Modeling Price Dynamics on Electronic Stock Exchanges with Applications in Developing Automated Trading Strategies

by

Darrin Matthew Gershman

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Rudolf Riedl, Chair
Professor of Statistics

Katherine Ensor
Professor of Statistics

Steve Cox
Professor of Computational and Applied Mathematics

Houston, Texas
January, 2010
Abstract

Modeling Price Dynamics on Electronic Stock Exchanges with Applications in Developing Automated Trading Strategies

by

Darrin Matthew Gershman

This thesis develops models for accurate prediction of price changes on electronic stock exchanges by utilizing autoregressive and logistic methods. Prices on these electronic stock exchanges, also called ECNs, are solely determined by where orders have been placed into the order book, unlike traditional stock exchanges where prices are determined by an expert market maker. Identifying the significant variables and formulating the models will provide critical insight into the dynamics of prices on ECNs. Whereas previous research has relied on simulated data to test market strategies, this analysis will utilize actual ECN data. The models recognize patterns of asymmetry and movement of the shares in the order book to formulate accurate probabilities for possible future price changes. On traditional stock exchanges, price changes could only occur as quickly as human beings could enact them. On ECNs, computerized systems place orders on behalf of traders based on their preferences, resulting in price changes that reflect trader activity almost instantaneously. The quickness of this automation on ECNs forces the re-evaluation
of commonly held beliefs about stock price dynamics. Previous strategies developed for trading on ECNs have relied mainly on price fluctuations to gain profits. This thesis uses the formulated models to design profitable strategies that use accurate prediction rather than price variability.
Acknowledgments

I would like to thank my advisor, Dr. Rudolf Riedi, and my other committee members, Dr. Katherine Ensor and Dr. Steve Cox. I would like to thank all of the past graduate and undergraduate students who took part in this project. I would also like to thank the ECNs Island and Archipelago for supplying the data and to former project member Premvir Jain for obtaining the data from the ECNs. A special thanks goes to my fellow statistics department graduate students, particularly Alejandro Cruz Marcelo, Kevin DeHoff, and Stephanie Hicks, without whose help I could not have completed this thesis. Another special thanks goes to all of my friends and family, particularly my brother, my mother, and my father.
# Contents

Abstract ................................................................. ii

Acknowledgments ......................................................... iv

1 Introduction .................................................................. 1

1.1 Introduction to ECNs ................................................ 1

1.2 Outline ................................................................ 5

2 Literature Review/Background ........................................ 7

2.1 ECNs .................................................................. 7

2.1.1 History of ECNs ................................................ 8

2.1.2 Advantages of ECNs .......................................... 10

2.2 The Order Book .................................................... 12

2.2.1 Order Book Information .................................... 15

2.3 Types of Traders .................................................... 18

2.4 Lifetimes of Orders ................................................ 22

3 Order Book Statistics ................................................... 26

3.1 Data ................................................................. 26

3.1.1 Standardizing Price .......................................... 29

3.2 Market Statistics .................................................... 30

3.2.1 Arrival Rates .................................................. 30

3.2.2 Considering Trades as Adds ................................. 34

3.2.3 Market Development ......................................... 36
3.2.4 Distribution of Added Orders and Prices on the ECNs ... 38
3.2.5 Correlated Stocks/ECNs ................................. 39
3.3 Distribution of Orders ..................................... 41
  3.3.1 Trades and Cancellations ............................... 42
  3.3.2 Order Book Asymmetry ................................ 44
3.4 Distribution of Shares .................................... 47
  3.4.1 Correlation/independence of price/volume ............ 48
3.5 Lifetimes of Orders ....................................... 49
3.6 Comparing Island and ArcaBook .......................... 52
  3.6.1 Arbitrage Opportunities .............................. 55

4 Models and Results 59
  4.1 Price Changes ............................................. 59
  4.2 Epochs .................................................... 63
    4.2.1 Changes in Inner Shares ............................. 65
  4.3 Model Setup .............................................. 74
    4.3.1 Time Scale .......................................... 76
    4.3.2 Variable Selection .................................. 77
    4.3.3 Evaluating Models .................................. 82
  4.4 Fixed-time Method ...................................... 85
    4.4.1 Model Diagnostics .................................. 92
    4.4.2 Model Variations ................................... 102
  4.5 Fixed-time means model ................................ 108
Chapter 1
Introduction

This chapter introduces electronic stock trading on ECNs, the systems that enable people to trade stocks via the Internet, and then provides an overview of the work done in this thesis. An understanding of ECNs, which stands for Electronic Communications Networks or Electronic Crossing Networks, enables the development of models for predicting price changes based on the information contained in the order placed by users on the ECNs. Models and strategies formulated for trading on the New York Stock Exchange and similar forums cannot be applied to ECNs due to the rapid nature of activity involved in automatic trading and the differences in market structure between ECNs and other types of exchanges.

1.1 Introduction to ECNs

Typically, when news organizations report on stocks, they report the closing price of a stock and its change since the previous day. When the “price” or “value” of a stock is reported as an individual number, it is the price at which the last transaction (trade) occurred. However, ECNs are double auction exchanges, where the users who wish to buy and sell a certain stock can simultaneously place orders. This setup differs from markets like the New York Stock Exchange where the price is determined by the market maker and from over-the-counter auction markets like NASDAQ where trade prices are negotiated between the buyer and seller. The
double auction nature of ECNs means that a pair of prices, the prices at which the stock can be bought and sold, represents the value of the stock better than an individual number does.

On ECNs, trades are executed automatically when there is a match of orders placed by users who wish to buy and sell a stock at the same price. When an order is added and there is not a corresponding order at the same price or a better price for a transaction to occur, then the order is placed in the order book where it waits to be traded or cancelled. A better price is defined as a higher price if the person wants to sell and as a lower price if the person wants to buy.

In standard OTC markets, like NASDAQ, there are two types of orders that can be placed by a trader: markets orders and limit orders. Both types of orders specify which stock and whether the trader wishes to buy or sell. Market orders are executed immediately at the best available price. On the other hand, limit orders specify a price and therefore may or may not be executed immediately. If there is no corresponding order to be matched with the limit order, the limit order is added to the order book. The limit order remains in the order book until it is either matched with an added order on the opposite side to create a trade or until the user cancels the order. Market orders guarantee users that their orders will be executed immediately but make no guarantees about price. Limit orders do not guarantee execution, but the users know the price of a trade if it occurs. ECNs treat all orders as limit orders, so a market order would be equivalent to a limit order with its price set to the inner price on the opposite side of the order book to
ensure immediate execution [KO03].

Obviously, each stock has its own order book. At any time before a trade occurs, the trader who added the order may decide to cancel their order while it remains in the order book. While orders wait in the order book, they are placed in position first by price, then by time, with priority going to the lowest sell orders and highest buy orders, then to the ones that arrived earlier. At any time, the lowest sell order remaining in the order book is the best ask or inner sell price, and the highest buy order in the order book is the best bid or inner buy price. At this time, a person may purchase the stock at the best ask price or sell it at the best bid price, limited of course by the number of shares available at that price. The difference between the best ask price and the best bid price is the (bid-ask) spread. The best ask price must be larger than the best bid, because if this were not the case, then somebody would wish to buy at a price larger than someone else wants to sell. If this had been true, the ECN system would have immediately traded these orders, and the orders would have been automatically removed from the order book.

When considering price movements, it is important to understand the concepts of a buyer's market and a seller's market. When the demand to buy a stock is lower than the supply being sold, the expected result is a price drop as the buyers have a choice of from who to buy their stock. The result is called a buyer's market. Once the price drops, the demand to buy will increase and the desire to sell will decrease, thus restoring equilibrium to the system. Likewise, when there is a demand to buy that is greater than the desire to sell, the price will be driven up. The result is a
seller's market, since the great demand to buy means that buyers will pay a higher price until equilibrium is restored. This thesis studies the volume and distribution of orders in the order book and changes to the order book to determine when a price change is about to occur. If an overabundance of orders enters on either the buy or sell side, a reactionary price change is likely to occur in order to restore balance to the system.

Throughout this thesis, the data used comes from the ECNs Island and Archipelago (ArcaBook). Most of the results and models are formulated using the Microsoft (MSFT) stock, which is the stock commonly used in competitions for testing market strategies. Other days with similar activity levels, such as Cisco (CSCO) and Yahoo (YHOO), will be used for models to verify the applicability of the models to multiple stocks. The general focus will be on the regular trading hours (9:30 AM - 4:00 PM), as the pre-market trading hours (7:00 AM - 9:30 AM) and after hours trading (4:00 PM - 8:00 PM) exhibit many anomalies that may adversely affect results.

Trading stocks via ECNs is a relatively recent development, so very little literature exists compared to the literature that exists about more established financial topics. None of the existing work uses order book dynamics to predict price changes. Although some literature exists on strategies that attempt to earn profit by trading on ECNs, those strategies rely on price fluctuation or other price trends instead of using predictive models.
1.2 Outline

This thesis is arranged as follows: First, an overview of existing literature on ECNs is presented. A summary of important results related to the order book follows. After that, models are developed that determine probabilities of different types of price changes. Using those models, strategies are designed for profitable trading on ECNs, and the results of those strategies are compared to existing strategies.

Chapter 2 begins with a history of stock trading and how it has developed from the early days of the New York Stock Exchange to the model usage of ECNs to trade stocks via the Internet. Then, existing literature on ECNs is summarized, featuring a variety of focuses and conclusions. Work on the distribution of prices, motivations of orders, and lifetimes of orders is presented. The chapter ends with an overview of existing market strategies and a summary of the market simulator used to test the profitability of those strategies.

Chapter 3 gives a detailed summary of the ECN order book. Orders are analyzed from their inception to their termination and at every step in between. The rate at which the orders arrive, how long orders last in the order book, and the distribution of shares are examined. The distribution of prices in relation to the inner prices and the distribution of orders in the order book is also explored in detail. In addition, the movement of inner prices on both Island and ArcaBook is discussed, and possible arbitrage opportunities are presented.

Chapter 4 formulates the models for predicting ECN price changes using the information gleaned in chapter 3. Before developing models, the appropriate time
scale and variables for the models are determined. The different models presented in this chapter perform a variety of functions. Some split the day into fixed widths of time and yield probabilities of different types of price changes in the next interval. Others yield probabilities of the next price change, regardless of when the change occurs. The models use a variety of order book information, including changes in shares, means of orders, and asymmetry of orders, to calculate the probabilities.

After the models are formulated in chapter 4, chapter 5 designs strategies that use the output of those models. The strategies buy and sell on the ECNs based on the results from the fitted models. These strategies employ a variety of techniques with regards to how long to hold onto a stock and limits imposed on the size of holdings. In addition, existing strategies are tested on the data to attempt to reproduce the results of the strategies that gained profit according to the authors. The profitability of the successful existing strategies is compared with the profitability of the strategies based on the models using the ECN data. The models and strategies are then extended to include data for multiple stocks or multiple days.
Chapter 2

Literature Review/Background

Although people have been trading stocks for hundreds of years, the rapid development and expansion of the internet in recent years has had a drastic impact on the way many people trade stocks. As with any new technology used in the business world, gaining an understanding of the dynamics may provide an advantage over others. This chapter introduces electronic stock trading and explains the importance of the information contained in the order book. In addition, a detailed look at the existing literature about relevant topics in stock trading is presented, focusing on trading via electronic stock exchanges. Much of the relevant literature has been published since the year 2000 due to the recency of the emergence of the internet as a tool for stock trading.

2.1 ECNs

The two main types of markets for stock trading are quote-driven markets and order-driven markets. A quote-driven market is one where a specialist, also known as a market maker, sets the bid and ask quotes. One famous example of a quote-driven market is the New York Stock Exchange (NYSE). The dynamics of quote-driven markets are generally slow, as the market maker personally adjusts prices based on trader activity and external information. In addition to setting the prices, market makers provide liquidity by trading for themselves or the parties that they
represent.

Order-driven markets do not have a market maker, and the traders place orders that provide the liquidity [Ran04]. The bid and ask quotes are determined by the highest price of any bid (buy) order and the lowest price of any ask (sell) order. Examples of order-driven markets are ECNs, which stands for Electronic Communication Network or Electronic Crossing Network [SR05]. Important ECNs include Island, owned by NASDAQ, and Archipelago (ArcaBook), owned by the New York Stock Exchange.

2.1.1 History of ECNs

Stoll [Sto06] describes the history of stock trading and how it has changed over the years, focusing on how ECNs have changed the way people trade stocks. The New York Stock Exchange was founded in 1792 by twenty-four brokers. All trading was done in person at a coffee house where the traders negotiated over a price. Today, there are 1366 seats on the NYSE. Each seat is owned by a specialist, a floor broker, or a brokerage firm. The cost of seats has risen dramatically over the years, reaching 2.15 million dollars in 2001, up from $225,000 in 1980. The orders used to be carried by hand, while now most actions are computerized with very little human interaction. Today, most of the orders that are sent to the NYSE are sent electronically via SuperDot, which sends orders directly to the specialist. Specialists are simultaneously acting as brokers for orders sent to them to be entered into the book while trying to trade for themselves, which has led to some controversy over a conflict of interests [Sto06].
The “traditional” stock exchanges, such as NYSE and NASDAQ, realized the value of electronic trading and merged with existing ECNs. NYSE merged with Archipelago in 2005, and NASDAQ merged with Instinet. Instinet and Archipelago are the two largest and most dominant ECNs by volume, as shown by Instinet’s merging with Island and Archipelago’s merging with Redibook. While stocks have a primary listing on one of the major markets, a stock may be listed on multiple markets at one time. By 2003, one out of every seven NASDAQ trades occurred on the Island ECN [KO03]. In 2004, 42% of the share volume in stocks whose primary listing was NASDAQ were traded on ECNs [Sto06].

For many years, NASDAQ operated under some controversy regarding the information available regarding price quotes. Often, certain quotes would be available only to certain dealers and would not be available to the general public. The Securities and Exchange Commission (SEC) decreed its order handling rules in 1997, which required the market makers on NASDAQ to reveal all customer limit orders. This information created greater competition between customers and dealers, resulting in smaller spreads. The order handling rules also forbade dealers from quoting better prices to other dealers than to ordinary traders [Sto06].

When trading on an ECN, where everything is automated, the time duration from adding an order until the completion of the transaction may be as short as micro-seconds. Of course, this time duration depends on whether there is a corresponding buy/sell order with which the order can be immediately matched. After a transaction occurs, the Depository Trust and Clearing Corporation is the
agency in charge of transferring the ownership of the shares and ensuring payment. This portion of the process is generally done within three days of the transaction on the ECN [Sto06].

In 1971, Fischer Black first suggested the idea of a stock exchange that was fully automated. Black is best known in the econometric world for the Black-Scholes equations. He envisioned a computerized stock exchange with lower trading costs that would not discriminate against either big or small traders. Although he died in 1995, his vision came to fruition a few years later when the SEC authorized the creation of ECNs in 1998 [Sto06].

2.1.2 Advantages of ECNs

The advantages of using these ECNs over traditional markets include charging lower costs for performing a transaction, offering longer trading hours (7:00 AM to 8:00 PM instead of 9:30 AM to 4:00 PM), and (most importantly) giving their members the ability to observe the entire order book. Instead of only seeing the inner prices at which the user can buy or sell the stock, the user gets to see every currently existing order in the order book at that time. Users must have an account directly with the ECN or have an account with a brokerage house that has access to the ECN [Sto06].

Barclay, Hendershott, and McCormick [BHM03] explain that there are many advantages for traders to use ECNs over market makers. There are lower trading costs, faster executions, greater operational efficiency, improved limit order exposure, and anonymity to other traders. Anonymity is important because certain
traders’ actions may have a dramatic impact on the actions of others. No such impact is felt if the orders are seen but the identities of the traders associated with those orders remain unknown [BHM03].

Stoll also provides several reasons why ECNs have advantages over other types of markets. The automation of the ECNs allow trades to be executed without the involvement of humans, such as dealers, market makers, and specialists who may have something to gain. Stoll also mentions the aforementioned advantage of anonymity, stating it would not be possible in traditional markets where orders were handed directly to a dealer. Stoll agrees with Barclay, Hendershott, and McCormick in noting that traders may prefer ECNs because they are cheaper and quicker than traditional markets. ECNs charge lower transaction fees, and the execution generally occurs in less than a second. The last reason given is that ECNs can be programmed by users to act only if certain conditions are met. For example, a trader, who may only wish to buy a stock for a certain price, has the ability to automatically place an order as soon as the stock is available at the specified price [Sto06].

The preceding papers discuss how order-driven markets like ECNs are faster and cheaper than quote-driven markets like the NYSE. Obviously, there are costs associated with any type of trading, including commissions paid to brokers and losses that may result from trading. On ECNs, the commissions paid to brokers are lower because the costs to handle orders are lower. In addition, the bid-ask spread is generally smaller on ECNs, creating smaller losses associated with a transaction.
Also, electronic trading happens much faster, so the costs associated with time and information lost because of delays are reduced [Sto06].

2.2 The Order Book

Thus far, the order book has been established as the place where buy/sell orders that have been added but not traded wait to either be cancelled by the user or matched with an incoming sell/buy order. However, much information about the stock is contained within the order book, and eliciting that information will be necessary to the price models developed later. This section examines the order book and summarizes others’ conclusions about information contained in the order book.

The bid-ask spread is a very important characteristic of a stock at any given time. Stoll [Sto06] describes how the spread has decreased over the years as the precision of the prices has changed. When the minimum tick size for price was 1/8 of a dollar, the minimum spread could be 12.5 cents. The minimum tick was reduced by half in 1997, resulting in 6.25 cents as the minimum spread. In 2001, the minimum was reduced to one cent, where it remains today. Stocks such as Microsoft that once had average spreads over 16 cents had their spread reduced to less than 2 cents after the changes that resulted in 2001 [Sto06].

One concept crucial throughout this thesis is whether a stock is “spread-minimal.” Although it sounds oxymoronic, this term is being coined to indicate that a stock’s spread, or difference between the inner sell and inner buy, is as small as it can get (1 cent) for at least 99% of regular trading hours. Examples of stocks with this
quality are Microsoft (MSFT) and Cisco Systems (CSCO). During regular trading hours, the spread between the inner buy and inner sell for Microsoft is greater than 1 cent for only 92.2 seconds, or 0.5% of the time, with a longest duration of 5.5 seconds. For Cisco, the spread is greater than 1 cent for only 5.9 seconds, or 0.03% of the time, with a longest duration of 0.8 seconds. On the other hand, the Google stock spends more than 99% of the day with a spread larger than 1 cent.

Why is it important for a stock to be "spread-minimal?" First of all, these stocks generally have very high activity, with Microsoft averaging 350 orders being added every minute and Cisco averaging 296 orders being added every minute during regular trading hours. This high level of activity allows for taking a subset of the data while still having enough information to draw valid conclusions. However, the main reason is the minimization of the spread between the inner buy and inner sell orders. When the spread is larger than 1 cent, it is possible for an order to be added on either the buy or sell side that then becomes the inner price. For a stock that is spread-minimal, a price change can only occur after a series of cancellations or trades that empties out the shares of the inner price on either the buy or sell side. The large quantity of shares at the inner prices ensure that any price changes observed are due to actual changes in the value of the stock and not due to random chance.

Knowing when the inner buy and sell prices are about to change is crucial for the development of price models and profitable market strategies later in this thesis. There are a few ways in which an inner price may change. The following are the
ways in which the best bid (highest buy order price) might change, but the ways in which the best sell will change are symmetric. First, there can be a new buy order added to the order book that is higher than the highest bid price and still smaller than the lowest sell order. This new order will take the place of the previous best bid. Second, there can be a new buy order that is higher than the previous best buy and is also higher than the lowest sell and is large enough in volume to empty all of the sell orders less than or equal to it in price. A third way to change the inner buy is by having the inner buy order either traded (by an incoming sell order of the same price) or cancelled and no other order of that same price currently exists to take its place.

For spread-minimal stocks, the spread is $0.01 for most of the day. Therefore, the first way for an inner price change is not possible. The only ways these active stocks can change inner prices are by the removal of those orders or an incoming order that empties out the other side’s inner price queue and becomes the new best price. This last way corresponds to changes in both inner prices at the same time. The very active stocks will generally have multiple orders at the inner prices. Therefore, it is wise to look at the number of orders or the volume of these orders to see when a change is about to happen by looking at the order book “emptying” at these inner prices. If the buy side of the inner prices begins to “empty,” then one expects to see a price drop. A shift in “value” of a stock is revealed even though the inner prices do not change.
2.2.1 Order Book Information

With all of the orders being added to the order book, one obvious question people have asked is “Which orders give the most information?” Menkhoff [MS08] presents some characteristics that may identify more informed traders. Some of these traits, such as proximity to a financial center, are irrelevant to this thesis. ECNs provide traders with anonymity, so no information regarding who placed an order or where it originated is available. Menkhoff also says that informed traders place “medium-sized” orders, have a large trading volume, trade early in the session, place orders when the bid-ask spread is wide, and place orders when the order book is thin. The medium-sized orders and large trading volume may sound contradictory, but Menkhoff explains that the informed traders place many orders, each with medium size, to maximize the number of shares traded while never placing any order so large that it disturbs the balance of the order book. Again, the lack of information as to who places an order limits the applicability of these conclusions. The intraday dynamics of the order book show that early in the trading day is when the order book is thinnest, and when the order book is thinnest is when the spread is generally widest, so these 3 characteristics are related. In fact, the before-hours trading (7:00 AM - 9:30 AM) is not open to the general public but is only open to certain firms, which are inherently run by the traders believed to be the most informed. Menkhoff drew his conclusions based on data from MICEX in Moscow from March 11-21, 2002.

Barclay and Hendershott [BH03] examine the properties of trading after hours
in order to find informed traders. The regular trading day goes from 9:30 AM to 4:00 PM (EST), but ECNs have longer hours. The users on ECNs can place orders and trade from 7:00 AM to 8:00 PM. The pre-market trading goes from 7:00 AM to 9:30 AM, while after hours trading lasts from 4:00 PM to 8:00 PM, although these time periods have restrictions about who may trade. Barclay and Hendershott believe that the after hours traders are more informed in general than traders during regular trading hours. Although there are far fewer trades after hours and therefore less information revealed, they believe that each individual trade contains more information than an individual trade during regular trading hours. Also, they explain that there are fewer uninformed traders after hours, so there is less noise and more signal in the information revealed by the trades after hours [BH03]. Although the information from pre-market or after hours trading provides insight into traders, the data for these hours contains much less activity. As a result of the thinner data, the spreads during those hours are generally higher than one cent, making the data too thin for the models that will be developed later. These models will need to be created from the regular trading hours where the data is rich, and spreads are minimal.

Dufour and Engle [DE00] conclude that there is an increased number of informed traders on the NYSE when the markets are most active, while Manganelli [Man05] finds no evidence of increased numbers of informed traders during active times. Wong, Tan, and Tian [WDT08] look for evidence of informed trading on the Shanghai Stock Exchange. They agree with Dufour and Engle, concluding that
there is an increase in informed traders when activity on the Shanghai Stock Exchange is higher. They also find that their price formulation model agrees with the one presented by Foster and Vaswanathan [FV93].

