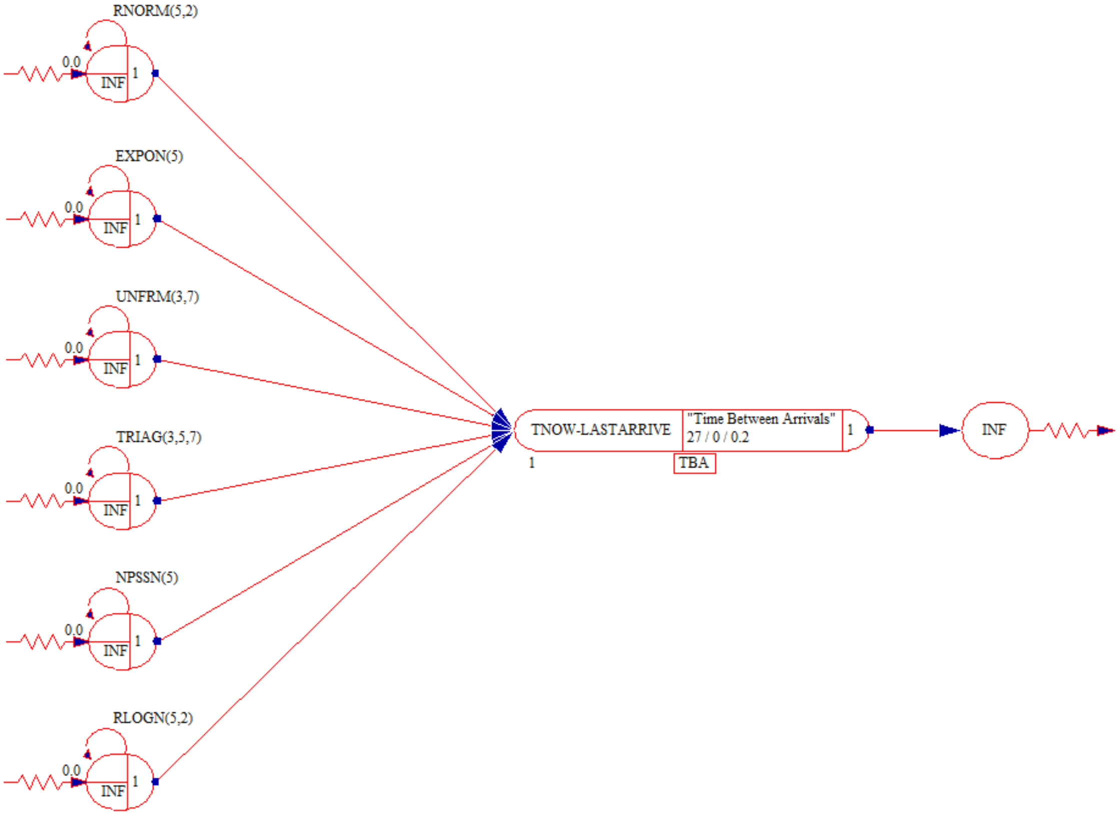


Figure 19: Simulation of a renewal process with 6 random distributions

After superposing those distributions, we get a distribution that follows a Poisson process as shown in the histogram below.

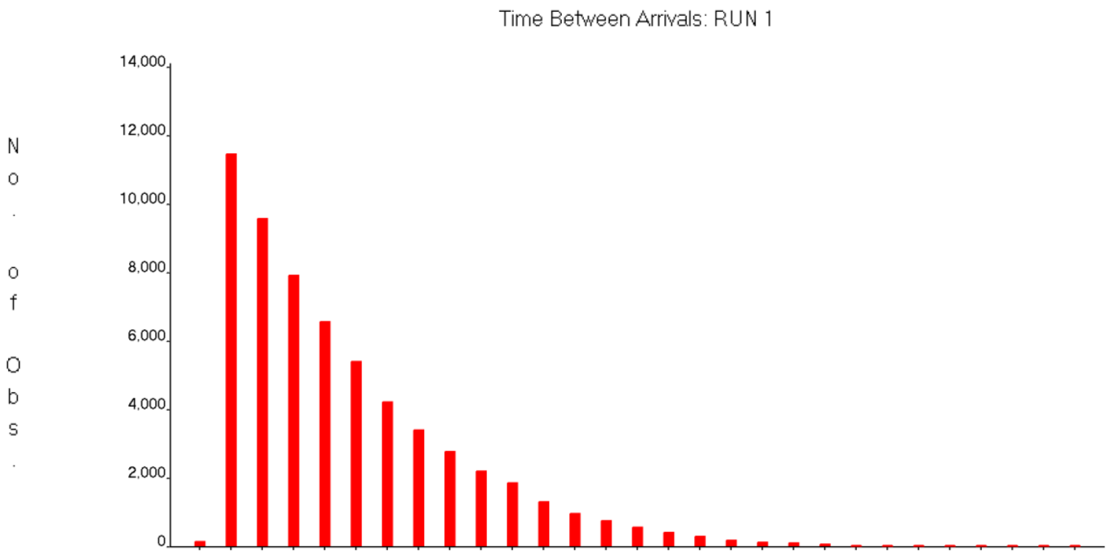


Figure 20: Time between arrivals for the superposed distributions

The resulting distribution follows a Poisson Process, which justifies its use for modeling order and cancellation arrivals to the limit order book.

The main validation measurement would be the mean square error (MSE), but since we do not have access to market depth historical data, this measure should be used for future work. It is reasonable to assume that the synthetic data generated by the above-described procedure possess the properties specified by the given set of parameters should the data pass the quality validation, and therefore could be used the same way as real world data to train and test learning models.

Assumptions:

1- Transaction costs and commission fees are not counted.

2- Historical best bid and ask prices and volumes are used to populate the level 1 data on the grid.

3- Price range from historical data (2/17/2015): [$2700, $2790] per contract.

4- 1 tick size is $0.001 per mmBtu ($10 per contract)

5- Trading time T = 5 hours (9:00 am to 2:00 pm) = 300 minutes= 18000 seconds

The following contents of this section reuses some of the concepts suggested in (Zhang, 2013), the difference being that the procedures and details are described for hybrid data generation instead of purely synthetic data generation.

Let T denote the whole duration of the simulation period in seconds, which follows the length of a complete trading day in real data. K is the length of price range with tick as unit size. For example, to cover a price range around [$2700, $2790], we take K = 9, and 1 tick = $10. Starting from 1 to 9, Each price takes 1 grid of the 9 grids from low to high, which corresponds to the lowest possible bid price to the highest possible ask price. Let i be the grid index for the price, thus $\mathrm{i} = 1$ corresponds to the price 2700 and i =9 is the price $2790. The grid $q_i = \{P_i, V_i\}$, stores the price P_i and volume $V_i$on the grid. New orders can arrive in any grid from $\mathrm{i} = 1$ to $\mathrm{i} = 9$, corresponding to the price range [$2700, $2790]. And by dividing

T into M equal periods of t, the order book data of T are generated by combining M windows of data simulated in each t_p, p = 1, 2, …, M, thus T=M×t.

The table below summarizes the key simulation parameters.

T (s)	t	M	K	Price Range ($)
18000	100	180	9	[2700, 2790]

Table 5: Simulation parameters

The simulation is performed during each simulation window t_p. In the first window t_1, the best ask price and bid price are initialized from the top of the book historical data and placed into the corresponding grids $q_{i_a}$and q_{i_b}. Thus for any grid q_i, the tick distance j to its opposite best quote is calculated as $j = |i - i_a|$ or $j = |i - i_b|$ depending on if it's on bid side $(i \leq i_b)$ or ask side $(i \geq i_a)$. When the simulation starts, all grids $q_i = (i = 1, \ldots, 9)$ generate their own order arrival and cancellation events by using Poisson processes independently with intensity which varies by j. The time stamp of each event is recorded. The top of the book data are used to populate the level 1 data in the grid. Market orders are not simulated in this framework but rather derived from historical data. For the grids q_i such that $i_b < i < i_a$, both limit ask order and bid orders are opened for placement.

The following table summarizes the key simulation events.

i	j	Possible events
$i = i_b$	$j_b = \|i_b - i_a\|$	limit bid order arrival and cancellation
$i = i_a$	$j_a = \|i_a - i_b\|$	limit ask order arrival and cancellation
$i < i_b$	$j_b = \|i - i_a\|$	limit bid order arrival; cancellation
$i > i_a$	$j_a = \|i - i_b\|$	limit ask order arrival; cancellation
$i_b < i < i_a$	j_a and j_b	limit ask and limit bid order arrival

Table 6: Simulation events

The intensity of Poisson process that generates events on grid q_i depends on the index i and the distance j. Two independent Poisson processes simulate the limit order arrival and cancellation process respectively.

Similarly, the distribution of cancellation rates also follows a Poisson process as shown below. By using the methods described above, we were able to reconstruct the market depth order book for natural gas futures prices for up to 10 levels.