Bouchaud, Potters, and Mezard discuss placement of orders on the Paris Bourse in their papers [BMP02] [PB03]. Their analyses reveal that the incoming limit orders follow a power-law distribution centered around the current inner bid and ask prices with a diverging mean. In addition, the average of the added orders has a maximum away from the inner price, and the distribution in the tails contains the information about the incoming orders. The volume of incoming orders have a Gamma distribution [BMP02]. They also claim that the dependence of price on volume is logarithmic [PB03].

Harris and Panchapagesan [HP05] study the role of specialists in the modern NYSE that combines traditional trading with a limit order book. Specialists serve a dual role on the NYSE, acting as brokers for clients while acting as dealers for their own accounts. This duality results in a conflict of interest, especially when considering that the specialists are able to see all of the limit orders coming into the order book, and they can use this information that is not available to ordinary traders for their own advantage. There is a bigger advantage to be gained for more actively traded stocks than less actively traded stocks. When the specialists can see the whole order book, the imbalance between the buy and sell sides of the order book can provide insight into future price changes. For this reason, the NYSE established rules preventing specialists from gaining such advantages. Buy and
sell orders provide the same amount of information, while large order convey more information than small orders [HP05].

2.3 Types of Traders

Ranaldo [Ran04] claims that previous studies had only focused on traders’ actions. His paper attempts to study the motivation behind these traders’ decisions. Ranaldo formulates hypotheses relating the volume (what he calls “thickness”) on the buy and sell sides of the order book to the aggressiveness of the orders being added. He concludes that added orders are more aggressive when the order book on that side is larger. Likewise, the added orders are less aggressive when the order book on the opposite side is larger. Perhaps the more significant conclusion from his paper is that the buy and sell order books are not symmetric.

Barclay et al. [BHM03] create the distinction between informed traders and uninformed traders. Informed traders need to trade before the market reflects the information they have, and therefore they place aggressive orders close to the inner price. Uninformed traders do not have any extra information and can afford to be more patient. They often place more cautious orders away from the inner price. In their paper, they conclude that ECNs attract a higher percentage of informed traders than of their market maker-driven counterparts. Due to the larger percentage of informed traders, the long-term price impact of an ECN trade is 50% larger than the long-term price impact for a trade transacted through a market maker.

Hollifield et al. [HMS04] examine the reasons and motivation for traders to place
their orders where they place them. They claim that the factors that determine
where someone places an order depends on the value the trader places on the stock
they wish to buy or sell, the current inner prices, the probability of the order being
executed, and the risks associated with the order being executed or not. The price
at which a trader places an order is a monotone function of the perceived value of
the stock according to the trader. When Hollifield, et al. tested their strategy on
the Stockholm Stock Exchange, they concluded that their monotone strategy works
on buy and sell orders separately, but it does not work when buy and sell orders are
jointly studied [HMS04]. The models developed in this thesis rely on information
contained at different prices in the order book. Thus, insight into the motivations
for the placement of orders provides insight into possible price movement.

Harris and Panchapagesan [HP05] categorize the traders as “pre-committed”
traders or “value-motivated” traders. The pre-committed traders already know
ahead of time what price they want and are only placing their order electronically
for cheaper transaction costs. They place their orders close to the inner price for
a high probability of their order being executed. If their order is not executed
and the inner prices move away from their order, they must “chase the market”
and add a new order at a worse price until a transaction occurs. On the other
hand, value-motivated traders place orders to get better prices, depending on the
value they place on the stock. They place orders close to the inner price for quick
execution and quick profit [HP05]. The main problem with this dichotomy of the
users is that there is no place for the traders who place their orders away from the
inner price. These traders will be important to the models developed later, making this categorization that omits them incomplete.

Bouchard, Mezard, and Potters [BMP02] argue that the majority of the incoming orders have prices away from the inner prices. They claim this behavior indicates more cautious traders who may be looking to buy or sell but are not willing to be so bold as to place orders right at the inner price. Although this may have been true when the paper was published, the market has evolved since then. The paper makes reference to active stocks that have 10 cancellations per minute. In the ECN data, during regular trading hours there are 230 cancellations per minute for Microsoft, one of the active, spread-minimal stocks [BMP02]. The difference in activity levels between the data sets indicates that the distribution of orders may also be different.

As these papers demonstrate, there are several motivations and strategies that people may use when placing their orders. There are more aggressive traders who will add an order close to the inner price with the hope of getting their order executed soon or immediately. Others are more patient and will place an order farther away from the inner price with the hope that the inner price will move in that direction and the order will be executed at a better price (a higher price if the user wants to sell and a lower price if the user wants to buy.) Under these assumptions, the distribution of the price of the added orders cannot be the result of a single distribution but perhaps will be better fitted with a mixture of the distributions from the different users.
The aggressive user can either place an order so that it will be immediately executed at the current inner price on the opposite side of the order book or place an order at the same inner price with the hope that the orders ahead will either be traded or executed so theirs will become first in line and traded. If the inner prices begin to move in the opposite direction, resulting in orders at “better” prices in front, the user may either wait and hope the price comes back or may cancel the order and place a new one. The models developed later in this thesis will provide information about the movement of stock price in the immediate future. Traders with access to such models would have insight into future price movements and may act accordingly.

The cautious trader places an order farther away from the inner price and hopes that as the prices move in that direction so they can have their order executed at a better price than the current inner price. The use of the word “cautious” to describe these traders does not mean they are any less crafty or less smart than the “aggressive” traders. In fact, they need to be very precise with the placement of their orders. They wish to place an order so that it gets traded at the best possible price for them. If they place an order too close to the inner price, they risk overpaying (if they are buying) or getting underpaid (if they are selling) for their shares. However, if they place an order too far from the inner price, they run the risk of their order never being executed and missing out. They may try to save a few cents per share, but the result may be missing a trade completely. In this case, they risk either not getting the stock they want or being forced to settle for a much
worse price.

The preceding conundrum is reminiscent of the optimization problems involving production of pieces of a certain length. If the desired length of produced pieces is 10 feet, then any pieces made less than 10 feet will have to be discarded, while pieces larger than 10 feet will have the excess removed as waste. The question posed is what should be the mean value of the produced pieces (if the variance is a given value) in order to minimize waste. For the cautious trader, the order placed too far from the inner price is the short (and therefore useless) piece because no trade occurs. The order placed too close to the inner price is analogous to the piece produced that was too large because there some margin for error that resulted in a less-than-optimal (but still somewhat satisfactory) situation. Therefore, producing the piece of almost exactly 10 feet is analogous to putting the price that will be reached by the inner price in the relatively immediate future but will not get greatly passed by. For example, if the inner prices are $25.00 and $25.01, and the user wishes to buy, an order of $24.95 may be entered. The trader may be frustrated if the inner price drops to $24.90 because he could have spent less. On the other hand, the trader may not get the order executed at all if the price drops to, for example, $24.98, and then increases away from $24.95.

2.4 Lifetimes of Orders

The lifetimes of the orders reveal important information about the order book. An order's lifetime is defined as the time between when the user adds it to the order book and when the order leaves the order book due to trading or cancelling. The
“age” of the orders in the order book tells whether the same orders remain in the order book or whether there is a constant flow of orders in and out of the order book throughout the day.

Several papers have been written exploring the lifetimes of orders in the order book. Hasbrouck and Saar [HS08] coined the term “fleeting orders” for those orders that spend fewer than 2 seconds in the order book. The paper explains that the “fleeting orders” behave more like market orders than limit orders and therefore justifies removing them when wishing to study the behavior of limit orders. They state that 27.7% of the orders (32.5% of the shares) in the data are fleeting orders, so less than \( \frac{3}{4} \) of the data is kept. However, these short-lived orders are often placed by aggressive traders close to the inner prices, and these orders provide important information. Therefore, for all results and models presented in this thesis, these short-lived orders will not be removed as was done in those other papers. Of all added limit orders, only 18.4% of them are eventually traded (partially or wholly). Almost 90% of the shares added are eventually cancelled [HS08]. The total is greater than 100% due to some orders being partially traded and partially cancelled.

Kenney [Ken07] and Chakrabarty [CHTZ06] use similar Cox Proportional Hazards (CoxPH) methods to formulate models for the time that elapses between when an order enters the order book and when it leaves the order book due to execution (trade) or cancellation. They perform some data cleaning that must be noted: All orders with prices more than 25 cents away from the inner prices are removed, as well as all fleeting orders.
The independent research presented by Kenney [Ken07] and Chakrabarty [CHTZ06] follow similar methods but yield different results. Kenney states that as an order is placed at a further distance from the inner price, the hazard rate for execution increases while the hazard rate for cancellation decreases. Chakrabarty states that the further an added order is placed from the inner price, the longer the time until execution and the greater chance the order has of being cancelled first. These results appear to be contradictory. Logically, an order will have a greater chance of being traded if it is added closer to the inner prices. If two orders are added on the same side of the order book and at different prices, the order closer to the inner prices must have a greater chance of being traded. Trades occur from the inner prices outward, so Chakrabarty’s results make logical sense while Kenney’s do not. Orders added further from the inner price cannot be traded until all of the orders with higher priority (closer to the inner price) are traded or cancelled first. These further orders will have longer to wait to be executed than the ones closer to the inner prices. Chakrabarty also concludes that the number of shares in an order does not affect that order’s probability of being traded [CHTZ06].

Lo [LMZ02] uses econometric and survival analysis techniques to model time until execution for limit orders. Orders can be partially cancelled or partially traded, leaving a portion of the shares in the order book, and Lo considers both the time from when an order is added to when it is first modified (“time-to-first-fill”) and the time from when an order is added until all shares are gone from the order book due to cancellation or trading or both (“time-to-completion.”) The
variables used for Lo's survival analysis model include the price of the limit order, the number of shares in the limit order, the bid-ask spread, and market volatility. The survival times of the orders depend on the price, but not on the number of shares [LMZ02]. Kooths et al. [KMR03] also use econometric models, but their models forecast inflation.
Chapter 3
Order Book Statistics

One of the definitive qualities of ECNs is that they are entirely order-driven, meaning that the prices are determined by orders placed by the users instead of by one individual person. Before price dynamics can be modeled, it is necessary to examine these orders more closely in order to determine what information will be useful for prediction of price changes. This chapter provides a detailed analysis of the orders and the order book that are the driving force for prices on ECNs. Among the questions to be answered are: At what rate are orders added, traded, and cancelled? How long do orders last (survive) in the order book? Is there any association between the price and shares of an order? Most importantly, what is the distribution of the prices and how does that information help predict future price changes?

3.1 Data

This thesis develops models for accurate prediction of price changes on ECNs using information contained in the orders. The models developed in this thesis utilize actual market data to evaluate how successful the predictions are. Before the formulation of models and market strategies, it is necessary to take a closer look at the data to be used in the formulation and evaluation of the models.

The data set that will be incorporated into the models comes from ECNs Is-
land (also called Inet), which is owned by NASDAQ, and Archipelago (ArcaBook), owned by the New York Stock Exchange (NYSE). The Island data includes all additions, deletions, and executions (trades) of orders. Every entry in the data is denoted with A, X, E, or P. An “A” indicates that an order was added. The data reveals whether the person adding the order wishes to buy or sell, the symbol for the stock the user wishes to trade, the price, and the number of shares. An entry with an “A” is added to the order book and waits to be matched with a corresponding order or to be cancelled by the user. An “X” indicates a cancellation and says how many shares the user wishes to cancel. If the number of shares being cancelled equals the number still in the order book, it is fully cancelled, and the order is no longer on the order book. If it is not fully cancelled, the order stays in the order book with the remaining shares. An “E” (for execution) indicates that an incoming order gets matched with an existing order already in the order book, thereby resulting in a trade. The data set tells which order it was matched with, the price, and for how many shares. Such a trade occurs when an incoming buy order is larger than or equal to the best ask price or an incoming sell order is smaller than or equal to the best bid price.

The trades indicated by an “E” are not the only trades that occur on the Island ECN. Entries in the data set denoted by a “P” reveal that a hidden or private order was executed. The order book does not show these orders, which only appear when traded. If one of these orders is added and then cancelled without being traded, there would be no record of its existence. These orders are usually large orders that
people wish to transact without being visible to other users for fearing of having a large impact on other traders’ decisions and future prices. Generally, the hidden orders denoted with a “P” are of little importance to this research. Hidden orders represent a very small percentage of the orders and trades, and they reveal no information about the current order book. Hidden orders are essentially invisible until traded, and their arrival times are left-censored. Due to the inability to know when the hidden orders are placed and the omission of hidden orders from the order book, these hidden orders will be omitted from the models throughout this thesis. This decision has negligible impact on results, as the hidden orders constitute a small percentage of all orders and provide almost no information.

As opposed to the Island data that contains all of the necessary information about added, cancelled, and traded orders, the data from ArcaBook only contains information about added orders and removed orders. These removed orders include both trades and cancellations but fails to distinguish between them. Trades and cancellations have the same net result on the order book, so this lack of differentiation between trades and cancellations will not have a significant impact on most of the analyses in this thesis. If it becomes crucial to know the difference between trades and cancellations for ArcaBook, the proportion and distribution of traded orders can be approximated from the Island data, and some of the orders removed from ArcaBook can be designated as trades.

Microsoft (MSFT) and Cisco Systems (CSCO) are two examples of spread-minimal stocks whose spread patterns and activity levels correspond to the desired
levels. Subramanian et al. [SRSK06] explain that the PLAT competitions that will be described in detail later use only Microsoft stock to test the entrants’ strategies. Unless otherwise noted, Microsoft will be used throughout the rest of this thesis to ensure consistency not only within the models but also between the results yielded here and the results from strategies tested on PLAT.

3.1.1 Standardizing Price

One point of interest is where the orders being added to and removed from the order book are placed in relation to the inner prices. However, looking at the raw data for prices is misleading, due to the movement of the inner prices throughout the day. One way to transform the data into a more useful form is by subtracting the price of added orders minus the inner price at that time. It makes sense to subtract the price of buy orders minus the inner buy price and sell orders minus the inner sell price to obtain the standardized prices. If the inner prices are represented by the ordered pair (inner buy price, inner sell price) = \( (p_b, p_s) \), added buy orders with price \( p^* \) are standardized to equal \( p = p^* - p_b \) and added sell orders with price \( p \) are standardized to \( p = p^* - p_s \). The main issue with this standardization is the inability to differentiate between \( p=0 \) created by a buy order added at the inner buy and a sell order added at the inner sell.

A more effective way to standardize prices involves adding or subtracting a cent to retain the difference between buy and sell orders that get added at the inner price. By using \( p = p^* - p_b - 0.01 \) for buy orders and \( p = p^* - p_s + 0.01 \) for sell orders, positive values correspond to sell orders and negative values correspond
to buy orders. Under this standardization, there will be a few 0's resulting from times during which the spread is 2 cents instead of the usual 1 cent spread. In this situation, the added order that brings the spread back to 1 cent will be inside the previous inner price, giving this added order a standardized value of $p=0$. These orders make up a small percentage of the total added orders.

3.2 Market Statistics

3.2.1 Arrival Rates

The most basic element of the ECN is the order. An order is placed by a trader who wishes to buy or sell a stock, also specifying the associated price and number of shares. Due to the high speeds associated with electronic trading via the internet, orders arrive to the ECNs at a very rapid rate. The arrival rate of orders to the order book is similar to a point process, but some important distinctions must be made. For point processes, one key assumption is that there is never more than one arrival at any time. In other words, the interarrival times are always positive. However, the time stamp on the ECN data used here gives accuracy to the nearest millisecond, so there are some milliseconds in which multiple orders are added. Although these orders may have appeared to be added at the exact same moment, there is a difference in the arrival times that goes beyond the accuracy of the data. Two orders appearing to be added at the exact same millisecond were in fact added a fraction of a millisecond apart, with the data maintaining the order at which they were added.
For point processes, the index of dispersion, defined as the ratio of the variance of the number of arrivals to the mean of the number of arrivals for a fixed time-scale, determines how a process relates to a Poisson process.

\[ I = \frac{\mathbb{E}[X]}{\text{Var}[X]} \]

Poisson processes have equal mean and variance, so the index of dispersion equals 1. The Poisson process has interarrival times that are exponentially distributed, giving the Poisson process the well-known "memoryless" property [Olo05].

During regular trading hours, an average of 350.25 orders are added every minute on Island for just Microsoft. The corresponding average number of orders arriving every minute on ArcaBook is 207.77. The variances are 59192.84 for Island and 21043.40 for ArcaBook. The index of dispersion is around 169 for Island and around 101 for ArcaBook. These are much larger than 1 as the variances are much larger than the means. The correlation between number of orders added per minute on Island and the number of orders added per minute on ArcaBook is 0.95, indicating a strong correlation between activity levels across the ECNs.

The previous observations indicate that the arrivals of orders to the electronic markets do not follow a Poisson process. Therefore, the arrival process is not Poisson, and the distribution of orders is not homogeneous throughout the day. Figure 3.1 and Figure 3.2 show the rates of added orders per minute over the entire day for Island and ArcaBook, respectively. Clearly, the pre-market and after hours trading have much less activity than the regular trading hours for both Island and ArcaBook. Restricting to the regular 9:30 AM to 4 PM hours, there are still parts
of the day that are generally more active than others. It is common for the early and late hours (9:30-11 AM, 3-4 PM) to be the most active, while the middle hours are less active. Also, the arriving orders appear to enter the system in clusters. In general, incoming orders generally lead to more and more incoming orders, which contradicts the memoryless property of a Poisson process. Due to the clustering of orders, the variance is higher than the mean, leading to an index of dispersion much greater than 1.

Now that the arrival rate for added orders has been examined, the next step involves looking at the rate at which trades and cancellation occur. Of course, ArcaBook does not distinguish between traded and cancelled orders, so these will have to be grouped together. For Island trades, the average is 66.42 per minute with variance 4128.59, yielding an index of dispersion of 62.16. For cancellations on Island, the mean is 301.3, and the variance is 41820.96, resulting in an index of
Figure 3.2: Arrival rates of added orders on ArcaBook

dispersion of 138.8. Again, these results indicate a non-Poisson distributed process.
For ArcaBook, the mean rate of order removal is 207.32 per minute with variance
21111.66, resulting in an index of dispersion of 101.83. It is noteworthy that the
index of dispersion for ArcaBook additions is very close to the index of dispersion
for ArcaBook order removals.

Figure 3.3 and Figure 3.4 show the number of orders that are traded or deleted
from Island and ArcaBook. When combining trades and cancellations on Island
to make them equivalent to ArcaBook and considering only regular trading hours,
there is a correlation of 0.95 between the rate at which orders are removed on Island
and the rate at which orders are removed on ArcaBook. The correlation between
the rates at which orders are added to Island and the rates at which orders are
being removed from Island is 0.993. Likewise, for ArcaBook, this correlation is
0.991. Therefore, intervals where there is a high level of activity for orders being
added to the order book are the same intervals where there is a high level of activity for orders being removed from the order book. Figure 3.5 shows the plot of the average amount of orders added per minute when splitting the day into 30-minute intervals. There is a visible convex (decrease then increase) shape, as well as an obvious correlation between Island (red) and ArcaBook (blue).

Figure 3.3: Rates of deleted/traded orders on Island

3.2.2 Considering Trades as Adds

At this point, it is important to realize that orders that get added to the order book are not the only orders added to the ECNs. Whenever a trade occurs, the traded shares are removed from the order book. However, this trade occurred as a result of an order added on the opposite side that got automatically matched with the existing order. Thus, there is a simultaneous add order that is never added to the order book and a traded order that is removed from the order book. If the volume
Figure 3.4: Rates of deleted/traded orders on ArcaBook

Figure 3.5: Averages numbers of arrivals in 30-minutes intervals
of the added order is greater than the volume of the existing orders with which it is to be matched, the orders at the inner price are traded and the remaining shares get added to the order book and become the new inner price. For example, if the inner prices are \((p, p+1)\) with volumes \((v_1, v_2)\), an added order to buy at price \(p+1\) for \(x\) shares \((x < v_2)\) would never appear in the order book. The result of the order would only show as a trade for \(x\) shares, and the new inner volumes would be \((v_1, v_2 - x)\). However, if \(x > v_2\), the new inner prices would be \((p+1, p+2)\) and the new inner volumes would be \((x - v_2, v_3)\), where \(v_3\) is the number of shares at price \(p+2\) that are waiting to be sold.

Around 14.6% of all orders that get added to the order book are eventually traded (either partially or fully). However, when considering orders that are immediately traded before they enter the order book, those constitute only 5.75% of all added orders. These orders make up a small percentage of the added orders, and when they are included, the index of dispersion values reported earlier do not show a significant change. They already factor into the rates at which trades occur, due to the automatic matching and removal of the order already in the order book.

### 3.2.3 Market Development

The only day for which data is available for both of the ECNs Island and ArcaBook is May 2, 2005. One concern with using data from May 2, 2005 is that the market may not have the same characteristics today as it did back then. The change in market characteristics as a result of increased activity over time is called “market development.” To measure market development, the relative activity levels on the
When looking at the number of orders added to the ECNs, there is a 10.8% increase between May 2nd, 2005 and May 23rd, 2007. This increase is due to the increased availability of ECNs to the general public. The densities and boxplots for the distributions of the amounts of adds per minute can be seen in Figure 3.6 and Figure 3.7. Let $\mu_1$ be the mean of the number of orders added every minute on May 2nd, 2005, and define $\mu_2$ accordingly for May 23rd, 2007. The p-value for testing whether these two distributions having equal means ($H_0 : \mu_1 = \mu_2, H_1 : \mu_1 < \mu_2$) is 0.12, so there is not sufficient information to conclude that the average arrival rates for May 23, 2007 are greater than those for May 2nd, 2005.

![Density of added orders per minute](image)

**Figure 3.6 : Measuring Market Development**

When comparing May 2, 2005 and May 23, 2007, the change in the rate of arrival rates was not found to be significant. Thus, the May 2, 2005 data can be
used without worry of significant difference in activity levels between the older and newer days of data. No conclusions can be made about whether the order book dynamics or models would be the same for the old and new data, merely that the market development is not significant.

3.2.4 Distribution of Added Orders and Prices on the ECNs

Now that the rate at which orders are added to the order book has been fully explored, it is necessary to look at how the prices of those orders are distributed. The added orders for Island and ArcaBook are centered around the same price, but the ArcaBook orders are spread out more than Island. To examine this closer, the means, medians, and standard deviations are calculated for each of Island and ArcaBook for each one-minute interval between 9:30 AM and 4:00 PM (the regular trading hours). Figure 3.8 shows the means (top), medians (middle), and standard
deviations (bottom) for Island (red) and ArcaBook (blue). Clearly, the means and medians follow very closely, with the means and medians having correlations of 0.959 and 0.981, respectively. The ArcaBook standard deviations are larger and more irregular than the Island standard deviations, and these have correlation of only 0.324. Clearly, orders on Island and ArcaBook have some important similarities yet also some stark differences.

![Graphs showing means, medians, and standard deviations of added orders on Island and ArcaBook.]

Figure 3.8: Means, Medians, and Standard Deviations of Added Orders on Island and ArcaBook

### 3.2.5 Correlated Stocks/ECNs

Another factor when examining changes in stock price is the overall movement of the market. If a stock is doing well, it may be a result of the market doing well as a whole. Therefore, stock prices should rise and fall in a correlated manner. The two spread-minimal stocks that have been presented so far are MSFT and CSCO. There is a 0.80 correlation between the inner buy price for Microsoft and the inner buy
price for Cisco. As expected, the correlation between inner sell price for Microsoft and Cisco is also 0.80. Figure 3.9 shows the distribution of inner prices for Microsoft and Cisco. Some parts of the day show obvious movement in the same direction, while others show one stock moving and the other one not moving.

![Inner Prices for MSFT](image)

![Inner Prices for CSCO](image)

Figure 3.9: Inner buy and sell prices and MSFT and CSCO

When considering the means, median, and standard deviations for MSFT and CSCO for each minute between 9:30 AM and 4:00 PM, there is a correlation of 0.789 between means, a correlation of 0.793 between the medians, and a correlation of 0.364 for the standard deviations between the two stocks. The means are less robust to outliers than the medians, so the means may have been influenced by very large or very small prices of added orders. The medians have a pretty high correlation, which, along with the relatively high correlation of inner prices, indicate that there is a positive correlation between MSFT and CSCO prices.

It is not surprising that the prices for MSFT and CSCO are positively correlated.
When the market is doing well, it is expected that stocks will generally do well as a whole and when the market is doing poorly, it is expected that stocks will generally do poorly as a whole. However, these stocks may be more correlated than most pairs. Both companies are technology companies, and many Microsoft products use Cisco systems and vice versa, so it is expected that as one goes, so goes the other. Also, because both are spread-minimal, the patterns of movement should be generally similar. The minimal spreads put a limit the number of ways the inner prices can move.

3.3 Distribution of Orders

Several papers mentioned earlier comment on motivations of traders and the different types of traders. The aggressive traders add orders close to the inner prices for quick execution and quick profit. More cautious traders place orders away from the inner prices with the hope of profiting from long-term price movement. This section examines the distribution of the orders within the order book.

At any time, the order book contains all of the information that is available on order-driven markets like ECNs. Knowing the distribution of those orders is crucial to gaining an understanding of the market. The orders are arranged in relation to the inner prices, which are at the center (all other buy orders are below, and all other sell orders are above). However, the inner prices move throughout the day, so in order to gain a more clear picture of how orders in the order book are distributed, it is necessary to standardize the prices by subtracting the appropriate inner price. However, one issue that arises from directly subtracting the corresponding inner
price is that all orders added at the current inner price get assigned the value 0 whether they are on the buy side or sell side. Therefore, adding $0.01 to all sell orders and subtracting $0.01 from all buy orders resolves this problem and distinguishes between buy and sell orders. If the inner buy is $25.00 and the inner sell is $25.01, an added buy order of $24.97 would be assigned the value -0.04, while an added sell order of $25.03 would be assigned a value of 0.03. Buy orders added at $25.00 would be assigned -0.01, and sell orders at $25.01 would be assigned 0.01. Before, both of these would be assigned value 0, making them indistinguishable from each other.

By applying this standardization, the distribution of all added orders for the entire day is shown in Figure 3.10. There are sharp peaks at $0.01 and -$0.01, which correspond to the inner buy and inner sell prices. There are also secondary peaks centered around $0.05 and -$0.05, which result from orders added four cents away from the inner prices. Previous papers have attempted to categorize traders by their strategies and goals. Based on this distribution, this dichotomy of traders appears to hold true. Aggressive traders place orders at the inner price (primary peak) with the hopes of achieving a quick trade, and cautious traders place orders a few cents away (secondary peak) from the inner price and hope the price moves towards them.

3.3.1 Trades and Cancellations

Thus far, much of the attention has been paid to where orders are added. However, it is important to remember that is only half of the story of an order. Every order
Figure 3.10: Distribution of orders in the Island and ArcaBook order books

that is added to the order book is eventually traded and/or cancelled. In order
to have any chance at gaining profit, it is not enough to add and cancel orders
but trading is necessary. Figure 3.11 shows the distribution of standardized prices
(at the time when the order is added to the order book) of those orders that are
eventually traded. There is an obvious majority at the inner prices, with more
than 89% of the orders that are eventually traded in those peaks. 98.9% of the
eventually traded orders are within 5 cents of the inner price, and more than 99.8%
of the eventually traded orders are within 10 cents of the inner price. The order
added the farthest from the inner price that was eventually traded was 27 cents
away from the inner price. These orders were added when the inner price was 27
cents away, but those people were patient enough to wait until the price moved and
reached that price.

Although by far the most orders that eventually get traded are placed at the
Distribution of standardized prices for orders that are eventually traded

Figure 3.11: Distribution of standardized prices for orders that are eventually traded

inner price, that location is also where most orders are placed in general. Looking at percentages of added orders that get traded for each standardized price may provide a better perspective, as seen in Figure 3.12. As expected, there is an obvious pattern of symmetry and monotonicity as the standardized prices move away from the inner prices. At the inner prices, 23.8% of orders added on the buy side and 23.7% of orders added on the sell side are eventually traded. These percentages drop rapidly as the standardized prices move away from inner prices, dropping down to around 1% for orders added 7 cents from the inner price.

3.3.2 Order Book Asymmetry

One obvious characteristic of the order book that requires further investigation is the symmetry between the buy and sell sides of the order book. The buy and sell order books exhibit similar shapes and have peaks at the same distances from the
Figure 3.12: Percentage of orders that are eventually traded.

inner prices. This symmetry occurs for all of the add orders between 9:30 AM and 4:00, of which there are over 136,000 of them. In order to get a closer look at the symmetry, it is necessary to focus on a smaller window of time. The first time window is a twenty-minute segment of time where the price monotonely increases 11 cents from $25.09 to $25.20. The distribution of standardized prices is shown in Figure 3.13. For this segment of time, the symmetry on the order books has been replaced by an obvious asymmetry. For added orders, a much larger peak appears at -$0.01 than at $0.01, and a much larger peak appears at $0.05 than at -$0.05. Conversely, the traded and deleted orders, a much larger peak appears at $0.01 than at -$0.01, and a larger peak appears at -$0.05 than at $0.05.

Next, a 4.5 minute window of time where the price drops 7 cents from $25.17 to $25.10 is investigated for possible asymmetry. As with the interval where the price increases, the distribution of added orders during the interval where the price
Figure 3.13: Distribution of added and traded/deleted orders during an interval of increasing price decreases shows asymmetry, as seen in graph Figure 3.14. Now, for added orders, the $0.01 peak is much larger than the -$0.01 peak, and the -$0.05 peak is much larger than the $0.05 peak. Likewise, for traded and deleted orders, the -$0.01 peak is much larger than the $0.01 peak, and the $0.05 peak is much larger than the -$0.05 peak. Therefore, the asymmetry reveals information about the price movement. For added orders, a larger peak on the sell (positive standardized price) side at the primary peak and a larger peak on the buy (negative standardized price) at the secondary peak indicate the price is decreasing. On the other hand, a larger peak on the buy side at the primary peak and a larger peak on the sell at the secondary peak indicate the price is increasing. The opposites hold when considering traded and deleted orders.
47

Figure 3.14: Distribution of added and traded/deleted orders during an interval of decreasing price

3.4 Distribution of Shares

When considering orders being added and removed from the order book, it is important to pay attention to the volume of shares in those orders, particularly when looking at orders that are partially traded or deleted. For example, a user who adds an order of 300 shares may delete that order 50 shares at a time. In this case, a net difference of added shares minus deleted shares would be 0. However, the net difference of added orders minus deleted orders is -5, which is less accurate than the difference of shares.

First, it is necessary to look at the distribution of the volumes for orders added to the order book, which is shown for Island and ArcaBook in Figure 3.15. There are some obvious peaks corresponding to popular shares values such as 100, 500, and 1000. The number of shares associated with an order is determined by the user, so these values tend to be round numbers. In fact, only 1.9% of orders added
on Island and only 2.8% of orders added on ArcaBook have a volume of shares that is not a multiple of 100.

Figure 3.15: Distribution of shares for added orders

3.4.1 Correlation/independence of price/volume

The correlation between an order's price and an order's volume is 0.005 for Island and 0.001 for ArcaBook. When considering buy and sell orders separately, the correlations are -0.016 for Island buy orders, 0.041 for Island sell orders, 0.054 for ArcaBook buy orders, and -0.029 for ArcaBook sell orders. There is extremely low correlation between an order's price and an order's volume, but the next step is to determine where there is independence.

To test independence between and order’s price and an order’s volume of shares, the orders are split into categories by their values. First, orders are divided into groups based on whether they are placed at the inner price ($|p| = 0.01$) or away from
the inner price \(|p| > 0.01\). Next, the orders are divided by whether their volume is less than 500 shares or whether their volume is greater than 500 shares. For Island, there are 31,812 orders added at the inner price with volume at most 500 shares, 17,584 orders added at the inner price with volume greater than 500 shares, 25,962 orders added away from the inner price with volume at most 500 shares, and 20,678 orders added away from the inner price with volume greater than 500 shares. The \(\chi^2\) value of 1.72 on 1 degree of freedom yields a p-value of 0.1897. These numbers indicate that an order with a smaller number of shares is more likely to be added at the inner price, and an order for a larger number of shares is more likely to be added away from the inner price. Therefore, independence does not hold between an order's price and its volume of shares. The results from ArcaBook agree with those for Island.

3.5 Lifetimes of Orders

The previous chapter discussed several papers that have been written regarding the lifetimes of orders, or the time duration between when the order is added and when it is removed for the last time due to trade or cancellation. The distinction of "for the last time" is made because orders can be partially traded or cancelled. If an order is added for 100 shares at 11 AM and later has 50 shares cancelled at 11:30 and 50 shares cancelled at 12:00, then the lifetime would be an hour.

The distribution for lifetimes of orders on Island and ArcaBook are shown in graphs Figure 3.16. Island and ArcaBook have similar distributions of lifetimes, with Island having a second peak that ArcaBook does not have. The orders are
short-lived, with 15.4% of orders added on Island having lifetime less than 1 second, 27% having lifetime less than 5 seconds, 69.2% having lifetime less than a minute, and over 90% having lifetime less than 5 minutes. For ArcaBook, there are 16.9% of orders with lifetime less than 1 second, 30.7% having lifetime less than 5 seconds, 67.1% having lifetime less than 1 minute, and 88.5% having lifetime less than 5 minutes. These results reveal the order book as "young" where 90% of the orders that are currently in the order book will no longer be there just five minutes from now.

![Figure 3.16 : Distribution of lifetimes for Island and ArcaBook orders](image)

Each order that gets added to the order book is eventually fully traded, fully cancelled, or a combination of traded and cancelled. The distribution of lifetimes of orders can be split into orders that are traded, those that are cancelled, and those that are partially traded and partially cancelled. This split can only be done for the Island data, since the ArcaBook data does not distinguish between trades
and cancellations. The distributions of lifetimes for orders that are only cancelled (left), only traded (center), and both traded and cancelled (right) are shown in graph Figure 3.17. The graphs show no visible difference in the distributions of lifetimes, so an order’s eventual method of removal from the order book does not affect how long the order spends in the order book.

![Figure 3.17: Distribution of log-lifetimes for traded and cancelled orders](image)

As was done before with price and volume, it is important to know if there is correlation or dependence between price and lifetime. The correlation between standardized price and lifetime is -0.00065 for Island orders and is 0.0017 for Arca-Book orders. Although these correlations are extremely low, uncorrelated does not imply independent. To test independence, the orders are split by whether they are placed at the inner price or away from the inner price and by whether their lifetime in the order book is smaller than or greater than 30 seconds. There are 30,564 orders placed at the inner price with lifetime shorter than 30 seconds, 19,290 orders
placed at the inner price with lifetime longer than 30 seconds, 18,481 orders placed away from the inner price with lifetime shorter than 30 seconds, and 26,793 orders placed away from the inner price with lifetime longer than 30 seconds. A $\chi^2$ value of 0.9482 on 1 degree of freedom yields a p-value of 0.33. Therefore, orders placed at the inner price are more likely to have shorter lifetimes than those placed away from the inner price. This result makes intuitive sense, as those orders placed at the inner price were likely placed there by aggressive traders for quick transaction.

Hasbrouck and Saar [HS08] tagged orders with lifetime less than 2 seconds as “fleeting” and removed them from their study. These orders constitute 19.3% of the added orders on Island and 21.9% of the added orders on Island. The distributions for the standardized prices of these orders are shown in Figure 3.18 and Figure 3.19. The distributions for these selected orders appear almost identical to the distribution of the standardized prices for all orders. Thus, the only impact of removing these orders would be in reducing the density of information available for accurate modeling. These “fleeting” orders do not have any anomalous characteristics other than their short life spans, and removing them would be an unnecessary loss of data.

3.6 Comparing Island and ArcaBook

Arbitrage opportunities arise when a stock is available to be bought at a certain price on one exchange and is available to be sold at a higher price on another exchange. This situation creates a no-risk profit for whoever is the first to act upon it. These types of events will be looked at closely later, but first it is important to
Figure 3.18: Distribution of standardized prices for fleeting orders and all orders on Island

Figure 3.19: Distribution of standardized prices for fleeting orders and all orders on ArcaBook
examine the behavior of inner prices for Island and ArcaBook. Figure 3.20 shows
the movement of inner buy and inner sell prices for both Island and ArcaBook
throughout the day. There are four time series: Island inner buy in red, Island inner
sell in blue, ArcaBook inner buy in green, and ArcaBook inner sell in purple. They
all move in relative unison. The Island and ArcaBook series are almost concurrent,
but Figure 3.21 shows a close-up on a smaller time window where differences are
more visible. Whenever one series moves, it either reverts back quickly or the
others soon follow. As a result, the correlation between Island inner buy price and
ArcaBook inner buy price is 0.9978. Likewise, the correlation between Island inner
sell price and ArcaBook inner sell price is 0.9981. As expected, these are extremely
high correlations, indicating that traders on Island and ArcaBook follow similar
behaviors.

Figure 3.20: Inner buy and sell prices for Island and ArcaBook

It is important to know which of the series tends to move first and thus “drives”
Due to the spread being one-cent under normal circumstances, a price increase will always occur first on the sell side, and a price increase will always occur first on the buy side. For price increases, Island moves before ArcaBook 151 times, while ArcaBook moves before Island only 57 times. For price decreases, Island moves before ArcaBook 157 times, while ArcaBook moves before Island only 56 times. Island moves first 73% of the time, so it appears that Island is more instrumental than ArcaBook in driving prices.

### 3.6.1 Arbitrage Opportunities

Whenever there is a new technological development, there will be people trying to take advantage of the situation. With ECNs, there will be people looking to make a profit by either legal or illegal means. The term arbitrage is applied to any situation where someone attempts to gain from differences in prices across markets.
For Island and ArcaBook, if the Island inner sell price is lower than the ArcaBook inner buy price, or if the ArcaBook inner sell price is lower than the Island inner buy price, then a person could buy the shares at the lower inner sell price and then sell them at the higher inner buy price, making an instant profit.

Using the data for both Island and ArcaBook, there were two instances when the Island inner sell price was lower than the ArcaBook inner buy price, and there were six instances when the ArcaBook inner sell price was lower than the Island inner buy price. For the two intervals of time where Island inner sell price was lower than ArcaBook inner buy price, each time the difference was just one cent. Both of the occurrences lasted less than a second and occurred shortly after 9:30 AM. These times are not surprising, as the influx of traders and orders into the market when the regular market hours begin at 9:30 might cause momentary irregularities. However, instances these do not last long. There were 6400 shares available to be traded at one of those times, so a savvy individual (or more likely, a well-programmed computer) could detect the difference across ECNs and make a profit of $64.

Of the six instances when the ArcaBook inner sell price was lower than the Island inner buy price, four of the six happened shortly after 4:00 PM, and two were shortly before 9:30 AM. The two that occurred before the regular trading hours lasted 309 and 87 seconds. As for the four that occurred after the regular trading hours, one lasted about twenty seconds, while the other three lasted less than one second. For these six, the number of shares to be traded was never larger
than 100, and all but one of the six had a price difference of one cent (the other was
two cents). Therefore, these occurrences do not provide as much potential profit as
the ones before.

There are eight instances where a quick profit can be made. These opportunities
occur early or late in the day, when there is less activity in the market and greater
chance for momentary irregularities. However, with the exception of one case, there
is not potential for much profit. Although the profit generated by these arbitrage
opportunities may be small, there is almost no risk involved if executed properly.
The only potential risk would result from a change in prices between the time when
the order to buy or sell is input and the time it reaches the ECN for execution. In
an ideal system, the executions would be immediate, resulting in risk-free profits.
Small yet consistent positive profits would produce a small standard deviation and
a high Sharpe ratio. The Sortino ratio would be infinite as a result of the lack of
negative returns.

Figure 3.22 shows one such arbitrage opportunity which occurs shortly before
9:30 AM. The Island inner buy price, shown in red, is greater than the ArcaBook
inner sell price. At that time, a trader with access to both ECNs could buy on
ArcaBook and immediately sell on Island at a price that is one cent higher. After
several seconds at those prices, the Island inner buy decreases by one cent and
becomes equal to the ArcaBook inner sell, ending the opportunity for arbitrage.
When the Island inner buy and ArcaBook inner sell are equal, a trader can buy
on ArcaBook and immediately sell on Island at the same price, resulting in no net
profit or loss.

Figure 3.22: Example of an arbitrage opportunity on Island and ArcaBook

More arbitrage opportunities can be found or created by legal or illegal means. An illegal way to create an arbitrage opportunity is by hacking into the computer system. By gaining information faster than other users, a trader could create many new opportunities for buying and selling (or short-selling then buying) to gain a quick profit. An example of legal arbitrage is successfully formulating a predictive model to give oneself advanced information about an impending price change. The next section creates such models, which will then be used for creating new profit-making strategies.
Chapter 4
Models and Results

Previous chapters drew conclusions about the shape and dynamics of an ECN’s order book. This chapter presents ways those conclusions can be used to formulate models for predicting price changes. The models developed will feature a variety of time series techniques, including autoregression and logistic regression, as well as modifications and variations on existing techniques. These models will then be used in the next chapter to develop strategies and algorithms for gaining profit by trading on ECNs.

Most of the preliminary analyses have used Microsoft stock data to draw conclusions about the order book, and most of the models developed in this chapter will use Microsoft data to formulate and test the models. However, other spread-minimal stocks, including Cisco and Yahoo, will be used for confirmation of the techniques and models developed here.

4.1 Price Changes

Before developing models for predicting price changes on ECNs, it is important to understand the different ways stock prices on ECNs can change. First, consider a stock whose bid-ask spread is n cents. If \( n > 1 \), a trader can add a new buy or sell order between the previous inner buy and inner sell prices, creating a change in one of the inner prices. However, stocks that are spread-minimal will have \( n = 1 \),
and the price cannot change simply by adding an order. The only way the price
will change is if all buy orders at the inner buy price or all of the sell orders at the
inner sell price are removed due to trading or cancellation.

Consider what happens when all of the shares at the inner buy are removed by
either trade or cancellation for a spread-minimal stock. All of the buy orders at the
inner price are either cancelled or matched with incoming sell orders at a price less
than or equal to the inner buy price. As soon as the inner buy shares empty, the
inner buy price has dropped by one cent. If a sell order is added at the previous
inner buy price, then the sell price has dropped by one cent as well, and a price
decrease has occurred. If the inner buy shares had been emptied by a trade with an
incoming sell order for more shares than were available at the inner buy price, the
remaining shares in the sell order would be added to the order book and become
the new inner sell price, completing the price change immediately. An analogous
case would occur for a price increase if all of the shares at the inner sell are emptied
out.

For example, if the inner prices are ($25.00, $25.01), with (300,500) shares, then
an incoming sell order for $25.00 with volume 400 shares will result in a trade for
300 shares at $25.00, and then the extra 100 shares will become the new inner sell.
Thus, the new inner prices will be ($24.99, $25.00) with volumes (c, 100), where c
is the number of buy shares at $24.99. The data represents the added 400 shares
as a 300 share trade on the buy side and a 100 share add on the sell side.

If the price were just an individual number, it would be simple to define a price
change. However, the prices are really ordered pairs, and it is possible for one of the elements in the price to change while the other one does not. It then becomes necessary to establish qualifications for which price changes to include.

The first type of price change for spread-minimal stocks occurs when either the inner buy or sell price changes, followed shortly by a price change in the same direction for the other side. These price changes will be called "true" price changes and are indicative of an actual change in the value of the stock. However, it is possible that after one side changes its inner price, that same side changes right back before the other side can change. In this situation, the ultimate result of the event is that the prices are exactly where they started. Therefore, these types of changes are not "true" price changes but instead are merely price "fluctuations."

Figure 4.1 shows an example of a true price change and a price fluctuation.

![True Price Change and Price Fluctuation](image)

Figure 4.1: Example of a true price change and a price fluctuation

The choice of which types of changes to include depends on which create op-
portunities to profit. In order to profit by a price increase or decrease, the user must purchase it at a lower price and sell at the higher price or short-sell at a higher price and then buy at a lower price. If someone buys low and sells higher, he or she must buy at the inner sell and will need the inner buy to increase. For a price increase, the inner sell increases first and then the inner buy moves up to return the spread to one cent. Likewise, if someone wants to short-sell at a higher price then buy it back at a lower price, the price decrease occurs at the inner buy first then the inner sell. Either way, it is the second price change that is the one necessary for profit. Therefore, it is the change that occurs on both sides in the same direction that is important to the traders. These changes are the true price changes, and generally these will be the only changes that will be considered for the models. Price fluctuations do not contribute to profitability and will therefore be omitted from the models unless otherwise indicated.

For Island between the hours of 10:00 AM and 3:00 PM, there are 466 true price changes where both the inner buy and inner sell prices move in the same direction by one cent. There are 100 price fluctuations on the buy side and 1070 price fluctuations on the sell side. At the same time on ArcaBook, there are 510 true price changes, 184 price fluctuations on the buy side, and 194 price fluctuations on the sell side. Although the inner prices for Island and ArcaBook move in relative unison, there are 44 more true price changes on ArcaBook than Island. This difference results from the times when the inner buy and inner sell prices on one ECN move in one direction and then back to their original values while the inner prices on the
other ECN do not move. This pair of changes does count as a pair of true price change, because both inner prices on a single ECN did change. Of the 466 true price changes on Island and the 510 true price changes on ArcaBook, 421 of them occurred in the same direction and in close time proximity to each other.