The simulated parameters and distributions seem to be a good approximation of the behavior of natural gas futures book. So far we have used the top of the book natural gas futures historical data which only reflect the best quotes. Ultimately, we need to have access to market depth natural gas futures historical data to be able to calibrate and validate the simulation model. The mean square error is a good way to validate the simulation model.

Monte Carlo simulation is also used to generate buy and sell prices. The log-return of the current natural gas futures price divided by the previous price can be modeled by a normal distribution:

$$\ln\left(\frac{S_t}{S_{t-1}}\right) \sim^{\Phi} \left[\left(\mu - \frac{\sigma^2}{2}\right)T, \sigma\sqrt{T}\right]$$

The mean of this distribution is the deviation (drift) minus half of the variance over a given period. The variance of this distribution is the volatility over a given period of time.

Here is a representation of the model following Monte Carlo simulation:

$$\ln\left(\frac{S_t}{S_{t-1}}\right) = \alpha + z_t\sigma$$

The log-return is the combination of two factors:

-The deterministic factor α is the deviation (drift) and represents the expected return

-The stochastic factor which depends on the volatility which is multiplied by the random variable z_t. This random number allows us to model price fluctuations as a stochastic process (Hull, 2008). This Brownian motion model can be considered as a “random walk” which reflects how financial markets work.

3.3. Predictive Modeling

The SVM Model

The process of building a support vector machine learning model is described in the figure below:

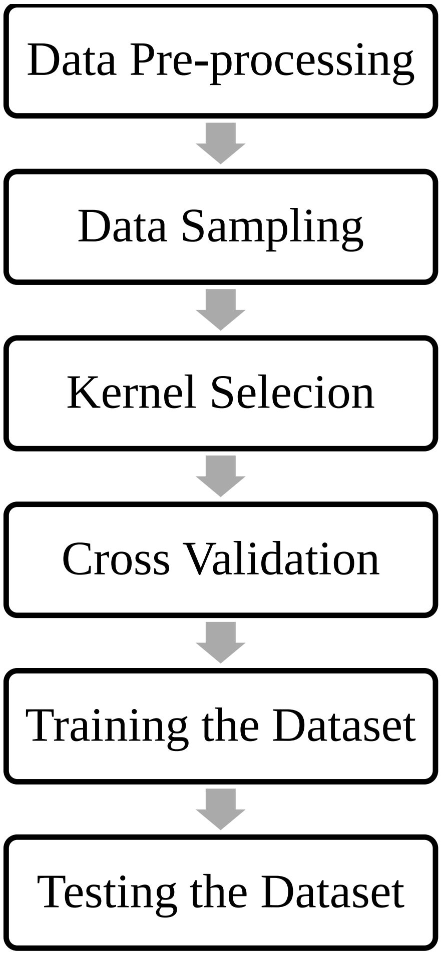


Figure 21: SVM prediction procedure

We start by Pre-processing the data by removing entries for natural gas futures other than the front month. We then select the predictors (best ask and bid prices and volumes).

The kernel function used in the proposed framework is the polynomial kernel $k(x, y) = (x^T y + c)^d$ (Vapnik et al., 1998, 2013). The radial basis function kernel $k(x, y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ is widely used for classification problems but is not suitable for this research. We apply the polynomial kernel and obtain the best cost and degree parameters (best c and d) by using cross validation. Then we use the obtained best parameters (c and d) to train the data set. We finally test the trained data set on the new data.

For degree-d polynomials, the polynomial kernel is defined as:

$$K(x, y) = (x^T y + c)^d$$

Where x and y are vectors in the input space, i.e. vectors of features computed from training samples and $c \geq 0$ is a free parameter trading off the influence of higher-order versus lower-order terms in the polynomial.

As a kernel, K corresponds to an inner product in a feature space based on some mapping φ:

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle$$

The nature of φ can be seen from an example. With $d = 2$ we get a quadratic kernel:

$$K(x, y) = (\sum_{i=1}^{n} x_i y_i + c)^2$$

From this the feature map is given by:

$$\varphi(x) = \langle x_n^2, \ldots, x_1^2, \sqrt{2} x_n x_{n-1}, \ldots, \sqrt{2} x_n x_1, \sqrt{2} x_{n-1} x_{n-2}, \ldots, \sqrt{2} x_{n-1} x_1, \ldots, \sqrt{2} x_2 x_1, \sqrt{2c} x_n, \ldots \sqrt{2c} x_1, c \rangle$$

Binary SVMs are not suitable for this model, given the proposed metrics based on 7 classes to be labeled. Thus, multi-class SVMs are built to label the 7 different classes given by the prediction metrics.

Two approaches are used: multi-class to binary reduction (Allwein et al., 1999) and multi-class optimization methods (Crammer et al., 2002).

3.4. Optimization

George Dantzig (1947) was the first to introduce linear programming. Since then, optimization became widely used to solve complex problems. We can distinguish between two different types of optimization methods: Exact optimization methods that guarantee finding an optimal solution and heuristic optimization methods where we have no guarantee that an optimal solution is found (Rothlauf, 2011). Some of the exact algorithms include simplex algorithm, Branch-and-Bound algorithm, Cutting-plane algorithm. Heuristic algorithms include tabu search, genetic algorithms, and simulated annealing (Pardalos, Rebennack, and Scheidt, 2009). Exact algorithms are preferred whenever we are dealing with polynomial-time problems. They are used in several areas of research to optimize supply chain performance, such as network optimization. If problems are NP-hard, the exact optimization methods need exponential effort (Rothlauf, 2011). Therefore, heuristic algorithms are used to find a near optimal solution in a reasonable time (Gill, Murray, and Wright, 1981). Most optimization problems in energy industry are one of those problems that cannot be solved in polynomial time (Rardin and Uzsoy, 2001).

In this research, we consider the case where a high frequency trader wants to minimize the short term inventory risk of his trading strategy. The primary purpose is to mitigate the inventory risk by managing the inventory levels. Several assumptions are made:

Assumption 1: A trader only buys long and sells short. In this case we can forecast the volumes to be sold at the best ask price.

Assumption 2: A small number of time periods is adopted to reflect the short term inventory risk management strategy.

Assumption 3: The time is discretized into separate events. This means that the decision variables will be used whenever a spread crossing happens.

Assumption 4: We assume that the inventory holding cost is $C(k) = \$10$ per contract per event.

The following is a description of the inventory model that focuses only on the buy side of the limit order book.

Variable Definitions:

Event k reflects when an upward spread crossing happens. We previously defined 3 labels in our learning model that represent 3 upward spread crossing amplitudes:

“oneup” (1 tick upwards) corresponds to:

$$0 < amplitude \leq 0.001 \qquad (1)$$

“twoup” (2 ticks upwards)corresponds to:

$$0.01 < amplitude \leq 0.002 \qquad (2)$$

“twreeup” (3 or more ticks upwards) corresponds to:

$$0.002 < amplitude \qquad (3)$$

The dimension considered is the number of natural gas contracts. The index k ranges from 1 to T=10 and refers to the event that is forecasted in the learning model.

$B(k)$: The number of contracts to be bought in event k

$I(k)$: Inventory level at the end of event k

Parameters:

$A(k)$: The number of contracts to be sold in event k. These values are forecasted using a simple moving average model.

$P(k)$: The price per contract for contracts sold in event k. These values are forecasted using a simple moving average model.

$I(0)$: Initial inventory level

$I(10)$: Final inventory level

Constraints:

The number of contracts sold must be equal to the number of contracts bought in event k plus the net change in inventory.

$$I(k-1) + B(k) - I(k) = A(k), k = 1, \dots, 10$$

This model is meant to be solved over and over as time advances and as parameters change.

The maximum inventory is set to M=10 contracts:

$$I(k) \leq 10, k = 1, \ldots, 10$$

The initial and final inventory conditions are treated as constraints.

$$I(0) = 0$$

$$I(10) = 0$$

We set $I(10) = 0$ for every iteration of the model. The value of $I(0)$ will depend on the previous iteration.

We also want the number of contracts bough and the number of contacts in position to be positive.

$$I(k) \geq 0, B(k) \geq 0 \; for \; all \; k = 1, \ldots, 10$$

Objective Function:

We want to minimize the total cost: buying contracts and keeping inventory.

Minimize $Z = \sum_{k=1}^{10} P(k) \times B(k) + 10 \times I(k)$

The following is a summary of the mathematical problem formulation.

$$\min \left[\sum_{k=1}^{T} (P(k) \times B(k) + C(k) \times I(k)) \right]$$

Subject to

$$I(k-1) + B(k) - I(k) = A(k), \forall\, k \qquad (1)$$

$$I(k) \leq M, \forall\, k \qquad (2)$$

$$I(k) \geq 0, \forall\, k \qquad (3)$$

$$B(k) \geq 0, \forall\, k \qquad (4)$$

$$I(0) = 0 \qquad (5)$$

$$I(T) = 0 \qquad (6)$$

The number of contracts bought vary based on inventory level as shown by constraints (1) and (2).

In order to implement the market making trading strategy, the same model described above from the buy side, is extended to the sell side. The events k in this case refer to the detection of a downward spread crossing. Thus, in the market making strategy, a single trader will be buying and selling simultaneously.