Of these 421 true price changes that are observed on both Island and ArcaBook, 208 were price increases, while 213 were price decreases. Of the 208 price increases, 151 occurred on Island before they occurred on ArcaBook while only 57 occurred on ArcaBook before they occurred on Island. Likewise, for the 213 price decreases, 157 occurred first on Island, and only 56 occurred first on ArcaBook. For the true price changes, they occur on Island before they occur on ArcaBook 73% of the time, indicating that information may reach Island before ArcaBook or that the traders on Island are driving prices more than on ArcaBook (perhaps due to the previously observed higher level of activity on Island than on ArcaBook). When the Island inner prices move before the ArcaBook inner prices, the time difference between these events has mean 0.451 seconds and median 0.068 seconds. When ArcaBook inner prices move first, the time difference has mean 2.083 seconds and median 0.341 seconds. Therefore, although the Island inner prices move first more frequently than ArcaBook, the ArcaBook inner prices move first by a greater time margin than when Island prices move first.

4.2 Epochs

Now that the qualifications for which price changes to consider have been established, it is necessary to discuss what happens during and between price changes.
An “epoch” is defined as the period of time between consecutive price changes. However, the inner prices are not necessarily constant during an epoch, as there may be fluctuations in the inner buy and inner sell prices. For a true price change, the inner buy and inner sell prices move in the same direction. An epoch begins at the second change in the pair and ends at the first change in the next pair. These changes occur within what is generally a small window of time, but they are not at the exact same time. Therefore, an epoch will begin at the second of the price changes that occur and end at the first change in the next true price change. The inner prices will be constant during an epoch except during fluctuations.

Price changes for spread-minimal stocks will only occur when the shares at either the inner buy or inner sell are emptied out due to trades or cancellations. The quantity of shares at the inner buy and inner sell at any given time can be thought of as a bivariate time series. Whenever one of the processes hits 0, one of the aforementioned price changes occurs. The quantity of shares at one of the inner prices hitting 0 causes a sudden jump to the number of shares at the new inner price. The series of the amount of shares at the inner prices go through a number of changes throughout the day whenever a trade occurs or when a user adds or deletes an order at the inner price. However, these are generally smaller than the sudden movements that occur when the inner price changes that result in the inner shares jumping from 0 up to another amount.
4.2.1 Changes in Inner Shares

The quantities of shares at the inner buy and inner sell are critical to understanding price changes. Whenever an order is added at the inner price, deleted from the inner price, or added at a price such that it creates a trade at the inner price, the result is a change in the quantity of shares at either the inner buy or inner sell. When the spread is one cent, the only way the price can change is for the shares at the inner buy or inner sell to empty out. Therefore, prediction of the movement of the inner buy shares and inner sell shares is crucial to the prediction of which direction the price is going to move next.

There are two special processes that need to be defined. The first process is the dual process of the number of shares at the inner buy and inner sell. Whenever one of these hits 0, there will be a sudden jump (called a “reward”) to the number at shares at the new inner price. However, true price changes occur on both sides of the order book. Thus, the other inner price will soon change as well, resulting in a jump in the inner shares on the opposite side.

The second process of interest can be thought of as a difference of differences of shares movement. Define

\[ v(p, t) = \text{volume of shares in the order book at standardized price } p \text{ at time } t. \]

Likewise, define

\[ d_t(p, t_1, t_2) = v(p, t_2) - v(p, t_1) - \lambda(p, t_1, t_2) \]

as the change in shares on the Island order book between times \( t_1 \) and \( t_2 \) at standardized price \( p \) cents, with the \( \lambda(p, t_1, t_2) \) term equaling the sum of all rewards.
that occur between times \( t_1 \) and \( t_2 \). This change is the number of shares added at standardized price \( p \) minus the number of shares traded or deleted at standardized price \( p \). The variable includes no provision for which whether the shares change occurs on the buy or sell side, because the sign of the standardized price reveals whether it is a buy \((p < 0)\) or sell \((p > 0)\). Likewise, define \( d_A(p, t_1, t_2) \) for changes on ArcaBook.

Now, let

\[
x_I(p, t_1, t_2) = d_I(p, t_1, t_2) - d_I(-p, t_1, t_2).
\]

For \( p=1 \),

\[
x(1, t_1, t_2) = d_I(1, t_1, t_2) - d_I(-1, t_1, t_2)
\]
equals the difference of the increments at the inner buy and sell prices. Essentially, this process represents the information contained in the dual process on inner prices as a single process with the effect of the rewards removed.

When looking at the size of the rewards, there are two jumps to consider. First, when one side empties, the shares jump up from 0 to however many shares are in the order book at the new inner price. However, the spread is now 2 cents, and a new order will quickly fill in the vacated slot. This new order causes the quantity of shares at the other inner prices to jump as well. However, this time the shares drop from the previous quantity to however many are added at the new inner price.

For Island, the first jumps have mean 35208 and median 33175, and the second jumps have mean -40156 and median -39339. The correlation between the size of the first and second jumps is -0.145. For ArcaBook, the first jumps have mean
21537 and median 22200, and the second jumps have mean -22988 and median -24227. The correlation between ArcaBook jumps is -0.66. The smaller jumps in shares on ArcaBook than on Island result from a generally lower level of activity on ArcaBook than on Island.

Why are the jumps on one side positive while the jumps on the other side are negative? The first jump occurs when the remaining shares at one of the inner prices are removed, and a new value becomes the inner price. Generally, that quantity of remaining shares was small for it to have been removed with a single order. The new value of shares is generally larger than the previous one, resulting in the large positive values observed for the first jump. The second jump moves from the shares at the inner price to the number of shares in the first order added to bring the spread back to 1 cent. Therefore, this jump is moving from a large amount of shares to the amount in one single order (generally a much smaller amount), resulting in the negative values. The boxplots of the first jump and the negation of the second jump can be seen in Figure 4.2 and Figure 4.3. The distributions are very similar, leading us to possibly hypothesize that the first and second jumps may be very highly negatively correlated. However, the correlation is only -0.11 for Island and -0.66 for ArcaBook. These correlations indicate that while the jumps are similarly distributed, the paired values of the first and second jumps are not necessarily strongly correlated.

It is important to realize that there are actually two rewards that occur at a true price change as a result of changes in both the inner buy and inner sell values.
Figure 4.2: Distribution of jumps in shares at Island price changes

Figure 4.3: Distribution of jumps in shares at ArcaBook price changes
The first change results when the inner shares at either the buy or sell side empties out and becomes 0. If there are 0 shares in the order book at that price, then that price is no longer an inner price, causing that immediate reward. The second change occurs when some new order (or the remaining shares in the traded order that resulted in the emptying out of the shares to cause the initial change) becomes the new inner price on the other side.

In addition to the size of these jumps, the time duration between consecutive changes is important. A careful distinction must be made when dealing with time: the time “within changes” is the time between when one inner price changes and the other side changes, while the time “between changes” is the time between consecutive price changes. The time between consecutive price changes is equivalent to the length of an epoch.

The period of time between two consecutive price changes is called an epoch. On Island, these epochs have mean duration 38.47 seconds and have median 18.045 seconds. The maximum duration of an epoch is 469 seconds (7.8 minutes), and the minimum duration is less than a millisecond. For ArcaBook, the epochs have mean duration 35.01 seconds and have median duration 15.89 seconds. The maximum duration for an epoch on ArcaBook is 463 seconds (7.7 minutes), while the minimum is again less than a millisecond. The epochs have slightly shorter time durations on ArcaBook than on Island due to the larger number of true price changes on ArcaBook. ArcaBook had 510 true price changes, as opposed to 466 for Island. More price changes create more epochs, which divide the day into smaller windows
of time.

The time within the price change on one side of the order book and the other has mean 0.109 seconds and median 0.003 seconds for Island inner prices. The maximum time within changes is 6.65 seconds, while the minimum is less than a millisecond. On ArcaBook, the mean is 0.234 seconds, and the median is 0.030 seconds. The maximum time within the changes is 14.38 seconds, and the minimum is less than a millisecond. The time within changes is longer for ArcaBook than for Island. The distribution of the log of these durations can be seen in graph Figure 4.4. These short durations agree with the expectations for actively-traded spread-minimal stocks.

![Density of log-durations within price changes](image)

Figure 4.4: Distribution of (log) durations within price changes

Next, consider the changes in the inner buy and inner sell shares. Whenever an order is added at the inner price, the change will be positive. When an order is traded or deleted at the inner price, the change will be negative. Figure 4.5 shows
the distribution of differences of only inner buy shares and only inner sell shares separately. There is no obvious difference in the distributions of buy and sell, so they can be combined into a single distribution in graph Figure 4.6. A little over 94% of the changes in the inner buy and inner sell shares are multiples of 100, so those that are multiples of 100 can be split apart from those that are not. The distribution of the multiples of 100 and the distribution of the rest of the changes is shown in Figure 4.7. For the multiples of 100, the peaks drop off as they move away from 100 shares, with 400 dropping below 500. There are also obvious peaks at 1000 and 2000. These peaks are expected, as the traders create these values when they place their orders. The differences that are not multiples of 100 show a more symmetric distribution centered at 0, although there are significantly fewer observations (6%) than for the multiples of 100 (94%).

Figure 4.5 : Distribution of changes in inner shares for buy and sell separately

Every time an order is added, traded, or deleted at an inner price, the shares at
Figure 4.6: Distribution of changes in inner shares combining buy and sell

Figure 4.7: Distribution of changes in inner shares (multiples of 100 or not)
the inner buy or inner sell will change. Whether there is a pattern to which side the change occurs or whether it is random will reveal the dynamics of the placement of orders. To determine whether it is random or not, the sequence of events is tested by looking at the runs of consecutive changes on the same side. The events can be split into two groups, buy side and sell side, or 4 groups: buy increase, buy decrease, sell increase, and sell decrease. Here, increase means a rise in quantity of shares as a result of an added order, and decrease means a drop in shares as a result of a trade or cancellation. Multiple changes that may occur from a single order must be taken into account. For example, if a large order is added such that it is immediately traded, it may be matched up with several smaller orders to fill up the entire order, which would show up as several consecutive small changes, when in reality it was one big order.

First, changes in the inner shares are classified by whether they occur on the buy or sell side. For now, whether it was an increase or decrease in shares is ignored. Consider the number of changes that occur on a single side of the order book between 2 consecutive changes on the opposite side. The minimum value is obviously 1, which occurs anytime a change occurs in the inner buy shares or the inner sell shares in between 2 changes on the opposite side. However, the maximum is 829, which occurs during a time of numerous changes to the inner sell shares but no activity to the inner buy shares. In fact, there are 5 streaks of more than 100 consecutive changes to one side’s inner shares without a single change to the other side’s inner shares. The median streak length is 2, and the mean is around 3.2.
Next, consider changes not only by which side of the order book but also direction. Now, an increase in the number of shares at the inner buy and a decrease in the number of shares at the inner buy would be considered different types of events. Now, the maximum length of a streak is only 59 in a row, which is still a sizable number for changes on one side and in the same direction. The mean streak length is 2.2, and the median streak length is 1, which means that more than half of the changes in the inner shares is followed by a change that differs either in direction (increase/decrease) or side of the order book (buy/sell.) Graph (17) shows the distribution of the log of the lengths of the streaks for both categorizations. The distributions appear similar, with the categorization into buy/sell (top graph) having larger tails than the categorization into buy/sell and shares increase/decrease (bottom graph.) This makes intuitive sense, since the periods of time where one side is undergoing all of the activity will tend to be broken up by both increases and decreases in shares.

4.3 Model Setup

Creating accurate price models requires several decisions to be made, presented in the following order:

- time scale
- variables (dependent and independent)
- diagnostic methods
As the ultimate goal of creating these models is the effective prediction of price changes, the dependent variable should be a function of price changes. Using logistic models, the probabilities of different price change outcomes (increases, decreases, no change) will be predicted. The independent variables will reflect the information from the order book that conveys insight into price changes. The time-scale will be chosen wisely to maximize the information contained in the model's variables. To both create and test the model, the data will be split into two (not necessarily equal) parts. The first part will be used to determine coefficients and parameters for the model, while the second part will test the output from the first part of the day to evaluate its accuracy and effectiveness of prediction.

First, the information in the order book that is important for price modeling and prediction must be determined. Obviously, using the entire order book would contain the maximum amount of information. The mean and median of orders added to and removed from the order book will be used in autoregressive models. However, orders at certain distances from the inner prices seem to “drive” the price movements. When looking at the distribution of standardized prices for orders in the order book, there are obvious peaks at one cent and five cents. Figure 4.8 shows the movement of shares at one cent (top), five cents (middle), and the inner prices (bottom). The shares at one-cent and five cents have correlation -0.93. Both sequences follow the inner price movement closely, with the one cent shares movement opposite to both the five cent shares and inner prices. Therefore, these distances appear to drive prices and will be featured in the predictive models.
4.3.1 Time Scale

As described by Engle and Russell [ER98], the selection of time-scale for a model is critical. Choosing a time-scale that is too small will result in many intervals with no activity, and the model will not be useful. However, choosing a time-scale that is too large will create intervals that may contain too much information, and individual events may be lost when combined with others within an interval.

To determine the optimal time-scale to use for the development of these price models, time-scales ranging from $\frac{1}{4}$ of a second to 30 seconds are tested for their ability to detect the correct direction of price changes. In the previous chapter, it was shown that during intervals where inner prices are increase, there is a greater quantity of shares added at 0.05 than at -0.05 and a greater quantity of shares added at -0.01 than at 0.01. Likewise, during intervals of price decreases, there is a greater quantity of shares added at -0.05 than at 0.05 and a greater quantity of...
shares added at 0.01 than at -0.01. To detect price changes for different time scales, an interval that exhibits the aforementioned qualities will be classified as such, and the percentage of price changes correctly identified is calculated.

The results of the percentage of price changes that were detected and their direction correctly determined are shown in Figure 4.9. It is important to remember that these price changes were not predicted ahead of time, but were detected by the order book movement. Once the optimal time scale is established, then predictive models can be created. In the graph, a time scale of one-second gives the best result at 88% of price changes detected. Smaller time scales do slightly worse at detecting price changes, as these large intervals combine price changes and lose information. One second time scale will be used for the models, as it is large enough to contain information about order book and price dynamics but is small enough not to lose individual events. A one-second time scale results in 18,000 intervals for a 5-hour period in which 466 true price changes were observed on Island and 510 observed on ArcaBook. Therefore, between 2.5% and 3% of the time intervals will exhibit a price change. When using other sets of data to create similar price models, the optimal time scale may vary, depending on the frequency of observations and the frequency of price changes or other events.

4.3.2 Variable Selection

Now that the time scale has been determined, the variables \( d_1(p, t_1, t_2) \) and \( x_1(p, t_1, t_2) \) no longer require two parameters for time. These variables can be parameterized
as

\[ d_I(p, t) = v(p, t) - v(p, t - 1), \]

the increment of shares at standardized price \( p \) between times \( t \) and \( t+1 \), and

\[ x_I(p, t) = d_I(p, t) - d_I(-p, t), \]

which now equals the difference in increments of standardized prices \( p \) and \(-p\) between times \( t \) and \( t+1 \).

Figure 4.9: Percentage of price changes detected when using different time scales

Now that the optimal time-scale has been established for a time-homogeneous autoregressive model, it becomes necessary to determine what variables should be included in the models. Before, the changes in shares at standardized prices with absolute value 0.01 and 0.05 were shown to closely follow the movement of the inner prices. However, it is important to determine the proper number of lags for an autoregressive model. Too few lags will result in an incomplete and less
successful model, but too many lags causes the model to be slow and less efficient. Speed is critical on ECNs, lots of data must be analyzed, so a faster model may be a more profitable model.

The first regression for determining significant lags will be an autoregression of the changes in shares at the inner buy and inner sell prices. Regressing changes in inner buy shares on previous lags of itself, the first three lags are significant at the 0.01 level. The same holds for regressing inner buy shares on inner sell shares, inner sell shares on inner buy shares, and inner sell shares on itself. When including both sets of independent variables in a single regression with inner buy shares as the dependent variable, the three lags of the inner sell shares are more significant (lower p-values) than the inner buy shares. Likewise, when regressing changes in inner sell shares on both inner buy shares and inner sell shares, the lags of inner buy shares have lower p-values than those of inner sell shares.

When regressing Island price differences on Island shares differences, the preceding three one-second time lags are significant (p-value less than 0.05). For ArcaBook, three lags are significant as well. Likewise, when regressing each of Island on ArcaBook changes on the lags of the opposite ECN, three lags are significant on each model. However, when combining the independent variables into a single model and regressing Island price changes on both Island and ArcaBook shares difference, the last 3 ArcaBook lags are more significant (lower p-values) than Island. Additionally, Island lags having lower p-values than ArcaBook lags when modeling ArcaBook price changes.
Three lags (three seconds) consistently appear as the set of significant variables for modeling price changes as a function of shares changes. In the previous chapter, the order book was noted as being “young” with a constant addition and removal of orders. The significance of only the previous three seconds agrees with the youthful quality of the order book and the information contained in it. Also, for both inner buy and inner sell shares and for Island and ArcaBook data, the lags on the opposite series are more significant than the ones for the series itself.

Other important quantities to include in the independent variables are shares in the order book at the inner buy and inner sell prices. Price changes only occur when either the shares at the inner buy or the shares at the inner sell get emptied out. Therefore, a smaller quantity of shares at an inner price indicates a better chance of these shares emptying and a price change ensuing, while a larger quantity of shares indicates a lower probability of a price changing.

Now that the variables to be used for the model have been established, but one problem persists that needs resolving before formulation of the model. If all of the available data is used for the model, then there is no way to test the effectiveness of the model for prediction. To allow for fitting and testing of the model, the day is split into two parts. The first part will be used to formulate and calculate coefficients. The parameters for the model will then be used to predict price changes for the second part of the day. The decision about when to split the day is an important one. If the time split is too early, then the coefficients were established using insufficient quantities of data. On the other hand, if the day is split too late,
then there is not enough time for sufficient testing of the model.

The best time to split the day is after the coefficients for the model are relatively constant. To examine this, formulate a model using increasingly larger portions of the day and look at the coefficients of the autoregressive model to see if they do indeed converge. Figure 4.10 shows the sequence of the 4 coefficients of an AR(3) model on shares movement at the inner price, utilizing an increasingly larger portion of the day starting from 10:00 AM. From the graphs, there is obvious fluctuation in coefficients for the first few minutes. However, after around 100 minutes, the coefficients are relatively stable and do not fluctuate. Therefore, after 120 minutes, at 12:00 noon, the day is split with the time before noon used to formulate the model and the time after noon used to test the model.

Figure 4.10: Sequences of coefficients over time
4.3.3 Evaluating Models

In the next chapter, profit strategies will be developed based on the models created in this chapter. In order to gain profit, it is not as important to predict the exact interval when a price change will occur as it to predict the direction of the next price change. Sometimes the order book conditions that cause the model to predict a large probability of a price increase or price decrease subside for a few seconds, and a price change does not occur until a few seconds later. In this situation, the interval flagged as being significant for a price change would be a false positive, and the interval that was not deemed significant but exhibited a price change would be a false negative. However, for the purposes of potentially gaining profit, the model was successful, if the predicted direction of the price change is the same as the direction of the observed price change. The word “potentially” is used here to reinforce that one price change is not enough to gain profit, but multiple price changes are needed, which will be important when creating profitable strategies.

To determine a model’s potential use in a profitable market strategy, a statistic is created that measures how effective the model predicts the direction of next price change. For an appropriate cutoff value, all of the intervals that have probability of price increase or probability of price decrease greater than the cutoff value are selected. Say there are $N_1$ intervals with a probability of price increase above the cutoff value and $N_2$ intervals with a probability of price decrease above the cutoff value, with $N = N_1 + N_2$. From those $N_1$ intervals that predict a price increase, let $n_1$ be the number of those significant intervals where the next price change that
occurs is an increase. Let \( n_1^* \) be the number of those significant intervals where
the price change is an increase an occurs in that exact interval. Define \( n_2 \) and
\( n_2^* \) similarly for the intervals that predicted a price decrease. The proportion of
significant intervals that correctly predict the next price change (PSIC) is defined
as
\[
\psi = \frac{n_1 + n_2}{N_1 + N_2}.
\]
Another useful measurement of model accuracy is the PSICE (\( \phi \)), or the proportion
of significant intervals that correctly predict the price change in that exact interval,
defined as
\[
\phi = \frac{n_1^* + n_2^*}{N_1 + N_2}.
\]
It is known that \( n_1^* \leq n_1 \leq N_1 \) and \( n_2^* \leq n_2 \leq N_2 \), so both \( \psi \) and \( \phi \) are bounded
by 0 and 1.

Obviously a higher PSIC corresponds to a model that performs better at pre-
dicting price changes, and PSIC values indicate models that could create profitable
strategies. It is not sufficient to be right more often than wrong, because ECN
prices need to move in the same direction twice to turn a one cent per share profit.
Consider a strategy based on a PSIC of \( p \), which means that the probability of
being right about the direction of the next price change is \( p \) and the probabiity of
being wrong is \( 1-p \).

Which values of \( p \) will show a profit, and which values of \( p \) will show a loss?
Consider a simple strategy that buys and sells based on the predicted probabilities
of a model. For simplicity, only one order will be placed, and the order will be held
no longer than two subsequent price changes. If the price moves in the opposite
direction from the one predicted, an event with probability 1-p, the trader unwinds
and takes a two-cent loss. If the price moves in the direction predicted, an event
with probability p, the trader holds it and waits to see the next predicted direction.
If the next predicted direction is not the same as the first price change, the trader
unwinds and earns 0 profit (but also 0 loss). If the next predicted direction is the
same as the first price change, the trader holds the stock until after that second
price change. If the stock moves in that direction again, the inner buy and inner
sell prices are two cents from where they started, and the trader can unwind for
a one cent profit. However, if the stock moves in the opposite direction, the inner
prices are back where they started, and the trader unwinds for a one cent loss.

The trader in the preceding example loses two cents with probability 1-p (one
incorrect prediction), loses one cent with probability \( \frac{1}{2}p(1-p) \) (one correct predic-
tion, one incorrect prediction, and the second change moves opposite first one), and
gains one cent with probability \( \frac{1}{2}p^2 \) (two correct predictions, second change moves
the same as the first). The expected value of the profit \( R \) is

\[
E[R] = (-2)(1 - p) + (-1)\frac{1}{2}p(1 - p) + \frac{1}{2}p^2 = p^2 + \frac{3}{2}p - 2.
\]

The expression for profit equals 0 for \( \sqrt{\frac{1}{4}} \), or around 0.85. Thus, a trader who
uses this strategy that makes decisions based on a model will only gain profit when
the PSIC is greater than 0.85.

The preceding example makes several assumptions. First, it assumes that price
increases and decreasing are equally likely to occur. Also, it assumes that there
will be an interval with significant predicted probability between the two changes. If the second price change occurs before a significant probability is predicted, then the person holds onto the stock. The outcome to this situation is equivalent to the outcome when predicting the price will keep moving in the same direction. Also, the example limits holding a stock to two epochs and limiting the position held to one share. In a situation where the stock is showing price movement in the same direction for several epochs, a much larger profit can be made by holding the stocks for longer and establishing a larger position.