The inventory level is updated at the end of the trading period for each iteration. Pending orders are cancelled. We also offset short sell positions to keep the inventory positive at all times for modeling purposes.

Chapter 4: Solution Approach

4.1. General Framework

In HFT, we are looking for ways to capture patterns in the data and predict those patterns in future instances in order to make profit. The limit order book has many levels corresponding to queues where orders (prices and matching volumes) queue up to be executed. The incoming orders are ranked depending on the price, and then if two orders enter the queue at the same price level, first in first out (FIFO) rule is applied. Two patterns are interesting: the mid-price movement and the spread crossing.

The mid-price movement is the direction of the mid-price which could be an indicator of a possible profit if detected accurately. But this will not guarantee making profit. The best ask and bid prices are particularly interesting. Using them is the only way to make a transaction happen. Placing a limit order does not guarantee its execution. On the other hand, placing a market order will guarantee the execution of an order. A market sell order is placed at the highest bid and a market buy order is placed at the highest ask. The spread crossing detects when the best ask and bid prices cross at different points in times or after a set number of events. This happens rarely, but when it happens, it guarantees profit because buying at best bid and selling at best ask when a spread crossing happens will guarantee profit. In this research, we further expand the notion of spread crossing to include the amplitude of spread crossing as an important metric. Therefore, performing predictive modeling that accurately predicts when these spread crossings happen is extremely critical to make profit. Hence, we focus on the spread crossing direction and amplitude (the number of ticks by which the spread was crossed) as a target for this research.

Trading strategies are then derived:

- If no spread ("stat") → do nothing
- If upward ("up") → buy at best ask then sell at best bid after a preset number of events (i.e. buy long sell short)
- If downward ("down") → buy at best ask and sell at best bid after a preset number of events (i.e. buy short sell long)

Market maker inventory optimization models can be used to determine optimal bid and ask quotes subject to inventory constraints.

4.2. Heuristic Approach

Due to the lack of market depth data for natural gas futures, we used the front month natural gas futures top of the book data to recreate the limit order book. A simulation grid is created to represent prices and volumes for 10 levels of data on both the bid ask sides.

A predictive modeling technique based on multi-class support vector machines is used. We first tried several techniques to detect when spread crossing happens. We calculated the spread crossings happening within a one second and a two second range. We were able to count few cases where the spread crossing happens. But, this did not allow us to consistently and accurately find the points in time or events when we observe the spread crossing. Hence, we discretized the trading framework as suggested by Jacobs et al. (2004) by basing our spread crossing detection method on the number of events k that separate two best bid and ask prices. We tried different values for the length of sliding windows ranging from 5 to 100 and selected k = 30 to be the optimal value. We then implemented the spread crossing calculations by adding a new column to the data set. It has three possible values for an upward spread crossing: “oneup”, “twoup” and “threeup”. Those values correspond to the different amplitudes of the spread:

“oneup” (1 tick upwards) corresponds to:

$$0 < amplitude \leq 0.001 \qquad (1)$$

“twoup” (2 ticks upwards)corresponds to:

$$0.01 < amplitude \leq 0.002 \qquad (2)$$

“twreeup” (3 or more ticks upwards) corresponds to:

$$0.002 < amplitude \qquad (3)$$

Three values for a downward spread crossing are defined in a similar fashion:

“onedown” (1 tick downwards) corresponds to:

$$0 < amplitude \leq 0.001 \qquad (4)$$

“twodown” (2 ticks downwards) corresponds to:

$$0.001 < amplitude \leq 0.002 \qquad (5)$$

“twreedown” (3 or more ticks downwards) corresponds to:

$$0.002 < amplitude \qquad (6)$$

Finally, one value for no spread crossing is defined:

“stat” (stationary) corresponds to:

$$amplitude \leq 0 \qquad (7)$$

Metrics (1), (2), (3), (4), (5), (6) and (7) are used as the target variables to be predicted.

Two types of trading strategies are derived: a strategy using Immediate-Or-Cancel orders where an order is totally or partially executed while the remaining is cancelled, and a strategy that limits the number of order’s cancellations. The later strategy uses an inventory optimization framework, where the buy and sell volumes are optimally determined. At the end of the trading period, Market orders are sent to liquidate any remaining inventory.

The figure below summarizes this approach.

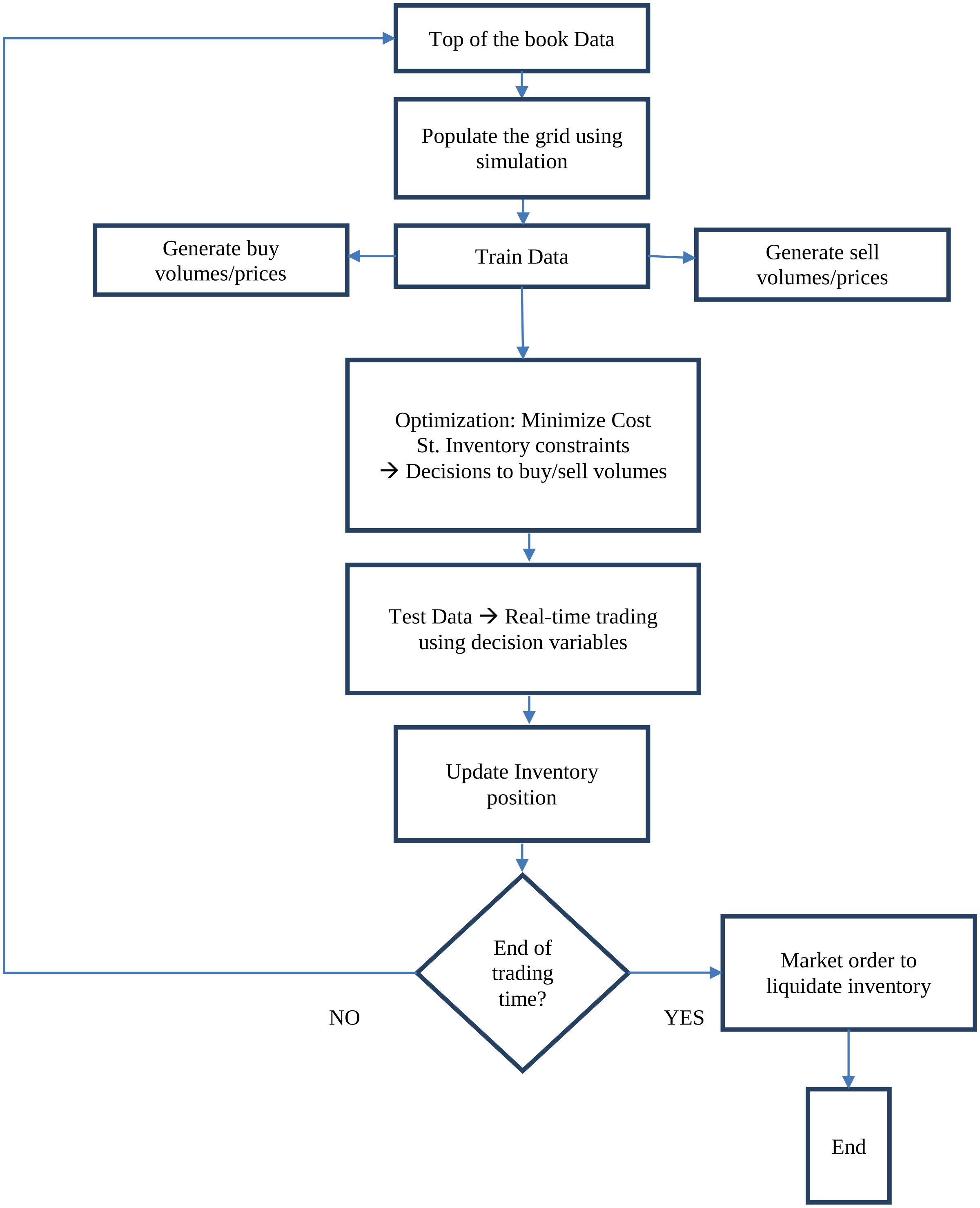

Top of the book Data
Populate the grid using simulation
Generate buy volumes/prices
Train Data
Generate sell volumes/prices
Optimization: Minimize Cost
St. Inventory constraints
→ Decisions to buy/sell volumes
Test Data → Real-time trading using decision variables
Update Inventory position
End of trading time?
NO
YES
Market order to liquidate inventory
End

Figure 22: Overview of the proposed approach

Chapter 5: Experimental Results

The heuristic algorithm developed in this research is capable of recreating the level 10 order book for natural gas futures using top of the book historical data. A variant of the multiclass support vector machine model developed by Kercheval and Zhang (2015) is trained on past data and tested on new data. Labels such as the amplitude and direction of spread crossing can be predicted with a high accuracy. Multiple trading strategies are then derived. An inventory optimization model is implemented to determine the best quotes to buy and sell. This limits the number of cancellation while still making profit. In these settings this model automates profitable trading strategies for natural gas futures.