4.4 Fixed-time Method

The purpose of the predictive models developed here is not to say what will happen next but instead to give probabilities for the different possible outcomes. Logistic autoregression fits a model whose output can be transformed by the logit function to yield probabilities of future events as a function of past information. The models created here will use information about movements of shares in previous time intervals to predict probabilities of events in the next time interval. The optimal time-scale was determined to be one second, so these models will predict the probabilities of the possible events in the next second. In any second, the possible outcomes are (1) price increase, (2) price decrease, and (3) no price change. The models that use information about changes in the shares of the order book to predict probabilities of events in the next time interval use what will be called the “fixed-time method.”

If the price increases by a cent and then decreases by a cent within an interval,
the net outcome is no price change. The time-scale of one second was determined as optimal to retain enough information in the intervals to create accurate models. The models presented here use order book information in the previous seconds to compute probabilities for the different possible events in the next second. However, some price changes result from one large order instead of as the result of a predictable pattern. Therefore, it is unrealistic to expect a model to be able to detect and predict every price change, but instead look for models that accurately predict a reasonably large percentage of them.

Logistic regression is often used in a variety of statistical fields, but it is often used for two possible outcomes, 0 and 1. Here, three possible outcomes for price changes are -1, 0, and 1, so multivariate logistic regression must be used. Logistic regression with outcomes 0 and 1 uses the inverse logit transformation \( v = \frac{e^u}{1 + e^u} \) [Tsa02] to obtain a probability between 0 and 1 that the event occurs. The probability that the event does not occur would be obtained by subtracting that value from 1.

There are three possible outcomes \( \{-1, 0, 1\} \) for the response variable \( Y \), so these probabilities are denoted \( P(Y = 1) = \pi_1 \), \( P(Y = 0) = \pi_0 \), and \( P(Y = -1) = 1 - \pi_1 - \pi_0 \). Thus, the response variable can be written as an ordered triplet, where \((y_1, y_0, y_{-1}) = (1, 0, 0)\) indicates \(Y=1\). In generalized linear models, the link function connects the random components (response variable \( Y \)) with the systematic components (explanatory variables) [Agr03]. The probability mass function is

\[
f(y; \pi_1, \pi_0) = \pi_1^{y_1} \pi_0^{y_0} (1 - \pi_1 - \pi_0)^{1-y_1-y_0},
\]
which equals
\[
(1 - \pi_1 - \pi_0) \left( \frac{\pi_1}{1 - \pi_1 - \pi_0} \right)^{y_1} \left( \frac{\pi_0}{1 - \pi_1 - \pi_0} \right)^{y_0}.
\]

Written as the natural exponential family, the pmf is
\[
(1 - \pi_1 - \pi_0) \exp \left[ y_1 \log \left( \frac{\pi_1}{1 - \pi_1 - \pi_0} \right) + y_0 \log \left( \frac{\pi_0}{1 - \pi_1 - \pi_0} \right) \right].
\]

The natural parameters are
\[
\log \left( \frac{\pi_1}{1 - \pi_1 - \pi_0} \right)
\]

and
\[
\log \left( \frac{\pi_0}{1 - \pi_1 - \pi_0} \right).
\]

When only two possible outcomes exist, the probability mass function is the Bernoulli distribution, and the natural parameter is the log odds \( \log \left( \frac{\pi}{1 - \pi} \right) \) [Agr03]. An extension to \( n > 3 \) outcomes would result in similar form to the case where \( n = 3 \), but the number of terms in the denominator would increase as needed [HS05].

Let \( \{X_1, \ldots, X_n\} \) denote the set of explanatory variables for a model, and let \( Y \) be the response variable defined earlier. Then
\[
\pi_1(x) = P(Y = 1|X_1 = x_1, \ldots, X_n = x_n).
\]
\( \pi_0(x) \) is defined similarly, and \( \pi_{-1}(x) \) is determined by subtracting the other sum of the other probabilities from 1. The components of \( X \) are sometimes called “risk factors” when used in the context of biostatistics, where positive coefficients in the fitted model indicate that the factor increases the risk of the event and negative
coefficients indicate that the factor decreases the risk of the event. The log-odds (logit) model takes the form

\[
\text{logit}[\pi_0(x)] = \log \frac{\pi_0(x)}{1 - \pi_0(x)} = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_n x_n
\]

and

\[
\text{logit}[\pi_1(x)] = \log \frac{\pi_1(x)}{1 - \pi_1(x) - \pi_0(x)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \quad \text{[Agr03]}
\]

Earlier, it was determined that between 2.5% and 3% of one-second intervals show a price change. Therefore, when multivariate logistic regression is performed, the probability of a 0 should be much larger than probabilities for 1 and -1. It then becomes necessary to interpret these probabilities and create criteria that signal when an interval is likely to contain a price change. To do this, a cutoff value is selected, and any time the probability of a price increase or price decrease goes above that cutoff value, that time interval is flagged as significant. The selection of this cutoff value is important. If it is too large, there will be very few intervals that are chosen, and therefore the model will fail to identify many price changes. On the other hand, too small of a cutoff value will result in the inclusion of too many intervals that do not contain price changes, creating a large false positive rate. This situation is similar to calculating sensitivity and specificity when determining a proper medical diagnosis.

The models generate probabilities of price changes on each of Island and ArcaBook using shares information from just Island, from just ArcaBook, and from both Island and ArcaBook. At first, the shares information will be the changes in shares at the inner prices for each of the last 3 lags (seconds). In all, six models
are created, with two different dependent variables (Island price changes and ArcaBook price changes) and three different sets of independent variables (last 3 lags on Island, last 3 lags on ArcaBook, last 3 lags from each of Island and ArcaBook). These models will reveal whether price prediction on an ECN is more effective using information from that ECN or from the other ECN and whether using information from both adds to accuracy.

The dependent variables are represented as $y_I(k, t)$ and $y_A(k, t)$ for Island price change and ArcaBook price change, respectively, between times $t$ and $t+1$. Note that only one parameter is needed for time now that the time scale of one second has been established. Here, $k \in \{-1, 0, 1\}$. These dependent variables are not the probabilities of the events, but are the inverse-multivariate logit of the probabilities. The independent variables include the current inner shares and previous lags of changes in inner shares. The current Island inner buy shares is represented as $z_{IB}(t)$, and the current Island inner sell shares is represented as $z_{IS}(t)$. Likewise, $z_{AB}(t)$ and $z_{AS}(t)$ are the ArcaBook inner buy shares and ArcaBook inner sell shares. The lags of changes in inner shares are represented by $x_I(1, t - k)$ for Island and $x_A(1, t - k)$ for ArcaBook. Later, when the shares changes at the secondary peak (standardized price $\pm 0.05$) are included, those variables will be $x_I(5, t - k)$ and $x_A(5, t - k)$.

When predicting Island price changes from Island shares information, the resulting model is denoted as

$$y_I(0, t) = \alpha_0 + \alpha_1 z_{IB}(t) + \alpha_2 z_{IS}(t) + \alpha_3 x_I(1, t - 1) + \alpha_4 x_I(1, t - 2) + \alpha_5 x_I(1, t - 3),$$
The fitted model is

\[ y_I(0, t) = 2.34 + .0001102 z_{IB}(t) - .00000784 z_{IS}(t) \]

\[-.0000367 x_I(1, t - 1) - .00001455 x_I(1, t - 2) - .0000122 x_I(1, t - 3) \]

\[ y_I(1, t) = 0.152 + .0000112 z_{IB}(t) - .000145 z_{IS}(t) \]

\[-.0000596 x_I(1, t - 1) - 0.0000313 x_I(1, t - 2) - .0000258 x_I(1, t - 3) \]

<table>
<thead>
<tr>
<th>( y_I(0, t) )</th>
<th>( \alpha_0 )</th>
<th>( \alpha_1 )</th>
<th>( \alpha_2 )</th>
<th>( \alpha_3 )</th>
<th>( \alpha_4 )</th>
<th>( \alpha_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>2.34</td>
<td>1.102e - 4</td>
<td>-7.84e - 6</td>
<td>-3.67e - 5</td>
<td>-1.45e - 5</td>
<td>-1.226e - 5</td>
</tr>
<tr>
<td>se</td>
<td>1.503e - 10</td>
<td>8.115e - 6</td>
<td>3.257e - 6</td>
<td>7.285e - 6</td>
<td>8.822e - 6</td>
<td>9.302e - 6</td>
</tr>
<tr>
<td>p</td>
<td>0.066</td>
<td>&lt; 2e - 16</td>
<td>1.76e - 14</td>
<td>6.03e - 9</td>
<td>0.00294</td>
<td>0.01680</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( y_I(1, t) )</th>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_3 )</th>
<th>( \beta_4 )</th>
<th>( \beta_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>0.152</td>
<td>1.12e - 4</td>
<td>-1.448e - 4</td>
<td>-5.964e - 5</td>
<td>-3.126e - 5</td>
<td>-2.576e - 5</td>
</tr>
<tr>
<td>se</td>
<td>2.024e - 10</td>
<td>8.894e - 6</td>
<td>1.217e - 5</td>
<td>1.211e - 5</td>
<td>1.308e - 5</td>
<td>1.381e - 5</td>
</tr>
<tr>
<td>p</td>
<td>0.084</td>
<td>4.27e - 10</td>
<td>3.69e - 7</td>
<td>3.43e - 13</td>
<td>5.7e - 4</td>
<td>5.89e - 4</td>
</tr>
</tbody>
</table>

Only two equations are necessary, as there are three variables whose sum is 1. The models estimate probabilities of price increase and no change, and the probability of price decrease can be calculated by simply subtracting from 1. Of course, \( y_I(-1, t) \neq 1 - y_I(0, t) + y_I(1, t) \), since the dependent variables in the fitted model are not the probabilities of the events.

As per the coefficients, the probability of predicting a price increase rises as the inner buy shares increases, as the inner sell shares decreases, and as any of the last
3 lags of shares changes decreases. The inner buy and inner sell shares reveal how much of a shift in the order book would be needed to empty out the shares at the inner prices. A larger quantity of shares at the inner buy and a smaller quantity at the inner sell indicates a smaller shift is needed for the inner sell to empty, making a price increase easier if a change in shares occurs in that direction. Therefore, the signs of the coefficients on the inner buy shares and inner sell shares agree with intuition.

The probability of a price increase rises as the values of the last 3 shares differences drops. When the shares difference is positive, the net difference in shares at the inner sell is larger than the net difference in shares at the inner buy. That positive net differences results from one of three possible situations: (1) a net increase in shares at the inner sell and a net decrease in shares at the inner buy, (2) a net increase in shares at the inner sell that is larger than the net increase in shares at the inner buy, or (3) a net decrease in shares at the inner sell that is smaller than the net increase in shares at the inner buy. For all three of these situations, the conditions for the emptying out of the inner buy shares is becoming more likely than the emptying out of the inner sell shares. Intuitively, a positive shares difference coincides with a decrease in the probability of a price decrease, agreeing with the signs on the coefficients. When the models that predict Island price changes using only ArcaBook data or both Island and ArcaBook data, the coefficients follow the same signs as the model using Island.
4.4.1 Model Diagnostics

Now that the model has been fitted, the logit transform can be applied to model estimates in order to yield probabilities of the different possible outcomes. The next step is to test how effective these probabilities are at predicting events. At each second, the probabilities of the three possible outcomes are calculated. To evaluate the success of the models and their output, a variety of techniques are applied here.

One simple way to evaluate the model's effectiveness graphically is through boxplots, which display the distribution of the predictive probabilities for different situations. Using the model that predicts Island price changes using only Island information, the first row of graph Figure 4.11 shows the boxplots of the probability of price decreasing for intervals where the price decreased (left), stayed the same (middle), and increased (right.) The second row of graph (12) shows the same boxplots but for probability of price increasing. The predicted probabilities of the price increasing are generally larger during intervals when the price actually does increase than when there is no change or a price decrease. The same is true for the predicted probabilities of a price decrease. Obviously, an accurate model would generate higher probabilities of a price decrease during intervals where prices do decrease and generate higher probabilities of a price increase during intervals where prices do increase. The boxplots indicate the model does exactly that. The boxplots of results for predicting Island price changes using only ArcaBook information and for predicting Island price changes using shares information from both Island and ArcaBook are shown in graphs Figure 4.12 and Figure 4.13, respectively.
93

Figure 4.11: Distribution of predicted probabilities for different outcomes using only Island data.

Figure 4.12: Distribution of predicted probabilities for different outcomes using only ArcaBook data.
Figure 4.13: Distribution of predicted probabilities for different outcomes using Island and ArcaBook data

ROC curves can evaluate the effectiveness of the predicted probabilities in the just as they can evaluate the success of a medical diagnosis decision rule. In the same way that a patient with an elevated level of a certain cell count is flagged as likely to have a certain disease, an interval of time can be flagged as likely to contain a price change based on the output from the model. ROC curves plot 1-specificity on the x-axis and plot sensitivity on the y-axis, and the various points represent the results when using different cutoff values. An ideal result would be close to (0,1), as this would correspond to both sensitivity and specificity equal to 1. ROC curves are useful in quickly determining which model has overall better accuracy of prediction by observing which model yields an ROC curve with greater area under the curve.

Graph Figure 4.14 shows the ROC curves for predicting Island price changes using information from Island (black), ArcaBook (red), and both (blue). Clearly
ArcaBook performs the worst, although it is still significantly above the line of chance \((y=x)\) for most of the cutoff values. Therefore, ArcaBook does contain some information about impending Island price changes, but not nearly as much as Island itself contains. The results for using Island by itself and in combination with ArcaBook are both good, with the results using both slightly outperforming Island by itself. Thus, although the Island information is much more informative than the ArcaBook data, the ArcaBook data does contribute and including it in the model improves predictive success.

Figure 4.14: ROC curves for predicting Island price changes

![ROC curves for predicting Island price changes](image)

Figure 4.15 shows the PSIC for different cutoff values \((x)\) that correctly predicted the direction of the next Island price change. Again, black indicates only Island data, red indicates only ArcaBook data, and blue indicates both. Here, black outperforms blue more obviously for many of the cutoff values. Both blue and black show monotone increasing trends as the cutoff value increases, which is expected as
using the higher cutoff values selects the intervals the model deemed more probable to change. Figure 4.16 shows the same PSIC plots along with the 95% confidence bands for all three models. Figure 4.17 shows the PSICE, which yields similar results between the models until cutoff value 0.1, when the model using Island by itself greatly outperforms the other two.

Figure 4.15: PSIC values using different models to predict Island price changes

The graph of the PSIC shows a general increasing pattern as the cutoff value increases. As a result, selecting higher cutoff values may appear to be an obvious choice when developing profitable strategies. However, higher cutoff values classify a smaller amount of intervals as significant, meaning that the lower cutoffs that result in more significant intervals but a smaller percent correct may end up being more profitable. For example, using Island and ArcaBook, a cutoff value of 0.1 results in 449 intervals with probabilities greater than 0.1, of which 366 (81.5%) are correct. Increasing the cutoff value to 0.2 increases the probability of being correct
Figure 4.16: PSIC values with estimation errors

Figure 4.17: PSICE values using different models to predict Island price changes
to 89.2%, or 150 out of 168. Although the probability has increased by increasing the cutoff value, the overall number correct has decreased significantly, which may decrease profitability.

An issue arises if both the probability of price increase and the probability of price decrease are above the cutoff value. Figure 4.18 shows the plot of probabilities of price increase \( y \) vs. probabilities of price decrease \( x \) generated from the models. There is an obvious pattern that shows that when one of the probabilities is large, the other one tends to be small. The conditions that make a price increase probable also make a price decrease improbable and vice versa, making this problem a very rare occurrence. In fact, when using a cutoff value of 0.02, there are 1324 intervals where the probability of an increase is significant, and 1727 instances where the probability of a decrease is significant. However, there is no overlap between those sets, as none of those intervals are significant for both directions.

![Figure 4.18: Probability of price increase vs. probability of price decrease](image-url)
The next set of models repeats the preceding analysis but predicts ArcaBook changes instead of Island changes. Once again, the different models use only Island, only ArcaBook, and both ECNs. The model predicting ArcaBook using only ArcaBook shares information is

\[
y_A(0, t) = 1.85 + .0001604z_{AB}(t) + .00000986z_{AS}(t) \\
- .0000567x_A(1, t - 1) - .0000181x_A(1, t - 2) - .00000105x_A(1, t - 3)
\]

\[
y_A(1, t) = 0.3324 + .0001325z_{AB}(t) - .0001826z_{AS}(t) \\
- .0001022x_A(1, t - 1) - .0000334x_A(1, t - 2) - .0000542x_A(1, t - 3).
\]

As with the models for predicting Island price changes, the inner buy shares has a positive coefficient while the inner sell shares has a negative coefficient, so the model makes sense. Likewise, the coefficients on the shares changes are all negative, and these patterns repeat when the Island data is included in the model.

The ROC curves are shown in Figure 4.19. As before, the model that uses
only the opposite ECN (in this case Island) performs the worst. However, Island does better at predicting ArcaBook price changes than ArcaBook did at predicting Island price changes, which agrees with the earlier findings that price changes that occur on both ECNs occur first on Island a large portion of the time. The remaining two models, using just ArcaBook and using both, performed similarly well, with the model that uses both ECNs to predict ArcaBook changes performing slightly better. This result agrees with the findings when predicting Island price changes.

![ROC curves for predicting ArcaBook price changes](image)

**Figure 4.19:** ROC curves for predicting ArcaBook price changes

Figure 4.20 and Figure 4.21 show PSIC values and PSICE values for predicting the next ArcaBook price change using Island (black), ArcaBook (red), and both (blue). Unlike when predicting Island, there is not one model that obviously outperforms the other two, although the blue graph (using both Island and ArcaBook to predict ArcaBook price changes) performs slightly better for larger cutoff values.
Figure 4.20: PSIC values for predicting ArcaBook price changes

Figure 4.21: PSICE values for predicting ArcaBook price changes
4.4.2 Model Variations

Thus far, the best models for predicting price changes on either ECN use shares data from both ECNs, with the next best model using shares data from only that ECN, and the worst model using shares data from the opposite ECN. However, there are several modifications that must be implemented for evaluation of their effectiveness in comparison to those models that have already been presented. First, instead of using the last three lags, only the most recent lag will be included in the model. Next, instead of using shares information at the inner price (primary peak), the models will use independent variables from the shares movement at the secondary peaks (five cents away). Last, interaction between the independent variables will be introduced to the models and their effectiveness evaluated.

Previously, it was shown that the last three lags were significant for price change prediction models using a one second time scale. However, it may be possible to achieve similar results using only the last lag and omitting the previous two. To test this, the models predicting price changes on Island and ArcaBook using information from both ECNs are refit using only the most recent lag. Figure 4.22 shows the ROC curves for predicting Island price changes using only one lag from Island and ArcaBook (brown) and using the last three lags from Island and ArcaBook (green). The curves appear identical for many values and only for high sensitivity is there a slight decrease in predictive accuracy when using only the most recent lag. Graph Figure 4.23 shows the same results but for predicting ArcaBook price changes. Once again, there is a slight decrease in predictive success from using only the most
recent lag (brown) instead of the last three lags (green).

Figure 4.22: ROC curves predicting Island changes using 3 lags vs. using 1 lag

Figure 4.24 and Figure 4.25 show the PSIC for Island (first graph) and ArcaBook (second graph) using just one lag of Island and ArcaBook (brown) and the last three lags for both (green). The two plots follow each other very closely, and there seems to be no predictive lost by using only one lag instead of three. This fact will be important for future model development, as well as increasing speed of computability by reducing the number of variables. When attempting to gain profit by implementing models that calculate probabilities every second, requiring fewer variables will be advantageous.

The next modification expands the models to include shares information from the secondary peak at standardized price ±0.05. To do this, models will be fitted and results will be compared for models predicting price changes on Island and ArcaBook using (1) only shares information at the inner peak, (2) only shares
Figure 4.23: ROC curves predicting ArcaBook changes using 3 lags vs. using 1 lag.

Figure 4.24: PSIC values for Island changes using 3 lags vs. using 1 lag.
Figure 4.25: PSIC values for ArcaBook changes using 3 lags vs. using 1 lag

information at the outer peak, (3) shares information from both peaks. To avoid
an inundation of variables, these models will only use the most recent lag, as this
change was shown to have negligible impact on the result.

Figure 4.26 shows the ROC curves for predicting Island price changes using
shares information at the inner peak (black), outer peak (red), and both peaks
(blue). The three curves are almost concurrent for most of the values, indicative
that the inner and outer peaks contain the same amount of information. The ROC
curves for predicting ArcaBook price changes also show near equivalence between
the models. However, plotting the PSIC in Figure 4.27 shows the model that
uses only the secondary (0.05) peak (red) performing slightly worse than the other
two for larger cutoff values. The movement at the inner peak results in the price
changes, so it is not surprising that the models using information from the inner
(primar) peak have a predictive advantage over the models using information from
the outer (secondary) peak. The results from the three models are very close in both the ROC curves and PSIC plots, so only slight differences exist across the models.

![ROC curves for Island price changes using different peaks](image)

Figure 4.26: ROC curves for Island price changes using inner peak, outer peak, and both peaks

Overall, the difference between models using variables based on changes in shares at the inner peak (± 0.01) and those using variables based on the outer peak (± 0.05) is minimal. The information resulting from the actions of the aggressive traders who place orders close to the inner prices for quick execution and that of the cautious traders who place orders a few cents away to wait for a better price have the same general impact on the prediction of price changes.

Thus far, all of the variables have been used in the models with no interaction. However, adding interaction terms may improve the predictive success of the models. Interaction terms between Island and ArcaBook shares information and between shares changes and amounts of inner shares are added to the model, and
the results are compared to the original model without interaction terms. The equations for the model with interaction terms are

\[
y_I(0, t) = 1.932(10^{-4}) + 7.296(10^{-5})z_{IB}(t) + 1.969(10^{-5})z_{AB}(t)
\]
\[+ 1.438(10^{-5})z_{IS}(t) + 8.144(10^{-5})z_{AS}(t)
\]
\[- 8.155(10^{-5})x_I(1, t - 1) - 6.181(10^{-5})x_A(1, t - 1)
\]
\[+ 2.426(10^{-9})z_{IB}(t) \ast z_{AB}(t) - 7.495(10^{-10})z_{IS}(t) \ast z_{AS}(t)
\]
\[+ 2.186(10^{-9})z_{IB}(t) \ast x_I(1, t - 1) + 2.798(10^{-9})x_I(1, t - 1) \ast x_A(1, t - 1)
\]

\[
y_I(1, t) = -1.499 + 4.988(10^{-5})z_{IB}(t) - 1.196(10^{-5})z_{AD}(t)
\]
\[+ 6.828(10^{-5})z_{IS}(t) + 1.105(10^{-4})z_{AS}(t)
\]
\[- 5.459(10^{-5})x_I(1, t - 1) - 1.145(10^{-4})x_A(1, t - 1) \ast x_A(1, t - 1)
\]
\[+ 4.37(10^{-9})z_{IB}(t) \ast z_{AB}(t) - 5.610(10^{-9})z_{IS}(t) \ast z_{AS}(t)
\]
\[+ 1.398(10^{-9})z_{IB}(t) \ast x_I(1, t - 1) + 2.607(10^{-9})x_I(1, t - 1) \ast x_A(1, t - 1).
\]
Figure 4.28 shows ROC curves for the original model (black) and the model with interaction terms (red). The two curves follow each other very closely, and the model without interaction terms performs slightly better for some of the values. When plotting the PSIC values, as shown in graph Figure 4.29, the two models yield similar results for cutoff values less than 0.1. However, between 0.1 and 0.2 the percentage for the model without interaction terms continues to increase while the percentage for the model with interaction decreases. The lack of improvement and loss of accuracy resulting from adding interaction terms demonstrates an absence of interaction between the independent variables.

![ROC curve for Island price changes using Interaction (red) or not (black)](image)

Figure 4.28: ROC curves for Island price changes using interaction terms (red) and no interaction terms (black)

4.5 Fixed-time means model

One variation on the fixed-time method uses means of added, deleted, and traded orders within the one-second intervals instead of differences in the quantities of
Figure 4.29: PSIC values for Island price changes using interaction terms (red) and no interaction terms (black)

shares as the independent variables. The outcomes \{-1,0,1\} remain the same. As was done before, the prices used to calculate the means are the standardized prices. Only orders within ten cents of the inner prices will be included in the calculation of means to remove the outliers that have very low probability of being traded but would have a great influence on means. Some people add buy orders at $0.01 (not standardized price) or sell orders at $1000, which have no chance of being traded.

The first set of models to fit using the fixed-time means model has the dependent variable as the price changes and independent variables consisting of the inner buy shares ($z_{IB}(t)$ for Island and $z_{AB}(t)$ for ArcaBook), the inner sell shares ($z_{AB}(t)$ for Island and $z_{AS}(t)$ for ArcaBook), the mean of added orders in the last second ($m_{I1}(t-1)$ for Island and $m_{A1}(t-1)$ for ArcaBook), and the mean of deleted or traded orders in the last second ($m_{I2}(t-1)$ for Island and $m_{A2}(t-1)$ for ArcaBook). Here, the means are weighted by the number of shares in the order being added or
The model fitted for predicting Island price changes using Island only is

\[
y_{I}(0, t) = 2.055 + 0.0001141 z_{IB}(t) - 0.00000484 z_{IS}(t) \\
\quad + 7.055 m_{I1}(t - 1) - 0.0118 m_{I2}(t - 1)
\]

\[
y_{I}(1, t) = 0.035 + 0.000114 z_{IB}(t) - 0.000146 z_{IS}(t) \\
\quad + 12.187 m_{I1}(t - 1) - 1.97 m_{I2}(t - 1).
\]

<table>
<thead>
<tr>
<th>(y_{I}(0, t))</th>
<th>(\alpha_{0})</th>
<th>(\alpha_{1})</th>
<th>(\alpha_{2})</th>
<th>(\alpha_{3})</th>
<th>(\alpha_{4})</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>2.055</td>
<td>1.141e - 4</td>
<td>-4.842e - 6</td>
<td>7.055</td>
<td>-0.0118</td>
</tr>
<tr>
<td>se</td>
<td>1.304e - 10</td>
<td>7.975e - 6</td>
<td>3.123e - 6</td>
<td>1.058e - 12</td>
<td>1.063e - 12</td>
</tr>
<tr>
<td>p</td>
<td>0.0365</td>
<td>&lt; 2e - 16</td>
<td>&lt; 2e - 16</td>
<td>7.48e - 6</td>
<td>2e - 12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(y_{I}(1, t))</th>
<th>(\beta_{0})</th>
<th>(\beta_{1})</th>
<th>(\beta_{2})</th>
<th>(\beta_{3})</th>
<th>(\beta_{4})</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>0.0353</td>
<td>1.14e - 4</td>
<td>-1.457e - 4</td>
<td>12.187</td>
<td>-1.97</td>
</tr>
<tr>
<td>se</td>
<td>1.970e - 10</td>
<td>8.680e - 6</td>
<td>1.208e - 5</td>
<td>3.697e - 12</td>
<td>1.026e - 12</td>
</tr>
<tr>
<td>p</td>
<td>0.0299</td>
<td>&lt; 2e - 16</td>
<td>&lt; 2e - 16</td>
<td>1.01e - 5</td>
<td>1.84e - 12</td>
</tr>
</tbody>
</table>

As with the previous fixed-time models, the probability of a price increase rises as the current inner buy shares increase and the current inner sell shares decrease. In addition, a larger mean of added orders and a lower mean of deleted/traded orders drives the probability of a price increase higher as well. These coefficients make intuitive sense as well, due to the secondary peak that results from added and deleted orders at a price several cents from the inner price. The activity in the secondary peaks follows the movement of the inner prices, so when the price
is increasing, more shares are being added at higher prices. These traders hope the price movement reaches their order after a few changes. This model follows a similar methodology as the SOBI strategy, which uses the means of the orders currently sitting in the order book to determine when to buy and sell. Here, the means use changes in the order book rather than the orders themselves.

Figure 4.30 shows the ROC curves for predicting Island price changes as a function of means of added and deleted orders from just Island (black), from just ArcaBook (red), and from both ECNs (blue). The ROC curves for the means models are similar to those for the shares models in that Island is predicted worst by just ArcaBook, and the other two are close with the model that uses both showing slightly better prediction results than the model that uses only means data from Island. Figure 4.31 shows the ROC curves when the dependent variable is ArcaBook price changes. The curves from the models using only ArcaBook and using both are very close, with the model using only Island performing significantly worse.

Figure 4.32 shows the ROC curves for predicting Island price changes using both Island and ArcaBook data for the fixed-time method (green) and the fixed-time means method (purple). The accuracy of prediction is slightly worse for these means models than for the previous models that used differences in shares.

Figure 4.33 shows the PSIC for Island price changes using data from Island (black), ArcaBook (red), and both (blue). The results are similar to the ROC curves, with ArcaBook showing the worst predictive accuracy. However, unlike
Figure 4.30: ROC curves for Island price changes using the fixed-time means method

Figure 4.31: ROC curves for ArcaBook price changes using the fixed-time means method
The previous models included the weighted means of both the added and removed (traded or deleted) in the independent variables. The next set of models will use only the means of the added orders or the means of the removed orders, and these results will be compared to the models that contain both. Figure 4.35 shows ROC curves for the models using means of only the added orders (black), only the deleted or traded orders (red), and both the added orders and deleted orders (blue). These curves are almost indistinguishable, which reveals that the
Figure 4.33: PSIC values for Island price changes using Island (black), ArcaBook (red) and both (blue)

Figure 4.34: PSIC values for ArcaBook price changes using Island (black), ArcaBook (red) and both (blue)
information gained from the means of orders entering the order book is the same as the information gained from the means of orders leaving the order book. The PSIC values for different cutoff values also show close similarity between the three models. Thus, no extra information is gained or lost by including or omitting either variable if the other one is already one of the independent variables.

Figure 4.35: ROC curves using the fixed-time means method for added orders (black), deleted/traded orders (red), and both (blue)

Figure 4.36 shows the ROC curves for the fixed-time method (brown) and the fixed-time means method (purple). The fixed-time method outperforms the fixed-time model, although the two curves are close. The fixed-time model only includes activity at selected values in the order book, while the means model includes all order book activity within a reasonable distance. The greater success of the model that uses information from fewer standardized prices shows the importance of that information at those prices with regards to price movement.
4.6 Epoch Method

The next model will still maintain a fixed time-scale for observations, but instead of predicting probabilities of what will occur in the next time, the model will predict probabilities for the next direction of price change. Therefore, the possibility of no change is removed from the outcomes, leaving only +1 and -1. The independent variables will be similar to those for the fixed-time method, but instead of using only the change in shares for the last second, the total change in shares since the last price change (in other words, since the start of the current epoch) will be used.

As with other models, the data is split into two parts, with the first half used to develop the model and the second half used to test the model. Each second, the model yields a probability that the next price change will be an increase and a probability that the next price change will be a decrease. These two probabilities will add up to 1, as these are the only two possible outcomes. The only problem
would arise if the trading day ends before the next change occurs, but the window of time used ends well before the trading day ends to ensure that another price change occurs.

Now that there are only two outcomes, the multivariate logit transform is no longer needed. Instead, the inverse logit transform $y = \frac{e^x}{1+e^x}$ can be used to find probabilities, with $P(Y = 1) = 1 - P(Y = -1)$. The variables $z_{IB}(t)$ and $z_{IS}(t)$ for the Island inner buy and Island inner sell shares remain the same. Now $w_{IB}(t)$ and $w_{IS}(t)$ represent the change in Island inner buy and inner sell shares since the previous price change. $w_{AB}(t)$ and $w_{AS}(t)$ are defined similarly for ArcaBook.

The resulting model is

$$y_{I}(t) = y_{I}(t) - 1.061 + 0.00006441z_{IB}(t) - 0.00005903z_{IS}(t) - 0.00003386w_{IB}(t) - 0.00004435w_{IS}(t).$$

<table>
<thead>
<tr>
<th>$y_{I}(0, t)$</th>
<th>$\alpha_0$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>-1.061</td>
<td>6.441e-5</td>
<td>-5.903e-6</td>
<td>-3.386e-5</td>
<td>-4.435e-5</td>
</tr>
<tr>
<td>se</td>
<td>0.126</td>
<td>3.459e-6</td>
<td>2.807e-6</td>
<td>3.32e-6</td>
<td>3.126e-6</td>
</tr>
<tr>
<td>p</td>
<td>&lt; 2e-16</td>
<td>&lt; 2e-16</td>
<td>0.0354</td>
<td>&lt; 2e-16</td>
<td>&lt; 2e-16</td>
</tr>
</tbody>
</table>

As with the fixed-time models, the probability of a price increase rises as the inner buy shares increases and the inner sell shares decreases. However, the coefficients on the change in inner buy shares since the beginning of the epoch and the change in inner sell shares since the beginning of the epoch both take negative values. Typically, these coefficients would take opposite signs, as in the fixed-time and fixed-time methods, because their behaviors indicate opposite movements in
the order book.

To evaluate the accuracy of the model, each observation can be designated as a predicted increase or decrease based on its predictive probabilities. Using 0.5 as a cutoff, every second there is a prediction as to whether the next change will be an increase or decrease. However, this cutoff value can be adjusted if the numbers of price increases and price decreases are not equal. As with previous models, ROC curves show the results when different cutoff values are used, as shown in Figure 4.37. As before, the black curve represents the model that uses only Island information, the red curve represents the model that uses only ArcaBook data, and the blue curve represents data using both Island and ArcaBook data.

![ROC curve for Island price changes using the Epoch method](image)

Figure 4.37: ROC curve for Island price changes using the Epoch method

Unlike the fixed-time method that uses information from one-second intervals to compute the predictive probabilities, the epoch method uses information since the last price change to compute the probabilities. Therefore, as the epoch pro-
gresses and the time since the last price change increases, there should be more and more information contained in the variables. To determine how accuracy of prediction changes over time, the predicted direction of the next price change (based on whether the predicted probability of increase is greater than or less than 0.5), the elapsed time since the last price change, and the time until the next price change will be needed for each second.

There are three methods for determining how predictive accuracy changes throughout an epoch. The first method uses the time since the previous price change, which is equivalent to the time elapsed since the beginning of the epoch. The second method uses the time until the next price change, which will not be known at the time but is known in hindsight. The third method considers the proportion of the epoch that has passed; that is, the time since last price change divided by the total time length of the epoch. Again, this information is not known at the time but is available in hindsight.

The first method looks at the percentage predicted correctly as a function of time since the previous price change. To do this, the observations are clustered using the greatest integer or "floor" function. To calculate the percentage correctly predicted at n seconds, include all points whose time since the previous change were between n and n+1 seconds. The plot of percentage correctly predicted versus time since last price change can be seen in graph Figure 4.38. The proportions correctly predicted remain relatively consistent over time, with a peak emerging at around 45 seconds. Of course, the later values will have fewer observations, as many epochs
last just a few seconds.

Figure 4.38: Proportion correctly predicted vs. time since last price change

Next, look at correctness probability as a function of time until the next price change. Although first guess may be to expect a trend similar but opposite of the trend observed in the plot for time since the previous change, it is important to keep in mind that the epochs have varying time durations. The plot of percentage correctly predicted versus time until next price change can be seen in graph Figure 4.39. This plot has a much more monotone pattern. As expected, the probability is maximum when the time until the next change is at a minimum. However, as the time moves further from the price change, the correctness probability drops. The probability shows a monotone decrease from above 0.9 to 0.5 as the time moves from 0 to 60 seconds. However, from 60 seconds to around 100 seconds, an increase in probability from 0.5 to 0.6 occurs. With the activity in the order book changing as rapidly as it does, the model's loss of accuracy as the time moves away from the
future price change agrees with the conclusion in a previous chapter that the order book is “young.”

Figure 4.39: Proportion correctly predicted vs. time until next price change

The next step calculates the proportion of distance from the last price change to the next price change. The plot of the probability of correct prediction versus the calculated proportion of distance is in graph Figure 4.40. Here, an obvious monotone increasing trend from the start until the end of an epoch is visible. Of course, that does not mean that will be true within an individual epoch, just in general for all epochs. From the beginning of the epoch until about 80% of the way into an epoch, the graph shows a trend that is almost linear, increasing from 0.5 to around 0.7. For the last 20% of an epoch, there is a much steeper slope, with the probability increasing from 0.7 to 0.9. This increasing trend agrees with the observations from the previous plots.

The preceding results provide a good idea of when the epoch models yield the
Figure 4.40: Proportion correctly predicted vs. proportion of epoch completed best predictive success. The epoch method has the best predictability later in an epoch, shortly before the next price change occurs. As an epoch progresses, the information about shares movement within that epoch gets greater and greater, resulting in the increased accuracy that was observed. The result when calculating predictive success versus time since the previous price change does not appear to agree with the conclusion of the other two graphs. The great variability of the lengths of epochs prevents any consistent pattern from emerging.

Thus far, the probabilities every second have resulted in categorization as either predicting the next change will be a price increase or a price decrease. However, as with the fixed-time method, profitable strategies will be developed by identifying certain times to buy and sell. Therefore, instead of using 0.5 as the cutoff value for significance, increasing the cutoff value would pick out only intervals where the models' output yields greater certainty about the direction of the next price change.
For example, if the cutoff value were 0.7, then only times when the probability of increase is greater than 0.7 (probability of decrease is less than 0.3) will be identified as predicting a price increase, and only times when the probability of decrease is greater than 0.7 (probability of increase is less than 0.3) are identified as predicting a price decrease.

Figure 4.41 shows the plots of the PSIC using the epoch method for predicting price changes on Island using data from only Island (black), only ArcaBook (red), and both ECNs (blue). When the cutoff value is between 0.6 and 0.8, the model using only Island data performs slightly better than the model that uses both, which in turn performs slightly better than the model that uses only ArcaBook. For cutoff values greater than 0.8, the model that uses ArcaBook shows drastic improvement while the other two show significant decline. Although these higher cutoff values will designate successively fewer and fewer intervals as significant, this stark difference is unlikely to occur due to chance. To test this hypothesis, these results are compared to the results when ArcaBook price change is the dependent variable using the same independent variables.

Figure 4.42 repeats the previous plots but for predicting ArcaBook price changes. These curves show the same pattern of extreme success when using ArcaBook information to predict and extreme failure when using Island data to predict. The same pattern of poor prediction for larger cutoff values when using Island data appears. Upon further investigation, a couple of problematic epochs late in the trading day cause the poor results. Near the beginning of these epochs, a shift in
Figure 4.41: PSIC values for Island price changes using the Epoch method

shares occurs that would indicate a likely price change in one direction. However, no price change occurs, and at some future time the shares shift in the opposite direction, and a price change occurs in the direction opposite of the one previously predicted. For most of these epochs, a price change in one direction is predicted with high probability, but all of these predictions turn out to be incorrect. One such epoch had an order added at the inner buy price for 90,000 shares, which resulted in the inner buy shares dwarfing the inner sell shares. However, as this 90,000 share order was slowly matched with incoming orders, the balance between the order books was restored. During this time, the large number of shares at the inner buy price had an impact on the resulting probabilities, which showed a high probability of a price increase during those intervals.

The fixed-time method did not have this same issue of being influenced by large orders early in epochs, because only shares changes in the previous few seconds have
an effect on the predicted probabilities of price changes. The epoch method has a much longer “memory” of past shares changes, which leads to the problems faced when the PSICs are calculated. For this 90,000 share order, the epoch method shows a large imbalance for many intervals after the order is added. However, the fixed-time method only uses the preceding three intervals, so this large order cannot influence as many probabilities as in the epoch method. Also, as the large order is traded, the fixed-time method would capture the removal of shares from the buy order book and the probabilities would reflect this movement. The epoch method would be more influenced by the large number of shares still remaining from the large order, because the epoch method considers cumulative changes since the beginning of the current epoch. Likewise, this issue helps explain why the coefficients of the epoch method do not match with intuition.

![Figure 4.42: PSIC values for ArcaBook price changes using the Epoch method](image)

In this section, the epoch method was presented as a way to utilize shares
information for the preceding portion of an epoch to predict the direction of the next change. One noteworthy result from the PSIC plots is that the fixed-time method showed greater success at predicting the direction of the next price change, even though the epoch method was designed to do exactly that. The fixed-time method only "remembers" for three seconds, but this short memory enables the model to ignore past aberrations in the order book. The epoch method did show some success, but it also kept too good a memory of past information that may no longer be relevant.

4.7 Model Updating

The models created using the fixed-time or epoch methods were formulated using one part of the day and tested on a later part of the day. Data from 10:00 AM to 12:00 PM established the proper coefficients for the variables in the model, and those coefficients were used to calculate probabilities of event for 12:00 PM to 3:00 PM. However, it is possible to include more and more data into the model's independent variables as the day progresses, a process called "model updating." For example, the model used to predict probabilities of price changes at 2:00 PM calculates its coefficients using "old" data from 10:00 AM - 12:00 PM. Perhaps including all of the data up to that time will have more accurate predictions than the ones that only use data from the beginning part of the day.

To test whether model updating is successful, a model predicting probabilities of price changes during the 1:00 PM - 2:00 PM hour using only the data from 10:00 AM - 12:00 PM (original model) will be compared to a new model created using
all of the data from 10:00 AM - 1:00 PM (updating model). A simple model using only the last lag of Island and the current inner shares will be used to see if the coefficients change significantly. The original model is

\[
y_I(0, t) = 2.211 + 0.0001129z_{IB}(t) - 0.000006719z_{IS}(t) - 0.00003877x_I(1, t - 1)
\]

\[
y_I(1, t) = 0.127 + 0.0001136z_{IB}(t) - 0.0001465z_{IS}(t) - 0.00006445x_I(1, t - 1),
\]

and the updating model is

\[
y_I(0, t) = 2.037 + 0.0001221z_{IB}(t) - 0.000002487z_{IS}(t) - 0.00004085x_I(1, t - 1)
\]

\[
y_I(t) = 0.446 + 0.0001068z_{IB}(t) - 0.0001447z_{IS}(t) - 0.00006881x_I(1, t - 1).
\]

In general, the coefficients do not differ greatly between the original model and the updating model. The largest difference occurs for the y-intercept of the \(Y_I(1, t)\) equation, where the updating model has a value four times that of the original model. The rest of the coefficients take similar values for the two models. The earlier plots of coefficients changing over time lead to the hypothesis of there being little difference between the results from the models, as the extra time included in the independent variables have little impact on the values of the coefficients. The
lack of a significant difference between the coefficients of the original model and the updating model agrees with this assessment.

Figure 4.43 shows the plot of the ROC curves for probabilities of events in the interval between 1:00 PM and 2:00 PM using data from 10:00 AM - 12:00 PM (black) and 10:00 AM - 1:00 PM (red). These curves are nearly concurrent for most of the cutoff values, indicating little difference between the predictive success of these models. Figure 4.44 shows the plots of the PSIC values versus different cutoff values for the two different models. Although they have similar PSIC values for smaller (below 0.10) cutoff values, the larger cutoff values show that the updating model correctly predicts a greater percentage correctly. At cutoff value 0.14, the original model predicts the correct direction of the next price change for 128 out of the 165 significant intervals (77.6%). The updating model predicts correctly for 115 out of 129 (89.1%). Although the updating model predicts fewer true positives (115) than the original model (128), the updating model significantly cuts down on false positives (14 instead of 37).

If the larger cutoff values reveal an improvement in PSIC between the two models, why was this not visible on the ROC curves? The PSIC measures the percentage of the significant intervals that correctly predict the next price change. The numerator for the PSIC is number correct, and the denominator is the total number of significant intervals. For the ROC curves, the y-coordinate is sensitivity, which is the number of true positives divided by the total number of events that occur. Looking closely at the portion of the ROC curves with high sensitivity, the
Figure 4.43: ROC curve using the original model (black) and updating model (red)

Figure 4.44: PSIC values using the original model (black) and updating model (red)
curve for the updating model (red) has higher sensitivity (lower false positives),
which agrees with the PSIC results. The difference on the ROC curves was harder
to see due to the smaller probabilities involved.

The results for predicting probabilities of events in 2:00 PM - 3:00 PM using the
original and updating models agree with the results when predicting probabilities
of events in 1:00 PM - 2:00 PM. For a given cutoff value, the updating models
designate fewer intervals as significant, but a large percentage of those significant
intervals correct predict the next price change. In the next chapter, profitable
strategies will be developed using the models created in this chapter, including the
updating model. As with the decision regarding cutoff values, the profitability of
the original model versus the updating model will depend on whether percentage
of correct intervals is more important than sheer number of correct intervals.

Updating the model as time progresses does yield slightly more accurate results
than the original model that was fitted without the updating. However, some
difficulties may occur when trying to create a system that updates the model during
the day. An automated agent that buys and sells according to the results of a model
may be locked into a strategy, and making adjustments to the model may not be
possible. Also, the time it takes to update the model is time lost, and these models
work on such a small time scale that no time can be wasted.

4.8 Asymmetry models

Thus far, all of the techniques used part of the data to determine the appropriate
coefficients for the model and used another part of the data to apply the model
and to determine its effectiveness at predicting future price changes. However, it is possible to find a statistic that represents the information about the changes in shares on the order book for each second. As with before, intervals would be deemed to predict a price increase if this statistic is above a certain cutoff value and would be deemed to predict a price decrease if below a certain other value. Doing this would enable prediction for every second on the entire set of data chosen rather than using part of the data for calculating coefficients. Using five hours instead of two or three hours will be very important when implementing profit strategies to increase the potential profit.

Static order book imbalance (SOBI) looks for times during the day when the mean of orders in the buy order book and the mean of orders in the sell order book differ in their distances from the inner prices by a significant margin. Specifically, when the difference between the mean of the sell orders in the order book and the inner sell price is greater than the difference between the mean of the buy orders in the order book and the inner buy price by some threshold value, a price increase is predicted. Likewise, when the difference between the mean of the buy orders in the order book and the inner buy price is greater than the difference between the mean of the sell orders in the order book and the inner sell price by some threshold value, a price decrease is predicted. The method created here will use a measurement of asymmetry between the buy and sell sides of the order book for orders added, traded, and deleted in the most recent time period. Therefore, this method can be thought of as a dynamic order book imbalance (DOBI) rather than a static order
book imbalance (SOBI) as the focus is on changes to the order book.