5.1. Simulated Limit Order Book

The table below shows what the simulated order book looks like.

	level 1				level 2				level 3				…
event	Ask price	Ask vol.	Bid price	Bid vol.	Ask price	Ask vol.	Bid price	Bid vol.	Ask price	Ask vol.	Bid price	Bid vol.	…
k-1	2.8569	6	2.8544	7	2.8571	8	2.854	15	2.8572	2	2.8538	22	…
k	2.8571	8	2.8544	7	2.8572	2	2.854	15	2.8574	2	2.8538	22	…
...	…	…	…	…	…	…	…	…	…	…	…	…	…
k+4	2.8571	8	2.857	6	2.8572	2	2.8544	7	2.8575	4	2.854	5	…
k+5	2.8571	8	2.857	6	2.8572	2	2.8544	7	2.858	10	2.854	5	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
k+8	2.8571	10	2.8544	7	2.858	10	2.854	5	2.8548	10	2.8538	22	…
k+9	2.8571	10	2.8545	8	2.858	10	2.8544	7	2.8581	10	2.854	5	…
k+10	2.8568	8	2.8545	8	2.8571	10	2.8544	7	2.858	10	2.854	5	…

Table 7: A sample of the simulated order book for natural gas

As outlined in Kercheval and Zhang (2015), a new entry in the message file causes one fresh record to be added into the order book.

For instance, the transaction event at the k-th row of the message file of Table 7,

- Execution of an ask order at the price $2.8569 with 6 contracts,
- Exactly cancels out the best ask price and its volume in Row k-1 of the order book,
- Making the next best ask price, $2.8571, become the new best ask price as shown in Row k of the order book.
- It can also be observed from the message file that multiple trading events could arrive within milliseconds as demonstrated from Row k-1 to k+10, leading to drastic fluctuation of prices and volumes in the order book.

Although a variety of "metrics" have been designed to capture the price fluctuation, in this research we select as a metric: the occurrence, direction and amplitude of bid-ask spread crossing.

The trading strategies presented are not adapted to the high frequency trading approach of quote stuffing that happens when natural gas inventory levels are released every Thursday at 9:30 am Central Time.

Case Study

Trading period: Tuesday February 17, 2015 from 9:00 am to 2:00 pm.

This is a regular trading day.

Figure 26 below describes the evolution of the best ask and bid front month natural gas futures prices for February 17, 2015 from 9:00 am to 9:55 am. Notice that the data are time stamped to microseconds. The corresponding ticker symbol is NGH5.

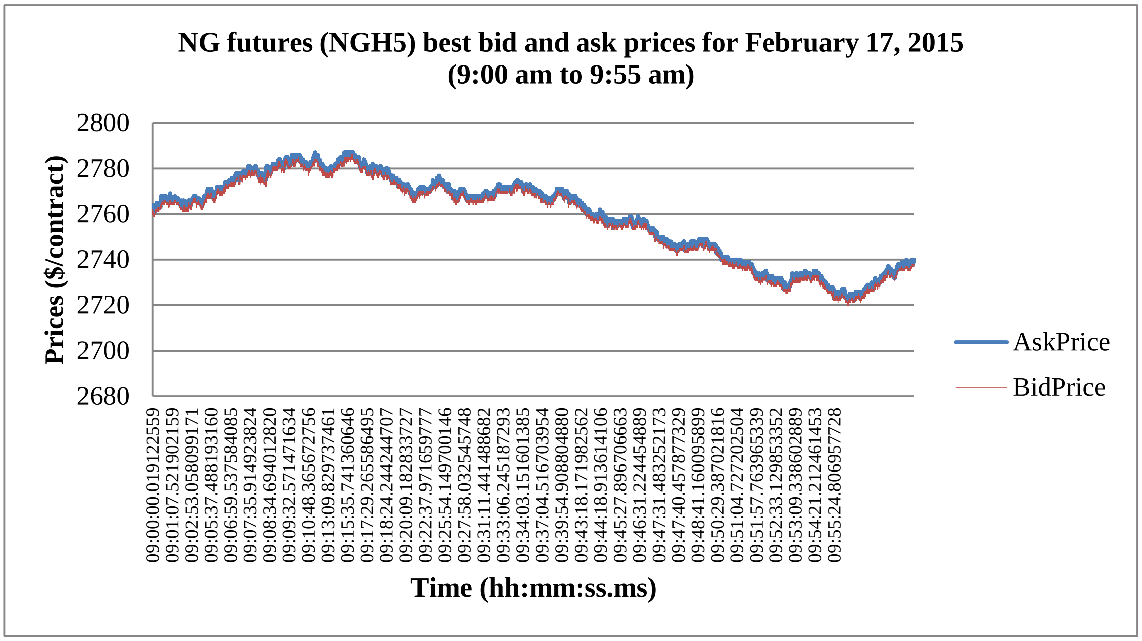


Figure 23: NG futures bid and ask prices in February 7, 2015

The figure below describes the evolution of the best ask and bid front month natural gas futures volumes for February 17, 2015 from 9:00 am to 9:55 am

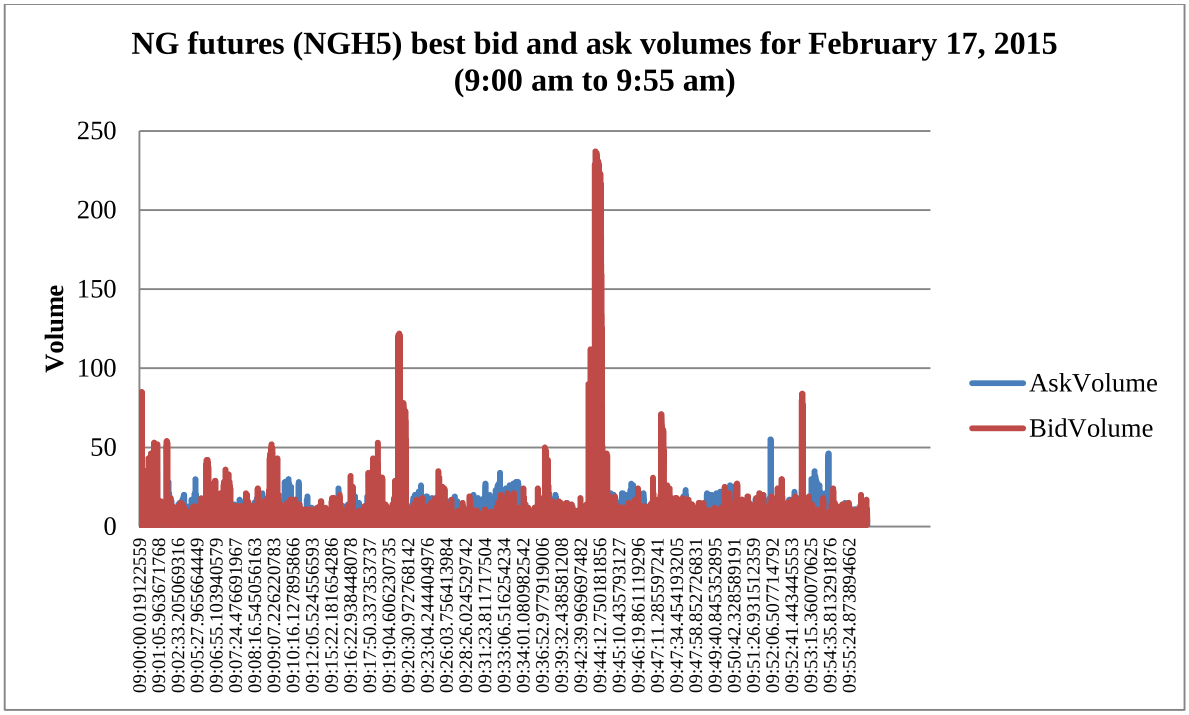


Figure 24: NG futures bid and ask volumes in February 7, 2015

5.2. Model Performance

Similar to to the proposed model of Kercheval and Zhang (2013). We initially have four available features for each level which are: “best ask price”, “best ask volume”, “best bid price”, and “best bid volume” which add up to forty features since we use level ten data. We computed two additional features to each level which are: “midprice” and “spread” and they can be calculated as shown below.

$$midprice = \frac{1}{2}(best\ ask\ price + best\ bid\ price)$$

$$spread = \ best\ ask\ price - best\ bid\ price$$

We obtain a set of 60 features that we use to train the data.

Metric	Description
$\boldsymbol{ASKp_i}$	Ask prices for all levels $i = 1, \dots, 10$
$\boldsymbol{ASKv_i}$	Ask volumes for all levels $i = 1, \dots, 10$
$\boldsymbol{BIDp_i}$	Bid prices for all levels $i = 1, \dots, 10$
$\boldsymbol{BIDv_i}$	Bid volumes for all levels $i = 1, \dots, 10$
$\boldsymbol{midprice_i}$	$midprice_i = \frac{1}{2}(ASKp_i + BIDp_i)\ \forall i$
$\boldsymbol{spread_i}$	$spread_i = ASKp_i - BIDp_i\ \forall i$

Table 8: Description of the predictors

7 labels were considered. They represent the direction and amplitude of the bid-ask spread crossing:

$$\begin{cases} oneup \\ twoup \\ threeup \\ onedown \\ twodown \\ threedown \\ stat \end{cases}$$

"oneup" (1 tick upward) corresponds to:

$$0 < amplitude \leq 0.001 \qquad (1)$$

"twoup" (2 ticks upwards) corresponds to:

$$0.001 < amplitude \leq 0.002 \qquad (2)$$

"twreeup" (3 or more ticks upwards) corresponds to:

$$0.002 < amplitude \qquad (3)$$

Three values for a downward spread crossing are defined in a similar fashion:

“onedown” (1 tick downwards) corresponds to:

$$0 < amplitude \leq 0.001 \qquad (4)$$

“twodown” (2 ticks downwards) corresponds to:

$$0.001 < amplitude \leq 0.002 \qquad (5)$$

“twreedown” (3 or more ticks downwards) corresponds to:

$$0.002 < amplitude \qquad (6)$$

Finally, one value for no spread crossing is defined:

“stat” (stationary) corresponds to:

$$amplitude \leq 0 \qquad (7)$$

1500 data points are trained then the new data are tested following a sliding window.

The machine that was used has the following specifications:

$$\begin{cases} Processor: Intel® \ Core™ \ Duo \ CPU \ T9900 \ @3.06 \ GHz \\ Installed \ memory \ (RAM) \ 8.00 \ GB \\ System \ type: 64 - bit \ Operating \ System \end{cases}$$

Training Time (seconds)	**Prediction Time (seconds)**
[2.635,6.846]	**[0.011,0.043]**

Table 9: Training and prediction time performances

The degree 2 polynomial Kernel was used in this model.

The cost variable is tuned and computed for each training set.

Parameter	Value
Type	C-svc C classification
Kernel	Polynomial kernel
Kpar (kernel parameters)	List (degree=2)
C (cost of constraints violation)	Parameter range [10^{-1} ,10^{1}]

Table 10: SVM model parameters

For natural gas futures contracts, we seldom get an amplitude of spread crossing greater than 2 ticks. For simplifying the results, we grouped the spread crossings by direction regardless of the amplitude.

An additional classification model based on Decision Trees is tested for comparison purposes. A total of two different classification models are built to predict the target variables. The first model is built using SVM with the parameters shown in the table above. The values of these parameters are obtained after tuning the model. The second model is built using Decision Trees.

The confusion matrix is computed for each model in order to calculate the predictions' accuracy. The table below describes the parameters of the confusion matrix.

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Table 11: Confusion matrix

The accuracy is calculated as follows:

$$Accuracy = \frac{TP + TN}{P + N}$$

Where:

$TP: True\ Positive$

$TN: True\ Negative$

P: Positive

N: Negative

The confusion matrix is extended to multiple classes following the same logic. The tables below describe the confusion matrix for the SVM and DT models. In this case we used the first 1500 data points generated from front month natural gas futures historical data on February 17, 2015 starting at 9:00 am.

SVM Model		***True* Classes**		
		Up	Down	stat
***Predicted* Classes**	Up	0	1	1
	Down	0	18	0
	Stat	43	0	1437

Table 12: Confusion matrix for the SVM model

DT Model		***True* Classes**		
		Up	Down	stat
***Predicted* Classes**	Up	0	0	29
	Down	4	16	278
	Stat	32	4	1137

Table 13: Confusion matrix for the DT model

As a result, the SVM model generates 157 support vectors and is more accurate as it has 97% prediction accuracy whereas the DT model has 77% prediction accuracy. Hence, the SVM model is more appropriate for detecting the direction of spread crossing.

Finally, the profit is obtained using the strategy rules outlined in the section below.

5.3. Trading Strategies

The next step is to implement a trading strategy based on the spread crossing direction:

At event i:

- If no spread (“stat”) → do nothing
- If upward (“up”) → buy at best ask at event i and sell at best bid at event i+30 (i.e. buy long sell short)
- If downward (“down”) → buy at best ask at event i+30 and sell at best bid at event i (i.e. buy short sell long)

Once we implement this trading strategy, we first verify that it does what we intended it to do (i.e. making profit). Then we validate it by implementing our predictive models.

We select support vector machines to fit our model because it has the highest prediction rate. After selecting the kernels and tuning the kernel parameters for the SVM, we train it with past data and test it on new data. Then we calculated its accuracy which is above 97%. We then validate the model by applying it on the upcoming test data. Our trading strategy leads to profit.

The prediction time horizon can impact the profit calculations as shown in the figure below.

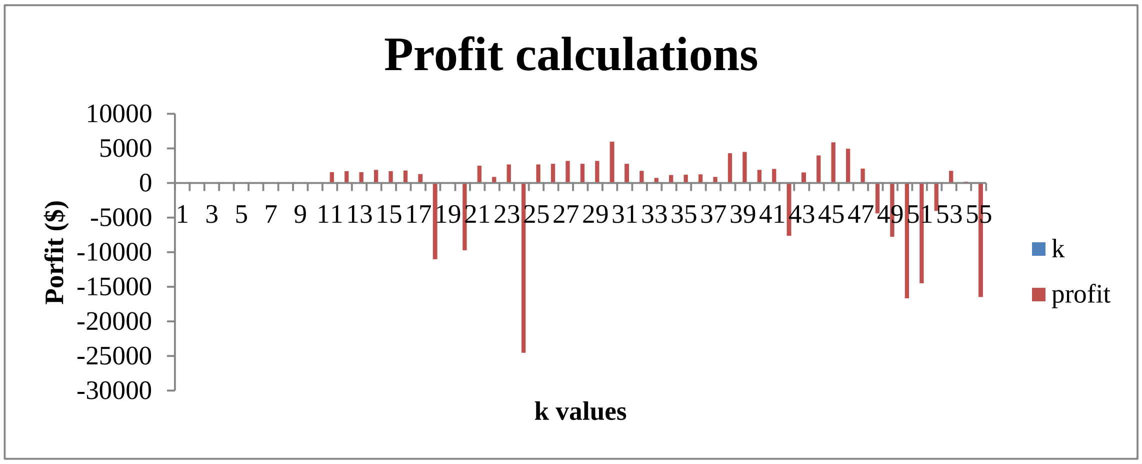


Figure 25: Profits obtained from different values of k (number of events)

Case Study

We consider the following trading time horizon: Tuesday February 17, 2015 from 9:00 am to 2:00 pm. We assume that a single trader sends bid and ask limit orders at the best bid and the best ask. We assume that these orders are executed immediately whenever the desired size is matched. We used a number of events k=30 events.

Trading strategy 1:

This is a simple strategy where a single trader trades 1 natural gas futures contracts when the spread crossing is "up" and 1 when it is "down".

Trading strategy 2:

This strategy is more elaborate than strategy 1 and uses the amplitude of spread crossing to determine the number of contracts to buy and sell. Below is the detailed strategy.

"oneup" (1 tick upward) corresponds to:

$$0 < amplitude \leq 0.001 \qquad (1)$$

➔ Buy long 1 contract, sell short 1 contract after 30 events.

“twoup” (2 ticks upwards)corresponds to:

$$0.01 < amplitude \leq 0.002 \qquad (2)$$

➔ Buy long 2 contract, sell short 2 contracts after 30 events.

“twreeup” (3 or more ticks upwards) corresponds to:

$$0.002 < amplitude \qquad (3)$$

➔ Buy long 5 contracts, sell short 5 contracts after 30 events.

Three values for a downward spread crossing are defined in a similar fashion:

“onedown” (1 tick downwards) corresponds to:

$$0 < amplitude \leq 0.001 \qquad (4)$$

➔ Buy short 1 contract, sell long 1 contract after 30 events.

“twodown” (2 ticks downwards) corresponds to:

$$0.001 < amplitude \leq 0.002 \qquad (5)$$

➔ Buy short 2 contracts, sell long 2 contracts after 30 events.

“twreedown” (3 or more ticks downwards) corresponds to:

$$0.002 < amplitude \qquad (6)$$

➔ Buy short 5 contracts, sell long 5 contracts after 30 events.

“stat” (stationary) corresponds to:

$$amplitude \leq 0 \qquad (7)$$

➔ Do nothing.

Trading strategy 3:

This strategy uses the inventory optimization model described previously.

Monte Carlo simulation is used to generate buy or sell prices when a spread crossing happens. We apply this model to predict natural gas futures prices (either best bid or best ask). The initial value of the natural gas futures contract being $2700.

We apply the standard inverse normal distribution on random numbers between 0 and 1. It provides us with a set of random numbers between -3 and 3 that are then used to predict the series of log-returns and the future values of the futures contract.

$$Logreturn(t) = Drift(mean) + Volatility \times z_t$$

$$Price(t) = Price(t-1) \times e^{Log-return(t)}$$

The values of the drift and the volatility are updated at beginning of each trading window and are derived from the previous trading window. With an initial price of $2700, a drift of 0.01% and a volatility of 0.2%, we obtain an instance of the price fluctuation for 10 events. The following figure shows how natural gas futures prices are generated using Monte Carlo simulation.