The measure of asymmetry will be defined as to be oriented with the price change. In order words, a positive value for asymmetry indicates order book conditions favorable to a price increase, and negative value for asymmetry will indicate conditions favorable for a price decrease. One such measurement is

$$a^*_I(t) = x_I(1, t - 1),$$

with the ArcaBook equivalent

$$a^*_A(t) = x_A(1, t - 1).$$

However, this value will generally take on large quantities, so scaling will be necessary. $z_{IB}, z_{IS}, z_{AB}$, and $z_{AS}$ are still defined as the inner buy shares and inner sell shares for Island and ArcaBook. A more manageable asymmetry value is

$$a_I(t) = \frac{a^*_I(t)}{z_{IB} + z_{IS}}.$$

Similarly,

$$a_A(t) = \frac{a^*_A(t)}{z_{AB} + z_{AS}}.$$

To use both Island and ArcaBook,

$$a_B(t) = \frac{a^*_I(t) + a^*_A(t)}{z_{IB} + z_{IS} + z_{AB} + z_{AS}}.$$

36.6% of the observations for $a_I$ have value 0 as a result of the intervals that show no changes in shares. 88.2% of the observations have absolute value below 0.1, so many of the non-zero values are very close to 0. Only 2.3% of the intervals have asymmetry absolute value above 0.5, and only 0.48% of the intervals have an
asymmetry absolute value greater than 1. However, these intervals show the largest change in the order book and are more likely to signal price changes.

As with the previous models, intervals will be deemed significant if they have asymmetry value \( a_I(t) \) for Island and \( a_A(t) \) for ArcaBook) beyond the cutoff value. For cutoff value \( c \), an interval will be significant for a price increase if the asymmetry value is above \( c \) and will be significant for a price decrease if the asymmetry value is below \(-c\). Unlike previous models, there is no probability value generated by a model, but the decision is based on a statistic measured directly from order book activity in the previous second. However, the same diagnostic tools, such as ROC curves and PSIC values, can be used to determine predictive accuracy. (Note that the techniques used in this section are sometimes called “models” although there is no actual fitting of models and determination of coefficients.)

The first step is to predict Island price changes using asymmetry values from Island, ArcaBook, and both. Figure 4.45 shows the plot of the ROC curves when using Island (black), ArcaBook (red), and both (blue). The three curves show the same order as the fixed-time models with both predicting better than only Island, and only ArcaBook predicts slightly worse than the others. However, one obvious difference is the lower area under the curves, indicating overall worse predictive success. The PSIC values are plotted in graph Figure 4.46. The predictive accuracy for the three different asymmetry values appears much different than in the ROC curves, with ArcaBook only predicting the best, followed by both, and then Island only. This difference may be due to the relatively small quantity of intervals for
which the asymmetry measurement take on extreme values. The PSIC plots show the same monotone increasing pattern noted for the fixed-time models.

![ROC curves for asymmetry method](image)

**Figure 4.45**: ROC curves on Island price changes for the asymmetry method using Island (black), ArcaBook (red), and both (blue)

Next, the asymmetry values from Island, ArcaBook, and both can be used to predict price changes on ArcaBook. The ROC curves show the same pattern of success with ArcaBook alone predicting the worst and the combination of both ECNS predicting the best. The PSIC plots show the same pattern as when predicting Island price changes, with ArcaBook predicting best and Island predicting worst. The asymmetry method yields lower PSIC values in general than the fixed-time method.

In general, the asymmetry method results in some successful predictions, but they perform worse than the fixed-time method. However, the asymmetry method does not devote a large amount of time to formulating the model’s coefficients. Instead, the entire day can be spent looking for significant order book changes that
may signal price changes. The next chapter develops strategies that strive to earn profit by making decisions based on the techniques presented here. The asymmetry method may turn out to be more profitable than the fixed-time method by using a larger part of the day to turn a profit despite lower predictive success.

4.9 Model Fitting for Other Stocks and Days

Thus far, almost all of the analyses and model-fitting used data from the Microsoft stock on May 2, 2005. Microsoft is an actively-traded, spread-minimal stock that is commonly used for evaluating the performance of automated trading strategies in competitions, making it the logical choice for developing models. However, Cisco (CSCO) and Yahoo (YHOO) have similar qualities to Microsoft in terms of activity level and spread.

May 2, 2005 is the only day for which data from both Island and ArcaBook is
available, making that day the only option when developing models that use both ECNs. However, data is available from Island for other days, most recently May 23, 2007. Unfortunately, the data is available through around 11:00 AM, omitting much of the trading day. Models can still be developed using this limited amount of data, as well as using the models from May 2, 2005 to predict the price changes in May 23, 2007.

The coefficients of the model resulting from fitting the fixed-time models with three lags for each of Microsoft, Cisco, and Yahoo for May 2, 2005 as well as Microsoft for May 23, 2007 are shown in the following table:

<table>
<thead>
<tr>
<th>$y_t(0, t)$</th>
<th>$\alpha_0$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
<th>$\alpha_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSFT (05)</td>
<td>2.34</td>
<td>$1.10e-4$</td>
<td>$-7.84e-6$</td>
<td>$-3.67e-5$</td>
<td>$-1.46e-5$</td>
<td>$-1.22e-5$</td>
</tr>
<tr>
<td>CSCO (05)</td>
<td>1.90</td>
<td>$1.09e-4$</td>
<td>$5.18e-6$</td>
<td>$-3.12e-5$</td>
<td>$-1.61e-5$</td>
<td>$-1.03e-5$</td>
</tr>
<tr>
<td>YHOO (05)</td>
<td>2.86</td>
<td>$2.61e-4$</td>
<td>$-5.59e-5$</td>
<td>$-1.15e-4$</td>
<td>$9.62e-6$</td>
<td>$-2.94e-5$</td>
</tr>
<tr>
<td>MSFT (07)</td>
<td>2.59</td>
<td>$1.08e-4$</td>
<td>$1.15e-5$</td>
<td>$-2.94e-5$</td>
<td>$-6.78e-6$</td>
<td>$4.60e-5$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$y_t(1, t)$</th>
<th>$\alpha_0$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
<th>$\alpha_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSFT (05)</td>
<td>0.15</td>
<td>$1.12e-4$</td>
<td>$-1.45e-4$</td>
<td>$-5.96e-5$</td>
<td>$-3.13e-5$</td>
<td>$-2.58e-5$</td>
</tr>
<tr>
<td>CSCO (05)</td>
<td>$-2.98$</td>
<td>$1.42e-4$</td>
<td>$-9.21e-5$</td>
<td>$-3.83e-5$</td>
<td>$3.18e-5$</td>
<td>$-3.11e-5$</td>
</tr>
<tr>
<td>YHOO (05)</td>
<td>$-0.16$</td>
<td>$4.27e-4$</td>
<td>$-4.41e-4$</td>
<td>$-2.12e-4$</td>
<td>$3.48e-6$</td>
<td>$6.67e-6$</td>
</tr>
<tr>
<td>MSFT (07)</td>
<td>$-1.34$</td>
<td>$1.19e-4$</td>
<td>$-4.75e-5$</td>
<td>$-5.36e-5$</td>
<td>$-6.87e-5$</td>
<td>$1.20e-5$</td>
</tr>
</tbody>
</table>

Figure 4.47 shows the plots of the ROC curves for the models developed using data for May 2, 2005 from Microsoft (black), Cisco (red), and Yahoo (blue). Cisco appears to yield the best ROC curve, followed my Microsoft and then Yahoo. The
Yahoo model clearly performs the worst of the three stocks, but the ROC curve still lies above the line of chance ($y = x$).

Figure 4.47: ROC curves for Microsoft, Cisco, and Yahoo

Figure 4.48 shows the PSIC values for Microsoft (black), Cisco (red), and Yahoo (blue). The results agree with the ROC curves, with Yahoo performing the worst, and Cisco and Microsoft showing the best results. Unlike the ROC curves, no obvious difference in accuracy exists between Cisco and Microsoft for the PSIC plots, with Microsoft higher for some values and Cisco higher for others.

The models developed for May 2, 2005 use data from 10:00 AM to 12:00 PM to formulate the coefficients and use the data from 12:00 PM to 3:00 PM to test the predictive success of the model. However, the models developed for May 23, 2007 use data from 10:00 AM to 10:30 AM to formulate the model and 10:30 AM to 11:00 AM to test the model. These models use much less time and much less data than the models from May 2, 2005.
Figure 4.48 : PSIC values for Microsoft, Cisco, and Yahoo

Figure 4.49 shows the ROC curve for the model using Microsoft data from May 23, 2007. The ROC curve shows that the model achieves some predictive success, but the results are not nearly as good as the model for May 2, 2005. The lack of success results from the shortage of data available to formulate an accurate model with accurate coefficients. Earlier, it was shown that the coefficients from the model converge after around 100 minutes, so the 30 minutes used for the 2007 models fall well short of the satisfactory quantity.

The lack of data causes an inability to develop an accurate model for May 23, 2007. Instead, to test how well models fitted from older days predict price changes in more recent days, the fitted model from May 2, 2005 is used to predict price changes on the portion of time available for May 23, 2007. Figure 4.50 shows the plots of the PSIC values for the model developed using data from May 2, 2005 to predict changes in May 23, 2007 (black) and the PSIC values for the model
Figure 4.49: ROC curves for Microsoft (May 23, 2007) using half of the May 23, 2007 data to predict the changes on the other half. The prediction is much better when using the model from May 2, 2005. Even though the data used to formulate the model is 2 years old, the larger amount of data used in the formulation of the model and calculation of the coefficients results in more accurate prediction.
Figure 4.50: PSIC values for predicting May 23, 2007 price changes
Chapter 5
Market Strategies

5.1 Intro

Presumably, the main reason people trade stocks on ECNs and other exchanges is to make money. The main tenet for gaining profit in stock trading is the oft-quoted “buy low, sell high.” Another interpretation of this expression is to buy before the price increases and to sell before the price decreases. The timing of when a trader decides to buy or sell the stock may result from any number of reasons, such as intuition, information (legal or illegal), or a predictive model. This chapter will look at the profitability of existing market strategies while creating some new ones based on the models in the previous chapter. As with the development of models in the previous chapter, the trading strategies presented in this chapter will focus on the Microsoft stock, but results for other stocks will follow afterwards.

Stocks are different from most items bought and sold in the world in that a person who does not actually own any shares can sell shares of a stock. Of course, the person must later buy back those shares, but the net result is that a person can gain a profit or incur a loss without having possession of any shares. This action is known as “short selling” or “going short.” Conversely, “going long” signifies the standard process of buying then selling. Regardless of which action is done first, the goal remains to buy at a lower price and sell at a higher price. When discussing market strategies, a positive net value of shares coincides with having bought more
shares than have been sold, and a negative net value means more shares have been sold than bought. The term “position” refers to whether the trader has bought or has sold more shares. A person with a positive position has bought more shares than that person has sold, while a person with a negative position has sold more than bought and must eventually buy back those shares to return to a neutral (0 shares) position.

ECNs are different from traditional quote-driven markets in that ECNs lack a market maker deciding the prices. Instead, stock prices on ECNs are determined entirely by the orders placed by users who wish to buy or sell the stock. As a result, buy and sell orders at the same price are automatically matched, and the bid-ask spread, defined as the inner sell price minus the inner buy price, will always be at least one cent. The semantics of the terms “inner buy” and “inner sell” must be carefully considered. The “inner buy” price is the highest price of any buy (bid) order in the order book, and the “inner sell” price is the lowest price of any sell (ask) order in the order book. However, if a trader wishes to buy the stock, the buy order must be matched with an existing sell order, so the best available price for buying the stock is actually the inner sell price. Likewise, the best available price to sell the stock is the inner buy. Therefore, at any time, the price at which a trader may buy a stock is always larger than the price at which the trader may sell, and the difference is the bid-ask spread.

For spread-minimal stocks, the bid-ask spread is one-cent for at least 99% of the regular trading hours (9:30 AM - 4:00 PM). Due to the large quantities of orders
at and near the inner prices, the prices of these stocks rarely move more than one cent at a time. The only exceptions are for really large orders that empty all of the shares at the inner price and at one cent away from the inner price. The result is that one price change is not sufficient to gain a profit. Consider a stock whose inner buy and sell are $n$ and $n+1$ cents, respectively. Someone can buy the stock for $n+1$ cents and then observe an increase in the inner prices to $n+1$ and $n+2$, respectively. Hence, the stock can then be sold for $n+1$ cents. The net result is no profit or loss (except transaction fees). Therefore, in order to gain profit, a trader must hold the stock for longer than a single price change.

The models developed using the fixed-time method predicted the probability of a price increase or price decrease occurring in the next second. The models developed using the epoch method predicted the probability of the next change being a price increase or decrease, regardless of how soon it occurred. These models rely on information about recent changes in the order book to formulate the probabilities of future events. However, due to the time duration between price changes, the models cannot predict more than one price change into the future. They can provide insight into market dynamics and determine appropriate times to buy, sell, and unwind current holdings.

Many crucial decisions must be made when developing a successful market strategy, including when to buy or sell, how large of holdings to keep at any time, and when to perform unwinding. Profit is achieved when the price at which the stock is sold is higher than the price at which it is bought, regardless of whether the buying
or selling is done first. However, it is important to keep in mind that the inner prices for spread-minimal stocks on ECNs must move two cents in order to earn a one-cent-per-share profit. It may be necessary to “cash out” after one price change and earn 0 profit rather than hold on to the stock too long. If the price moves in one direction then reverts back to the previous price, the sell price is once again one cent below the buy price, and one cent is lost per share.

Another important factor is the number of shares to buy or sell at a time. A trader’s aversion to or preference for risk influences the number of shares that would be appropriate to buy or sell at a time. When testing the strategies on historic data, it is impossible to know how the extra orders and transactions resulting from the strategy would affect the future events on the market. Therefore, any realistic strategy must keep the number of shares involved in the transactions at a reasonably small value. The Sharpe ratio is a useful measuring tool, because it yields the same value no matter how many shares are bought or sold at a time.

5.2 Existing Strategies

Although ECNs are a recent technological development, some strategies have been developed to attempt to gain profit. Many of these strategies were discussed in previous chapters, but this section revisits those and applies them to the data used to develop the price models. Later sections will use the price models for creating new strategies, and those results will be compared to the results for the existing strategies.

Between 2003 and 2005, the Penn-Lehman Automated Trading (PLAT) project
held competitions on its Penn Exchange Simulator (PXS) where it invited people to

test their automated strategies against each other. Several papers have been written

about the papers that competed (and often succeeded) in those competitions. Some

of the strategies did not provide specifics about time scales, quantities of shares,

and other key factors. However, the ECN data can test these strategies with a

variety of possible values chosen for the missing parameters.

The strategies are often presented along with a Sharpe ratio to evaluate their

effectiveness at gaining profit. However, the Sharpe ratio requires several sets of
data for testing the strategy. Therefore, the data available will be split into 30-
minute intervals, and Sharpe ratios will be calculated from these windows of time.

In addition, few of the strategies provide details about the time between consecutive
decisions and observations. For strategies that look at price movements, one minute
will often be chosen as a logical time-scale that is big enough to show price changes
but not too big as to combine them. One minute may seem like a large time-scale,
considering that the models presented in the previous chapter and the strategies
developed from them later in this chapter use a time scale of one second. However,
the models use changes in shares on the order book, which happen at a much greater
rate than changes in the inner prices. A one-second time-scale for price movements
would result in a vast majority of values equaling 0.

According to the Efficient Market Hypothesis (EMH), it is impossible for a
strategy to perform better than the market as a whole. It also states that there
is no pattern of behavior in the market that can be exploited for consistently out-
performing the market [SRSK06]. However, many strategies have been published that perform a variety of techniques to produce profitable outcomes. The section describes several strategies that have been designed for use on ECNs, with focus on those that are automated to buy and sell under certain conditions.

Obviously, the long-standing tenet for profiting from the stock market is to buy when the price is low and to sell when the price is high. If a trader expects the price to increase, he or she can purchase shares now and sell them later if the price did indeed rise. However, what can a trader do if he or she expects the price to drop? The main way to profit from that situation is “short selling.”

Short selling is a strategy to make money when you think a stock is about to drop in price. A trader can “sell” shares of a stock that he does not actually own, under the condition that he or she will buy the shares back at a later time, hopefully at a lower price than the price at which it was sold. This way, a user can sell shares of a stock he or she expects will drop, then buy it a short time later at a lower price, with the shares going to the buyer. In this situation, the short seller makes a profit without ever actually owning the stock. There are large risks associated with short selling, usually resulting from the price increasing and being forced to buy shares at a higher price than they were sold. Most exchanges require all outstanding debts to be settled at the end of the day, forcing a short-seller to possibly be stuck buying back at a higher price and suffering a loss. Short selling is an important concept for market strategies presented in this section as well as those developed later by allowing a trader to profit when anticipating a price drop.
Before models for price prediction can be created and developed into potentially profitable market strategies, it is necessary to examine strategies that already exist. Very few of these strategies attempt to predict future price changes in order to gain profit. Instead, they use order book imbalances, price trends, or market volatility to achieve a profitable market strategy. The papers do not only present their strategies but also explain how the profitability of the strategies were tested and evaluated. Most strategies were tested on the Penn-Lehman Automated Trading (PLAT) project’s Penn Exchange Simulator (PXS). The returns yielded when the strategies were tested on the PXS were evaluated using the Sharpe ratio. The “returns” defined here are not returns in the formal definition but are used as the amount of profit or loss shown in a given time interval using the strategy. The Sharpe ratio is defined as the mean of returns $R$ divided by the standard deviation of returns.

$$S_{Sharpe} = \frac{\mathbb{E}[R]}{\sqrt{Var[R]}}$$

A less common variation on the Sharpe ratio is the Sortino ratio, which replaces the standard deviation of returns in the denominator with the standard deviation of only the negative returns.

$$S_{Sortino} = \frac{\mathbb{E}[R]}{\sqrt{Var[R^+]}}$$

[SRSK06]

The Penn Exchange Simulator (PXS) [KO03] uses both real orders from the Island ECN and orders from the designers’ algorithm to create a realistic stock market simulator where traders can test their market strategies. PXS is a major
development from the Penn-Lehman Automated Trading (PLAT) project. PXS enables users to add their own orders to the actual orders that were in the order book at a given time either in the past (historical mode) or the present time (live mode) [SR05].

Several techniques using a variety of strategies have been developed for gaining profit on ECNs. Many of these strategies competed in the PLAT Competition and provide the results of how their strategy performed. Most of the strategies presented in published papers performed well in the competition, as the less successful strategies probably required adjustments before they could be implemented.

The PLAT competitions encourage strategies that yield positive returns but also minimize risk. Shares that are bought or sold must be liquidated by the end of the trading day, so having a large positive or negative quantity of shares may be unwise due to the need to perform transactions before the end of the day to return back to 0 net shares. This action of reverting back to 0 net shares is called unwinding or "cashing out" [KO03]. The need for unwinding wisely is crucial to developing a successful strategy.

Kearns and Ortiz [KO03] discuss the competitions held on the PXS. The first two competitions required that participants remain within 100,000 shares, which the organizers considered to be "generous," but did not impose any other restrictions. Cumulative profit accumulated during the competition was the sole factor in awarding victory for the first two competitions. The Platinum Platter Competition (PPC 2003), the third competition held on the PXS, used a much more intricate
system of criteria to evaluate performance in the competition [KO03].

Silaghi and Robu used the PXS to test how existing market strategies fare when used alone or in combination with other strategies. These strategies are (i) static order book imbalance (SOBI), (ii) volume average weighted prices (VWAP), (iii) trend following (TF), and (iv) reverse strategy (RS). SOBI computes two volume-weighted averages of the prices contained in the buy and sell order books. If the sell-side difference is larger than the buy-side difference by a certain cutoff amount ($\theta$ dollars), the belief is that prices will rise and a buy order is placed. Likewise, if the buy-side difference is larger than the sell-side difference by at least $\theta$ dollars, the hypothesis is that prices will fall and a sell order is placed [SR05]. The results for the SOBI strategy are provided to the participants of the PXS competitions as a benchmark to compare their own strategies.

Among the strategies presented here, the SOBI strategy is most similar to those that will be developed later in this thesis. SOBI uses the order book as the sole factor in evaluating the current market conditions and in deciding whether to buy, sell, or do nothing. However, it is important to notice the static nature of SOBI (which is what the S stands for in SOBI) that looks at a snapshot of the order book at a given time. The models presented later will focus on changes in the order book, giving those models a dynamic element. In fact, one of the models performs similar functions to SOBI except order book changes replace static order book imbalances.

The SOBI strategy as presented by Silaghi/Robu [SR05] causes some concerns. It is likely that the occurrences where the order book imbalance is greater than
some threshold value will be clustered together, resulting in a likely large positive or negative position. If there are no limits set on the size of the positive or negative position that may be taken or on the number of trades that may occur within a window of time, then there is a greater risk for a large loss if the price moves in the direction opposite of the expected direction. Common sense dictates that strategies that capitalize on predicting price changes on a small time scale would benefit from limiting the size of the shares positions obtained and from performing regular unwinding. Unwinding regularly limits the possibility for large losses and quickly capitalizes on predicted changes.

The Volume Weighted Average Price (VWAP) strategy looks at weighted averages by volume. If the average price for first orders in the buy order book is higher than market average, the trader places a sell order, and vice versa. Trend Following uses linear regression with respect to time to create 2 trend lines from last “ticker” prices from 2 different time windows. The slope of lines can be used to determine the direction of the price development. Last, the reverse strategy says the trader should sell when price is rising and buy when price is falling. Although it sounds counterintuitive to the standard “buy low, sell high” paradigm, this strategy utilizes micro-movements and price corrections [SR05].

Silaghi and Robu [SR05] look at these four existing strategies as well as combinations of the strategies and test them using historical data on the PXS market simulator. They employ a two-thirds majority rule, so if at least 2 out of the 3 methods show a significant result and if the results suggest placing a buy or sell
order, then an order is placed there. Using the simulated data, the most profitable strategy was a two-thirds majority rule combining SOBI, reverse, and TF. Another measure of success is the empirical daily average of returns divided by standard deviation, also known as the Sharpe ratio. The best Sharpe ratio resulted from using the reverse strategy by itself [SR05]. The strategies discussed here will return later when models are formulated and the success of the strategies are evaluated.