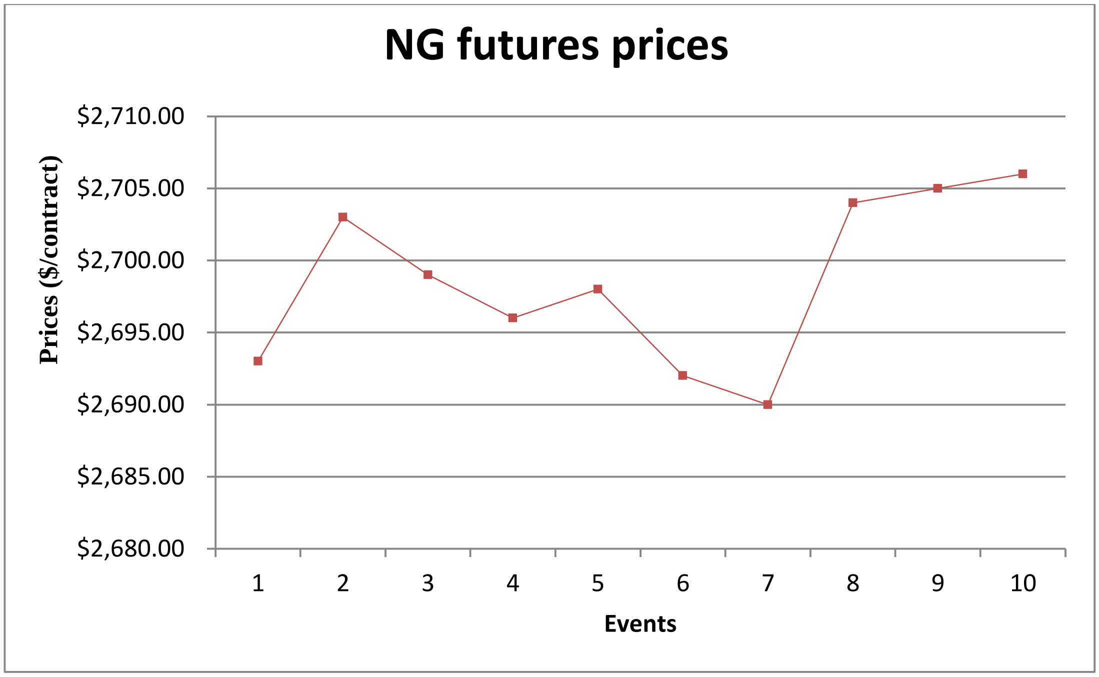


Figure 26: Monte Carlo simulation example

The corresponding volumes for each price level are obtained by using a random generator to produce a random sequence of numbers ranging from 1 to 20. The maximum volume is set to 20 contracts. This is conservative estimate that reflects the maximum number of natural gas futures contract that could be traded during a given event. By using a fixed seed, the sequence of numbers can be reproduced.

The simulated values of prices and volumes are then plugged into the optimization model. Table 14 below shows the input to the optimization model for 10 events.

Event	1	2	3	4	5	6	7	8	9	10
Prices	2693	2703	2699	2696	2698	2692	2690	2704	2705	2706
Volumes	5	1	6	5	4	18	4	3	18	9

Table 14: Input for the optimization model

The output for the optimization model is as follows:

Event	0	1	2	3	4	5	6	7	8	9	10
Volumes		12	0	0	9	0	18	14	3	17	0
Inventory	0	7	6	0	4	0	0	10	10	9	0

Table 15: Output for the optimization model

From the output of the optimization model, we observe that the optimal volumes to buy are generated. The inventory constraints are satisfied at all times. The corresponding optimal objective function is Z= $196,839. This would be the optimal cost for buying natural gas contracts and keeping the inventory for 10 events. In reality, traders buy on margin, thus, only a small portion (around 10%) of the contract’s value is needed to buy a contract. For each trading window, the optimal volumes computed are matched with the actual volumes from the top of the book data. We assume in this single trader model that orders made have the priority to be fulfilled, the inventory levels are updated and pending orders are cancelled. In addition, in a market making strategy, traders buy and sell simultaneously.

The table below summarizes the results obtained from the 3 strategies. We assumed that there are no transaction costs or fees incurred by the trader.

Strategy #	Number of executed orders	Number of cancelled orders	Total Profit
1	1645	2423	$3170
2	2289	3562	$4910
3	2177	1895	$2260

Table 16: Comparison between three trading strategies

From the table above, the most profitable strategy is strategy 2. Thus predicting the amplitude of bid-ask spread crossing on top of its direction leads to more profit. Strategy 3 used the least number of order cancellations but was the least profitable strategy. This happens because orders would wait longer in the order book before being executed. In case a fee on order's cancellations is imposed, strategy 3 could be more profitable. This is relevant because regulators worldwide are starting to impose fees on high-frequency transactions. For instance, Italy established a levy on trades occurring every 0.5 seconds and faster.

Several parameters can result in different values for the profit:

- The trading strategy considered by the user.
- The number of events k that is considered in fitting the model as well as calculating the profit.

- The number of contracts to be traded when different amplitudes of spread crossing are detected.

The aforementioned models as well as the trading strategies can be applied on real data from the stock market. See Appendix A for a description of historical data.

Chapter 6: Conclusion and Future Work

The objective of this model was to study the limit order book for high frequency trading using simulated natural gas futures data. We used 60 predictors for 10 level data based on best ask and bid prices and volumes from front month natural gas futures historical data. Our main target was to predict the direction and amplitude of bid-ask spread crossing. We developed a machine learning model that uses multiclass support vector machines to predict our target, considering k number of events. We compared the performance of the SVM model to the one of a Decision tree model. The SVM model had a higher accuracy and outperformed the DT model.

We implemented trading strategies for a single trader based on the spread crossing direction and amplitude in addition to the selected number of events which resulted in making profit. The profit amount can be affected by considering different trading strategies, thus changing the number of events and the number of contracts impacts the total profit. This finding shows that faster traders can adversely select slower traders leading to an unfair market. In case a fee on order's cancellations is imposed, the inventory optimization strategy would be more profitable. This is relevant because regulators worldwide are starting to impose fees on high-frequency transactions. For instance, Italy established a levy on trades occurring every 0.5 seconds and faster.

It is our future plan to investigate other machine learning techniques such as ANNs, logistic regression and genetic algorithms. They should be easier to implement using the existing framework.

Other plans include studying the impact of speed based taxes on HFT. For instance, implementing a tax on cancelled orders might reduce the liquidity but also might make the market fairer.

References

Ait-Sahalia, Y. and Saglam M. (2013). High-Frequency Traders: Taking Advantage of Speed. Working Paper.

Alden, W. (2014). High-Frequency Trader Charged With Manipulating Commodity Prices. DealBook.

Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning Research*, *1*, 113-141.

Amihud, Y. and Mendelson, H. (1980). Dealership market: Market-making with inventory. *Journal of Financial Economics*, *8*(1), 31-53.

Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217-224.

Baron, M. D., Brogaard, J., Hagströmer, B., and Kirilenko, A. A. (2015). Risk and return in high frequency trading. *Available at SSRN 2433118*.

Baruch, S., & Glosten, L. R. (2013). Fleeting orders. *Columbia Business School Research Paper*, (13-43).

Bates, J. (2015). Post Flash Crash, Regulators still use bicycles to catch Ferraris. Traders Magazine.

Bernanke, B. (2009). *The Crisis and the Policy Response.* Federal Reserve.

Bernanke, B. (2013). *2013 Monetary Policy Release.* Federal Reserve.

Biais, B., Foucault T., and Moinas S. (2011*). Equilibrium High Frequency Trading*. Working Paper.

Blanco, C., Soronow D. and Stefiszyn P. (2002). *Mulitfactor models of the forward price curve II*. Commodities Now.

Blanco, C. and Stefiszyn. P. (2002). *Valuing natural gas storage using seasonal principal component analysis*. The Risk Desk.

Bodie Z., Kane A. and Marcus A. (2009). *Essentials of Investments*, MCGraw-Hill/Irwin Series in Finance.

Boehmer, E., Fong, K. and J. Wu. (2014). International Evidence on Algorithmic Trading. Working Paper.

Boogert, A. and de Jong C. (2008). *Gas storage valuation using a Monte Carlo method*. Journal of Derivatives 15 81–98.

BP. (2012). *Statistical Review of World Energy*.

BP. (2015). *Energy Outlook 2035.*

Bringedal, B. (2003. "Valuation of Gas Storage: A real options approach."Department of Industrial economics and technology management, NTNU.

Brogaard J., Hendershott T. and Riordan R. (2012). *High Frequency Trading and Price Discovery*, Review of Financial Studies 27 2267-2306.

Browning, E.S. (2007). *Exorcising Ghosts of Octobers Past*. The Wall Street Journal (Dow Jones & Company), C1–C2.

Brush, S., Schoenberg, T. and Ring, S. (2015). How a Mystery Trader with an Algorithm May Caused the Flash Crash. BloombergBusiness.

Cartea and Figueroa, *Pricing in Electricity Markets: A Mean Reverting Jump Diffusion Model with Seasonality*, University of London, December 2005.

Cafiero, C., Bobenrieth, E.S.A., Bobenrieth J.R.A. and B.D. Wright. (2009). “The Empirical Relevance of the Competitive Storage Model.” Journal of Econometrics, special issue.

Cespa, G. and Foucault, T. (2008). Insiders-outsiders, transparency, and the value of the ticker. *Available at SSRN 1117581*.

Chaboud, A., Chiquoine, B., Hjalmarsson, E., and Vega, C. (2009). Rise of Machines: Algorithmic Trading in the Foreign Exchange Market. Internation Finance Discussion Papers 980.

Chen, C. H., He, D., Fu, M., & Lee, L. H. (2008). Efficient simulation budget allocation for selecting an optimal subset. *INFORMS Journal on Computing*, *20*(4), 579-595.

Chen S., Chen H. and Zhou Y. (2010). Order Book Simulator and Optimal Liquidation Strategies, June 2010, Stanford University MS7E 444 Investment Practice Course Project.

CME Group. (2015). Collateral Acceptance Criteria for Exchange Traded Funds and Stocks. Cmegroup.com/clearing

Covel, M. W. (2009). The Complete Trutle Trader. How 23 Novice Investors Became Overnight Millionaires.

Crammer, K., & Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, *2*, 265-292.

Cvitanic, J., and Kirilenko, A. A. (2010). High frequency traders and asset prices. *Available at SSRN 1569075*.

Dantzig, G. (1963) Linear Programming and Extensions. Princeton University Press.

de Jong, C. and Huisman, R. (2002). Option formulas for mean-reverting power prices with spikes. Rotterdam School of Management, Erasmus University Rotterdam, Working Paper.

Dicolo, J. A. and Rogow, G. (2012). Gas Market Stung by Rapid Traders. The Wall Street Journal.

Dudley B. (2012). BP Group Chief Executive, *BP Statistical Review*.

Dudley B. (2015). BP Group Chief Executive, *BP Statistical Review*.

ElBoghdady, D. (2013). High-frequency trading firm Panther Energy fined in 'spoofing' case. Washingtonpost.

Energy Information Administration. (2014). Gas Storage report: Overview, data, analysis and projections. http://www.eia.gov/naturalgas

Energy Information Administration. (2015). Annual Energy Outlook 2015 with projections to 2040.

Energy Information Administration. (2016). Naural Gas Spot and Futures Prices (NYMEX). https://www.eia.gov/dnav/ng/NG_PRI_FUT_S1_D.htm

Eydeland, A. and Wolyniec K. (2003). *Energy and Power Risk Management: New Developements in Modeling, Pricing, and Hedging.* Hoboken: John Wiley & Sons.

Fennel, M. (2011). Natural Gas Transportation, Storage and Use (Science of Electricity).

Fischer, B. (1976). The pricing of commodity contracts. Journal of Financial Economics 3 : 167-179.

Fodra, P. and Labadie, M. (2012). High-frequency market-making with inventory constraints and directional bets. *arXiv preprint arXiv:1206.4810*.

Foucault T., Hombert, J., and Rosu I. (2015). *News Trading and Speed*. The Journal of Finance.

Garman, M. B. (1976). Market microstructure. *Journal of financial Economics*, *3*(3), 257-275.

Geiger, K. and Mamudi, S. (2014). High-Speed Trading Faces New York Probe Into Fairness. BloomberBusiness.

Gerald, P. and O'Driscoll, Jr. (2007). *Subprime Monetary Policy, Investors Have Come to Bank on the Fed's Backing of Risky Ventures*, Volume: 57, Issue: 9.

Gill, P. E., Murray, W., & Wright, M. H. (1981). Practical optimization.

Gillespie, P. (2015). What interest rate increase mean for real people. CNN Money.

Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J., and Howison, S. D. (2013). Limit order books. Quantitative Finance, 13(11), 1709-1742.

Gray J. and Khandelwal P. (2004). Towards a realistic gas storage model. Commodities Now.

Gross,D. and Harris,C.M.(1985). *Fundamentals of queuing theory*. New York: John Wiley & Sons. 2nded.

Guéant, O., Lehalle, C. A., & Fernandez-Tapia, J. (2012). Optimal portfolio liquidation with limit orders. *SIAM Journal on Financial Mathematics, 3*(1), 740-764.

Guo, X., de Larrard, A., and Ruan Z. (2012). Optimal order placement in a limit order book. Preprint, UC Berkeley

Guo, X. and Ruan Z. (2013). Dynamics of the order position in a limit order book. Working paper, UC Berkeley

Gutierrez, A., Nafidi, A., Gutierrez Sanchez, R. (2005). *Forecasting total natural-gas consumption in Spain by using the stochastic Gompertz innovation diffusion model*. Applied Energy; 80(2), 115–124.

Harrell, C., Bowden, R., and Ghosh, B. K. (2000). *Simulation using promodel*. McGraw-Hill Higher Education.

Hasbrouck, J., Saar, G. (2009). Technology and liquidity provision: the blurring of traditional definitions. Journal of Financial Markets 12, 143–172.

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

Hawkes, A. G. and Oakes, D. (1974). A Cluster Process Representation of a Self-Exciting Process, Journal of Applied Probability, Vol. 11, No. 3, pp. 493-503

Hecht, A. (2016). Natural Gas: The Magnet Of Lower Price In A Fuel That Loves To Toast And Roast. Seekingalpha.

Hendershott, T., and Seasholes, M. S. (2006). Market maker inventories and stock prices. *Available at SSRN 890860*.

Hendershott, T., Jones, C. and Menkveld, A.J. (2011). *Does Algorithmic Trading Increase Liquidity?* Journal of Finance 66:1-33.

Hendershott, T. and Riordan, R. (2012). *Algorithmic Trading and the Market for Liquidity*. Journal of Financial and Quantitative Analysis, forthcoming.

Hendershott, T., and Menkveld, A.J. (2014). *Price Pressures*. Journal of Financial Economics 114:405-423.

Hippert, H.S., Pedreira, C.E. and Souza, R.C. (2001). *Neural networks for short-term load forecasting: a review and evaluation*. IEEE Transactions on Power Systems; 16(1), 44–55.

Ho, T. and Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial economics*, *9*(1), 47-73.

Holland, A. 2008. *A decision support tool for energy storage optimization*, in Proceedings of 20th IEEE International Conference on Tools with Artificial Intelligence, 2: 299–306.

Hull, J.C. (2008) *Fundamentals of Futures and Options Markets*, Pearson Prentice Hall, Upper Saddle River, NJ: Prentice Hall.

Hult, H. and Kiessling, J. (2010). Algorithmic trading with Markov chains, Doctoral thesis, Stockholm University, Sweden.

Ince, H.; Trafalis, T. (2008). *Short term forecasting with support vector machines and application to stock price prediction*, International Journal of General Systems 37(6): 677–687.

Jacobs, B. I., Levy, K. N., & Markowitz, H. M. (2004). Financial market simulation. *The Journal of Portfolio Management*, *30*(5), 142-152.

Jang, J.S., Sun, C.T., Mizutani, E. (1997). Neuro-Fuzzy and Soft Computing: *A Computational Approach to Learning and Machine Intelligence*. Prentice-Hall International, Inc.

Jarrow, R. A., and Protter, P. (2012). A dysfunctional role of high frequency trading in electronic markets. *International Journal of Theoretical and Applied Finance*, *15*(03), 1250022.

Jensen, G.R., Johnson, R.R. and Mercer, J.M. (2000). Efficient Use of Commodity Futures in Diversified Portfolios. Journal of Futures Markets, 20 : 489-506.

Jensen, G.R., Johnson, R.R. and Mercer, J.M. (2002). Tactical Asset Allocation and Commodity Futures. Journal of Portfolio Management, 28 : 100-111.

Jerry, B. (1984). *Discrete-event system simulation*. Pearson Education India.

Kercheval, A. N., and Zhang, Y. (2015). Modelling high-frequency limit order book dynamics with support vector machines. *Quantitative Finance*, *15*(8), 1315-1329.

Kidnay, A. J., Parrish, W. R., & McCartney, D. G. (2011). *Fundamentals of natural gas processing* (Vol. 218). CRC Press.

Kimoto, T.; Asakawa, K.; Yoda, M. and Takeoka, M. (1990). *Stock market prediction system with modular neural network*, in Proceedings of the International Joint Conference on Neural Networks, 1: 1–6

Khotanzad, A., Elragal H. and Lu, T, L. (2000). *Combination of artificial neural-network forecasters for prediction of natural gas consumption*. IEEE Transactions on Neural Networks; 11(2), 464 – 473.

Kyle, A. (1985). Continuous Auctions and Insider Trading. Econometrica 53, 1315–1335.

Lati, R. (2009). The Real Story of Trading Software Espionnage. Advanced Trading.

Lauricella, T. (2011). "Traders Exit High-Speed Lane". The Wall Street Journal.

Law, A. M., & Kelton, W. D. (1991). Simulation modeling and analysis, McGraw-Hill. *New York*.

Lin, Tom C. W. (2013). *The New Investor*. 60 UCLA Law Review 678 (2013).

Lorenz, J. M. (2008). *Optimal trading algorithms: Portfolio transactions, multiperiod portfolio selection, and competitive online search* (Doctoral dissertation, Technische Universität München).

Luckock, H. (2001). A statistical model of a limit order market. *Sidney University preprint (September 2001).*

Maria, A. (1997, December). Introduction to modeling and simulation. In *Proceedings of the 29th conference on Winter simulation* (pp. 7-13). IEEE Computer Society.

Markowitz, H.M. (March 1952). "*Portfolio Selection*".The Journal of Finance, 7 (1): 77–91.

Markwat, T. D., Van Dijk, D. J., Swinkels, L., & De Zwart, G. J. (2008). The Economic Value of Fundamental and Technical Information in Emerging Currency Markets.

Moallemi, C. (2013*). High-Frequency Trading and Modern Market Microstructures*. Talk at the Graduate School of Business Columbia University.

Moallemi, C. C., and Saglam, M. (2012). Dynamic portfolio choice with linear rebalancing rules. *Available at SSRN 2011605*.

Mokhatab, S., & Poe, W. A. (2012). *Handbook of natural gas transmission and processing*. Gulf Professional Publishing.

Moniz, E. J., Jacoby, H. D., Meggs, A. J. M., Armtrong, R. C., Cohn, D. R., Connors, S. R., ... & Kaufman, G. M. (2011). The future of natural gas. *Massachusetts Institute of Technology report*.

Murry, D. and Zhu, Z. (2004). EnronOnline and informational efficiency in the U.S. natural gas market. The Energy Journal, 25(2), 57-74.

O'Sullivan, A. and Sheffrin, S. (2003). *Economics: Principles in action*. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall, pp. 197, 507.

Pagnotta, E. (2010). Information and liquidity trading at optimal frequencies.*Available at SSRN 1684297*.

Pagnotta, E. and Philippon, T. (2011). Competing on speed. The National Bureau of Economic Research. Working Paper.

Pardalos, P. M. (2009). *Optimization in the energy industry* (p. 533). S. Rebennack, & M. Scheidt (Eds.). Berlin: Springer.

Phua, P.K.H., Ming, D. and Lin, W. 2000). *Neural network with genetic algorithms for stocks prediction*, in The 5th Conference of the Association of Asian-Pacific Operations Research Societies. Singapore.

Popper, N. (2012). High-Speed Trade Giants to Merge, The New York Times.

Pritsker,A.A.B, and O'Reilley,J.J.(1999).Simulation with Visual SLAM and Awesim. New York: John Wiley & Sons.

Radcliffe, R. C. (1997). *Investment: Concepts, Analysis, Strategy*. Addison-Wesley Educational Publishers, Inc. p. 134.

Rardin, R. L., & Uzsoy, R. (2001). Experimental evaluation of heuristic optimization algorithms: A tutorial. *Journal of Heuristics*, *7*(3), 261-304.

Roberts, N., Andersen, D. F., Deal, R. M., Garet, M. S., & Shaffer, W. A. (1983). *Introduction to computer simulation: the system dynamics approach*. Addison-Wesley Publishing Company.

Rose P. and Marquis M. (2007). *Money and Capital Markets*, MCGraw-Hill/Irwin Series in Finance.

Ross, K. (2012). Trading Platform PDQ Entreprises LLC. The Wall Street Journal.

Rothlauf, F. (2011). Summary. In *Design of Modern Heuristics* (pp. 221-225). Springer Berlin Heidelberg.

Sanchez-Ubeda, E.F. and Berzosa, A. (2007). *Modeling and forecasting industrial end-use natural gas consumption*. Energy Economics; 29(4), 710-742.

Schapiro, M. (2010). *Testimony Concerning the Severe Market Disruption on May 6, 2010.*

Schuyler, J. R. and Newendorp, P. D. (2013). Decision Analysis for Petroleum Exploration.

Sharf, S. (2014). Fed Sets October End Date For Monthly Asset Purchases. Forbes Investing.

Shek, H. H. S. (2011). Modeling high frequency market order dynamics using self-excited point process. *Available at SSRN 1668160.*

Shroeder P. (2014). Senate panel to dig in to high-frequency trading. thehill.com

Strasburg J. and Paterson S. (2012). *The Wall Street Journal*.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B. and Vandewalle, J. (2002). *Least Squares Support Vector Machines*. Singapore: World Scientific.

Suzuki, K. and Turner, D. (2005). *"Sensitive politics over Japan's staple crop delays rice futures plan*". The Financial Times.

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. New York: Random House.

Thompson, M. 2011. *Counterparty credit risk pricing and measurement of swaption portfolios*. Working paper, Queen's School of Business.

Thompson, Matt, Matt Davison, Henning Rasmussen. 2009. *Natural gas storage valuation and optimization: A real options application*. Naval Research Logistics 56(3) 226–238.

Thompson, M.S. (2003). *Real options valuation and optimization of energy assets*. Ph.D. thesis, University of Western Ontario.

Trafalis, T., Ince, H.and Mishina, T. (2003). *Support vector regression in option pricing*, in Proceedings of Conference on Computational Intelligence and Financial Engineering. Workshop in soft computing for financial analysis. Hong Kong, China.

Tsai, C.F. and Wang, S.P. (2009). *Stock price forecasting by hybrid machine learning techniques*, in Proceedings of the International MultiConference of Engineers and Computer Scientists, 1: 755–760.

Vangheluwe, H. (2004). Modelling and Simulation Concepts. *MacGill University, Canada*.

Vapnik, V. (2013). The Nature of Statistical Learning Theory. Springer Science & Business Media.

Vapnik, V. N., & Vapnik, V. (1998). Statistical Learning Theory (Vol. 1). New York: Wiley.

Virtamo, J. (2005). Queueing theory. *Lecture Notes*.

Witten, I. H., Fran, E. and Hall, M. A. (2005). Data Mining : Practical Machine Learning Tools and Techniques.

Zhang, Y. (2013). Modeling High-Frequency Order Book Dynamics with Support Vector Machines (Doctoral dissertation, Florida State University).

Appendix A: Applications to the Stock Market

In this Appendix, market depth historical data for the stock market are described. The definitions below are based on LOBSTER, an online limit order book data tool designed with the goal of providing researchers with easy-to-use, high-quality limit order book data (LOBSTER, 2014).

Data description

Records of high-frequency trading activity are organized into a database with two major components, the "message" and the "orderbook" files (LOBSTER, 2014). The "message" file stores basic information on each trading event, such as time of occurrence and transaction type, price, volume and direction. The "orderbook" contains limit orders information for bid and asks events. Each entry of the order book groups ask and bid events on n different price levels (we take n = 10) along with their volume sizes. The best ask and best bid are listed first, and the next best second, etc. The following is a description of the message file and the orderbook file.

Message file

Time (sec)	Event Type	Order ID	Size	Price	Direction
⋮	⋮	⋮	⋮	⋮	⋮
34713.685155243	1	206833312	100	118600	-1
34714.133632201	3	206833312	100	118600	-1
⋮	⋮	⋮	⋮	⋮	⋮

Variable explanation.

- Time: Seconds after midnight with decimal precision of at least milliseconds and up to nanoseconds depending on the period requested
- Event Type:
    - 1: Submission of a new limit order
    - 2: Cancellation (partial deletion of a limit order)
    - 3: Deletion (total deletion of a limit order)
    - 4: Execution of a visible limit order
    - 5: Execution of a hidden limit order
    - 7: Trading halt indicator (detailed information below)

- Order ID: Unique order reference number
- Size: Number of shares
- Price: Dollar price times 10000 (i.e. a stock price of $91.14 is given by 911400)
- Direction:
    - -1: Sell limit order
    - 1: Buy limit order
    - Note: Execution of a sell (buy) limit order corresponds to a buyer (seller) initiated trade, i.e. buy (sell) trade.

Order book file

Ask	Ask	Bid	Bid	Ask	Ask	Bid	Bid	
Price 1	Size 1	Price 1	Size 1	Price 2	Size 2	Price 2	Size 2	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1186600	9484	118500	8800	118700	22700	118400	14930	...
1186600	9384	118500	8800	118700	22700	118400	14930	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Variable explanation

- Ask Price 1: Level 1 ask price (best ask price)
- Ask Size 1: Level 1 ask volume (best ask volume)
- Bid Price 1: Level 1 bid price (best bid price)
- Bid Size 1: Level 1 bid volume (best bid volume)
- Ask Price 2: Level 2 ask price (second best ask price)
- Ask Size 2: Level 2 ask volume (second best ask volume)
- ...