Silaghi and Robu [SR05] use existing strategies by themselves and in combination with the other strategies to determine which returns the best profit. The four strategies they use are (1) static order book imbalance (SOBI), (2) volume average weighed prices (VWAP), (3) trend following (TF), and (4) reverse strategy. When used alone, SOBI yields a Sharpe ratio of 0.2, VWAP yields a Sharpe ratio of -0.823, TF yields a Sharpe ratio of 0.14, and reverse strategy yields a Sharpe ratio of 0.92. When combining all four strategies such that the agent only acts when all four agree, the Sharpe ratio is 0.3797. Using a combination where SOBI and reverse agree yields a Sharpe ratio of 0.261, and a combination of SOBI, reverse, and TF such that two out of three of the strategies must agree yields a Sharpe ratio of 0.65.

Out of all of the strategies and combinations of strategies tested by Silaghi and Robu [SR05], the best Sharpe ratio resulted from using the reverse strategy by itself. However, the trend-following and reverse strategies are the same techniques presented by Feng et al. [FYS04]. The reverse strategy showed negative results and a negative Sharpe ratio when applied to the ECN data. This disparity would
indicate that results are not reproducible and may be the result of good luck rather than an effective strategy.

The results for testing the SOBI strategy on Island data is shown in Figure 5.1, and the results for testing the VWAP strategy is shown in Figure 5.2. When testing SOBI for different cutoff values, the Sharpe ratios show a generally increasing pattern, starting out negative for smaller cutoff values and becoming positive for larger cutoff values. The total profit increases then decreases, with the maximum reached around cutoff value 0.008. The SOBI strategy tested here is not a true SOBI strategy, as extreme values have been removed to prevent outliers from influencing the results. VWAP yields consistent negative Sharpe ratios for the different cutoff values, while the total profit shows a consistent upward trend as the cutoff values increase. This trend results from less trading due to fewer significant values for larger cutoff values, resulting in smaller negative returns.

![Figure 5.1: Sharpe Ratios and Total Profits when testing the SOBI strategy](image-url)
Figure 5.2: Sharpe Ratios and Total Profits when testing the VWAP strategy

Savani and Veal [SV05] describe their market strategy, which won the 2005 PLAT competition. Their Sharpe ratio in the competition was 3.87, easily defeated their competitors, as the next highest was 1.33. Their ingenious strategy, called “Jump and Dump,” exploited a loophole in the competition framework. Unlike in previous years, the 2005 competition did not include real stock market data, but instead relied solely on automated “agents.” These agents were programmed to place buy and sell orders distributed around the inner prices. The Jump & Dump strategy bought all shares in the sell order book, essentially cornering the market. Then, the Jump & Dump system would place buy and sell orders for one share each at some price larger than where the shares were bought. Due to the fact that there were no other orders in the sell order book other than the one just added, the automated agents believed this new value to be the true price and added orders around the new price. The Jump & Dump strategy then sells its shares to those
added orders at the artificially high price. The agents and the other competitors are all pre-programmed, so once the competition started, there was no way for other competitors to change their programs to counteract the Jump & Dump strategy.

Although the Jump & Dump strategy performed very well in the 2005 competition and is a clever strategy, it is impractical to attempt such an endeavor in real life. Real traders cannot be fooled into adding orders at an artificially created value like the automated agents were. Generally, the shares being traded at any time on an ECN represent a small percentage of the total shares in existence, so there is no practical way to corner the market and drive the price up. Also, there are huge risks associated with buying so many shares on a stock. Therefore, the Jump & Dump strategy displays an important lesson: Just because a strategy works on a simulator does not mean it is practical and can be used in real life.

Sherstov and Stone [SS05] present strategies for automatically trading on ECNs and describe their results on the PXS simulator in the PLAT competition. They also compare the results of their strategies to the result of the Static Order-Book Imbalance (SOBI) strategy, which the PLAT competition provides to its competitors as a way of comparing their strategies' relative profitability. The competitions held in December 2003 and April 2004 ranked its participants by their Sharpe ratio, which is the mean of returns divided by standard deviation. Using the Sharpe ratio instead of total returns places emphasis on consistency between days [SS05].

Sherstov and Stone [SS05] describe the details of a Trend Following (TF) Agent. This method calculates the rate of change of the prices ($P'$) and the rate of change
of the rate of change \((P^\prime)\) of the prices. These values are not explicit first and second derivatives but can be thought of as an analog for these derivatives. When \(P' > 0\) and \(P'' > 0\), both the price and the rate of increase are increasing, so the TF Agent buys 75 shares at the inner price in anticipation of further price increases. Likewise, when \(P' < 0\) and \(P'' < 0\), the price is decreasing at an increasing rate, so the TF Agent sells 75 shares at the inner price in anticipation of further price decreases. The strategy performs unwinding whenever the signs of \(P'\) and \(P''\) do not match, preventing the accumulation of large amounts of shares in either the positive or negative direction [SS05].

Sherstov and Stone [SS05] describe two of their strategies: the trend-following agent and the market-making agent. The trend-following agent, which buys when the price is increasing at an increasing rate and which sells when the price is decreasing at a decreasing rate, yielded a Sharpe ratio of -0.4573. The market-making agent, which buys and sells under similar conditions to trend-following but also places a corresponding order on the opposite side, yielded a Sharpe ratio of 0.229. The SOBI strategy was used as a benchmark strategy, and SOBI yielded a Sharpe ratio of 0.7014 [SS05].

Figure 5.3 shows the Sharpe ratios when testing Sherstov and Stone's [SS05] trend-following strategy for different time scales. None of the time scales results in a positive Sharpe ratio. However, the paper reported a negative Sharpe ratio, and the results when using the ECN data agree with the lack of profitability when using this strategy.
The trend-following strategy performs very poorly, but Sherstov and Stone [SS05] achieved better results for their market-making strategy. This strategy requires a time scale and profit margin. To test the profitability of trend-following, the time-scale will be set to one minute, and the profit margin will vary. A smaller profit margin would reduce the profit made on every trade, but a larger profit margin leaves a large risk of having orders unmatched and incurring losses from unwinding at a significantly worse price. Figure 5.4 shows the Sharpe ratios and total profits for different profit margins. None of the profit margins yield positive results, although a profit margin of five cents almost breaks even. The positive Sharpe ratio claimed by Sherstov and Stone was not reproduced when using the ECN data.

The next strategy is the Market-Making (MM) Agent, which modifies the TF Agent to gain profit from price fluctuations, not trends. As with the TF Agent, the MM Agent calculates $P'$ and $P''$ and looks for instances when their signs match.
Figure 5.4: Sharpe Ratios and Total Profits when testing Sherstov/Stone’s Market Making strategy

When $P' > 0$ and $P'' > 0$, the MM Agent still buys 75 shares at the inner buy price. However, this time a sell order for 75 shares is simultaneously placed at a fixed distance (called the profit margin) from where the shares were bought. If the shares were bought at price $B$, then the sell order would be placed at price $B + PM$. Likewise, when $P' < 0$ and $P'' < 0$, the agent sells 75 shares at the inner sell (call it $S$) and places a buy order for 75 shares at $S - PM$ [SS05].

The different strategies are tested on ten days in November 2003, classified by the daily price behavior as "monotone" (M), "substantial fluctuation" (F), "zigzag behavior" (Z), and "mixed (other)" (O). The TF and MM Agents yielded Sharpe ratios of -0.457 and 0.229, respectively. SOBI outperformed the MM Agent with a Sharpe ratio of 0.701. In the competition, the strategies placed orders of size 15 shares instead of 75 shares and used smaller profit margins to achieve safer, more consistent profits and to avoid share imbalances. The MM strategy won the PLAT
Ramamoorthy [Ram03] introduces a strategy based on moving averages that takes a safe or risky approach based on current amounts of cash and stock in possession of the user. His goal is to maximize the Sharpe ratio while considering fees and possible unwinding penalties. The returns (R) are defined by the equation
\[ R = \text{Cash} - \text{Unwinding Penalty} - \text{Trading Fees} + \text{Trading Rebates}. \]
The unwinding penalty is assessed by valuing any remaining positive shares at $0 and requiring that any negative shares be bought back at double the cost to buy back shares at the end of the trading day. These penalties greatly encourage unwinding and discourage keeping large positive or negative amounts of shares at any time. The trading fees and rebates are charges of $0.003 per share for removing liquidity via a trade and $0.002 per share for the order removed from the order book via a trade [Ram03]. These fees and rebates exist to encourage liquidity on ECNs by giving money back to the trader who added the order to the order book while simultaneously charging the trader whose order caused the trade and removed the shares from the order book.

Ramamoorthy [Ram03] explains the algorithm by which his strategy is implemented as well as several variations. The basic algorithm calculates values MA₁ and MA₂, which are moving averages of the price of the stock over the past time interval of length N₁ and N₂ with N₂ > N₁. For some threshold value (call it k), if MA₁ > MA₂ + k, then the algorithm sells m shares at the best price available (the inner buy price). Likewise, if MA₂ > MA₁ + k, then the algorithm buys m shares.
at the inner sell price.

The first modification to the algorithm takes unwinding into account by performing the actions based on the moving averages only if time is before a certain cutoff. If the time is after the cutoff and the share holdings are positive, then the modified algorithm sells \( m \) shares, and if the time is after the cutoff and the share holdings are negative, then the modified algorithm buys \( m \) shares. The original algorithm bought and sold shares \( m \) at a time, so \( m \) must be a fraction of the total share holdings at any time [Ram03].

Ramamoorthy [Ram03] used moving averages of the prices to determine trends, selling when those moving averages were increasing and buying when the moving averages were decreasing. There are two time parameters that are not specified which represent the lengths of time to measure the movement of the prices. Their results claim a Sharpe ratio of 0.0647 for the basic strategy, which then rises to 0.5432 when they introduce a genetic element that decided whether to be safe, regular, or risk-taking. Their work does not mention the time scale used for the time between price observations. Figure 5.5 shows the Sharpe ratios that result from testing Ramamoorthy's MA strategy on the ECN data for various time scales. None of the time scales shows a positive Sharpe ratio, although there is an obvious improvement in Sharpe ratio as the time scale increase. However, this improvement is due to the detection of fewer significant times to buy or sell and not any improvement in the model. The lack of details provided by Ramamoorthy regarding time scale hinders replication of his results, although none of the chosen values yielded
results similar to his results.

![Sharpe ratios for Ramamoorthy's MA](image)

Figure 5.5: Sharpe ratios when testing Ramamoorthy’s Moving Average strategy

Ramamoorthy contributed to the work by Subramanian et al. [SRSK06] that develops “evolutionary algorithms” to select an optimal market strategy dependent on the user’s holdings of cash and stock. They claim that “technical indicators” yield greater profit when used in conjunction with other trading rules than by themselves. These “genetic algorithms” and “genetic programs” are tested on the PLAT simulators, because testing them on real markets would be too risky. On the PLAT simulator, their strategy is tested using data from Microsoft (MSFT) stock. Sixty days worth of data was used, some of which were used for fitting the model and others for testing the model.

Subramanian et al. [SRSK06] define “agents” and “strategies” and how they differ. An agent is the program that places orders based on the programmer’s wishes. On the other hand, the strategies make decisions based on an algorithm
and on current market conditions. Basically, the strategy is the course of action to be taken, while the agent is the person or the program carrying out that action [SRSK06].

The genetic algorithm places weights on other algorithms to decide whether to buy, sell, or do nothing. The genetic algorithm is risk averse when there is a large positive or negative position and is risk seeking/neutral for smaller positions. The other algorithms used in the genetic algorithm include the moving averages strategy presented by Ramamoorthy [Ram03], a volume-based strategy similar to VWAP, the “basic” strategy (opposite of “reverse” strategy), and the price channel breakout strategy. The price channel breakout strategy uses Bollinger bands, which create maxima and minima to enclose the price. Define the upper price band limit (UPBL) as the maximum of the price over the last n ticks and the lower price band limit (LPBL) as the minimum of the price over the last n ticks. If the current price is greater than the UPBL, then buy. If the current price is lower than the LPBL, then sell. This breakout strategy only performs a transaction when the price “breaks out” of the “price channel” formed by the Bollinger bands [SRSK06].

The genetic algorithm results in a Sharpe ratio of 0.701, while the genetic program yields a Sharpe ratio of 0.285 [SRSK06]. Subramanian explains how these genetic algorithms and genetic programs yield consistent profits by designing safe agents that avoid decisions that may potentially be unsafe [Sub04].

Feng et al. [FYS04] describe an order book oriented market-making strategy that places orders in pairs to capitalize on market fluctuations. If the inner buy
and inner sell prices are \( x \) and \( y \), respectively, then the trader places a buy order at \( x-c \) and place a sell order for the same number of shares at \( y+c \). If both orders get executed, then profit is earned. Their primary strategy uses "fixed gaps," meaning that the profit margins on the buy and sell sides are equal. A variation on this strategy, called the Price Control Method, prevents extreme positions by lowering the sell price to encourage selling when the number of shares currently being held is positive. Likewise, when the number of shares currently being held is negative, the buy price is raised to encourage buying back the shares [FYS04].

One important decision involves the size of the profit margin \( c \). If the trader acts cautiously and picks \( c \) to be a small value, then the profit made on each pair of transactions is also small. However, if the trader gets greedy and chooses a large \( c \), then larger fluctuations are required in order for both orders to be traded, resulting in a higher probability that at least one of the orders does not get executed [FYS04]. The profitability of the strategy is contingent on the execution of both orders. If the pair of orders is placed and then the price shifts upward or downward, one order may be traded while the other just sits in the order book as the price moves away. The trader may then have to unwind the shares of the order that was executed and take a large loss. It would take several successfully executed pairs of orders to recoup the losses from one of these major losses. This possibility for large losses makes this strategy very risky when using a large profit margin.

Feng et al. [FYS04] also present a reverse strategy, also called "contrarian", that buys shares when the price is dropping and that sells when the price is rising
over a preceding fixed time interval. Their reverse strategy contrasts to what they call the “basic strategy” that buys when the price is rising and that sells when the price is dropping. Both the market-making strategy and the reverse strategy were entered in the PLAT competition. Feng et al. [FYS04] extol the benefits of testing their strategies on the PLAT simulator without having to deal with the risks associated with testing an unproven strategy using real money. They outline a few of the drawbacks of using the PLAT simulator, such as the inability to adjust one’s strategy once the simulation has begun and the time lag behind the real market that could theoretically provide an advantage to those with access to real-time data. Nevertheless, they seem satisfied with the results of the competition. Both the market-making strategy and the reverse strategy advanced from their respective pools, and the reverse strategy won the competition despite losing money in the final round. The reverse strategy emerged victorious as it lost less money than its competitor in the final match. They explain the poor performances of the strategies in the final round as the result of “uncommon price patterns” [FYS04].

The reverse or contrarian strategy presented by Feng et al. won the PLAT competition, even though the it yielded negative results in the final round. A strategy that performs no trading whatsoever and returns $0 profit would have outperformed the two finalists over the five-day period used in the final round of the competition. The reverse strategy that won was one of two strategies entered by Feng et al. in the competition, and other papers have described as many as three strategies entered into a single PLAT competition. These multiple entrants raise the
concern of whether the success experienced by some of the strategies results from a truly successful strategy or from a random chance. This chance of having at least one successful strategy increases with more strategies entered in the competition. Traders should take caution before proceeding with a strategy that shows highly inconsistent results.

Feng, Yu, and Stone [FYS04] describe a market-making strategy, a "basic" strategy, and a "reverse" strategy. The market-making strategy places orders in pairs at distances of profit margin \( c \) from the inner price. This strategy is similar to the market-making strategy of Sherstov and Stone [SS05] but here there are two orders that need to be executed instead of only one. As a result, using the same value for the profit margin would result in making greater profit but also increases the chance that the orders are not executed. If one order is executed but the other is not, the trader would have to unwind those shares at a less than ideal price and may incur potentially large losses.

Figure 5.6 shows the Sharpe ratios and total profit for the market-making strategy presented by Feng et al. A few of the larger profit margins yield positive returns, while the lower ones yield negative returns. The smaller profit margins do not earn enough profit from the orders that are executed in pairs to make up for the losses resulting from unwinding the orders that are not matched. The larger profit margins result in fewer pairs of orders executed, but the profits made dominate the unwinding losses.

This market-making strategy requires sizeable price fluctuations to turn a profit.
The previous market-making strategy presented by Sherstov and Stone [SS05] also required price movement, although only one order needed to be executed as a result. Days that show a lot of price movement will generally result in lots of matching of orders and larger profits. Days that show a more monotone pattern of price movement may result in devastating losses for these strategies, so one should proceed with caution when implementing such strategies.

The basic strategy and reverse strategy use the same general premises but apply them in different ways. When the price is increasing, the basic strategy buys shares under the conclusion that the price will continue to increase, while the reverse strategy anticipates a price correction and sells shares. Likewise, when the price is decreasing, the basic strategy sells, and the reverse strategy buys. According to the results presented by Feng et al. [FYS04], the reverse strategy outperformed the basic strategy on its way to winning the PLAT competition despite showing
negative overall returns. When the strategies are tested on the ECN data, the basic strategy yields a Sharpe ratio of -1.769, and the reverse strategy yields a Sharpe ratio of -0.946. These results agree with the published results, with the reverse strategy performing better than the basic strategy despite of negative results.

Sohn et al. [SBK05] define the Relative Strength Index (RSI) as a value between 0 and 100 that measures the strength of a stock compared to its past performance. A high RSI indicates that the average of price increases are greater than the average of price decreases. When the RSI is high, the stock is beating past performance, and the trader should buy. When the RSI is low, the stock is doing worse than past performance, and the trader should sell. However, they explain that RSI-based strategies are generally less successful than other strategies and that reverse strategies tend to be successful [SBK05].

5.3 Setting Up Model-Based Strategies

When formulating a trading strategy for an ECN based on a market model, several decisions must be made before proceeding. The first decision is what model and variables to use and how to implement the model wisely. The previous chapter introduced a variety of methods for creating models and a number of variations that could be introduced to the models. The fixed-time and epoch methods yielded probabilities for the different price change outcomes. One option for a profit-making strategy is to observe the model output, and if it yields a probability greater than some pre-determined cutoff value, then the strategy buys or sells accordingly. A significant probability of a price increase would be a signal to buy, and a significant
probability of a price decrease would be a signal to sell.

Although the models yield accurate probabilities for price changes, it is necessary to use some of the data (and thus some of the day) to determine the coefficients for the models. As such, some of the day is spent towards building the model and not spent trading. Therefore, it may be advantageous to use the asymmetry values to decide when to buy or sell. As with the fixed-time and epoch models, the asymmetry values measure the underlying dynamics of the order book to determine the probable direction of price movement. Unlike those models, the asymmetry values do not require coefficient evaluation and can be used immediately.

The use of asymmetry values to decide whether to buy or sell is similar to the SOBI strategy described by Silaghi and Robu [SR05]. The SOBI strategy uses the average prices on the buy and sell sides to determine order book imbalance to infer price change directions. The asymmetry values also calculate an imbalance in the order book, but instead use differences in the order book from one time to the next. Instead of being “static” (which is what the S in SOBI stands for), the asymmetry values are dynamic.

A trader on an ECN earns profit from a successful combination of buying and selling, which means knowing the best time to unwind or “cash out” is crucial to success. Therefore, the decision of when to perform unwinding is an important factor when developing a market strategy. Some strategies wait until shortly before the end of the trading day and perform unwinding all at once. This plan can be very risky, especially if left with a large positive or negative holding when it is time
to unwind. A safer strategy would involve picking times during the day at which to unwind, either at regular intervals or based on the models.

Rather than unwinding completely by liquidating the entire position at once, it may be advantageous to liquidate orders one at a time. One option is to hold onto a position for two epochs and then liquidate. The models only predict one change into the future, and the time between changes is generally long enough to create difficulties in predicting multiple changes in the future. Another option holds onto the position for a fixed amount of time instead of a fixed number of epochs. Again, this plan experiences the conflict between holding onto a position for a while to gain profit and the inability to predict more than a little time into the future.

Instead of holding onto a position for a fixed amount of time or a fixed number of epochs, the same model that decides the appropriate times to buy or sell the stock can be used to decide when to unwind. The shares bought when the probabilities from the models’ output were significant for a price increase should be sold when the probabilities from the models’ output are significant for a price decrease and vice versa. If the trader trusts the model to make the decisions about when to begin ownership of the stock, then logically the same model should be trusted to decide when to end ownership of the stock.

Some of the strategies described before place a limit on the size of the position (quantity of shares held at a time), while others place no such limit. This position could refer to either long (positive) or short (negative) shares held before unwinding. Obviously, placing limits on the size of the position minimizes the amount of money
that could be potentially gained or lost while trading.

5.4 Strategies Using the Models

The first strategy buys and sells based on the predicted probabilities from the fixed-time model output. The model uses Island price changes for the dependent variable and uses one lag of inner shares changes from Island, ArcaBook, or both. No limit is placed on the size of the position, and unwinding is only performed at the end of the time interval. The values for the Sharpe ratio and total profit at various cutoff values are shown in Figure 5.7. Again, the black graph uses Island only, the red graph uses ArcaBook only, and the blue uses both. With the exception of one observation using only Island, the Sharpe ratio is negative for each cutoff value, and each cutoff value shows a negative total profit. The one observation that shows a positive return shows a very small positive return. The amount of money lost decreases as the cutoff value rises due to fewer intervals being tagged as significant for buying or selling. Figure 5.8 shows the results when the strategy is applied to ArcaBook price changes. Trading on ArcaBook yields similarly poor results and negative returns with only a couple of exceptions that yield slightly positive results. Therefore, allowing unlimited buying and selling of the stock during the day does not succeed like a successful strategy should, and modifications must be made to this strategy.

The next strategy follows a similar paradigm of buying or selling without limit based on the recommendation of the model. However, this strategy unwinds every time the decision based on the model switches to the other side. As long as the
Figure 5.7: Sharpe ratios and total profits using unlimited buying and selling on Island

Figure 5.8: Sharpe ratios and total profits using unlimited buying and selling on ArcaBook
significant probabilities all fall on the same side, the trader keeps buying or selling as per the model. As soon as the opposite side shows a significant probability, the model has detected a change in the dynamics of the order book and the entire position is unwound, regardless of the size of the position. This unwinding maximizes the profit by cashing out before the price reverts back to a worse price.

Figure 5.9 shows the plot of Sharpe ratios and total profit for different cutoff values using the same model for predicting Island price changes. The Sharpe ratio remains consistently between 0.3 and 0.7 for all cutoff values between 0.05 and 0.2. However, the total profit shows a sharp drop between cutoff values 0.05 and 0.1, showing a much slower drop between 0.1 and 0.2. This pattern results from the much greater number of orders being traded for the lower cutoff values. The consistency of the Sharpe ratios despite the large drop in overall profit demonstrates the large variability among the resulting profit for the lower cutoff values. Trading based on the larger cutoff values is much less active, resulting in lower but more consistent profits. This strategy has shown to be very successful based on the overall positive values for the profits and Sharpe ratios. The model that uses data from both Island and ArcaBook shows lower total profits than the model that uses only data from Island but shows very consistent Sharpe ratios. The model that uses only data from ArcaBook yields negative Sharpe ratios for the lower cutoff values but yields large positive Sharpe ratios for the higher cutoff values.

The results for this strategy when applied to ArcaBook are shown in graph Figure 5.10. The Sharpe ratios are similar for the larger cutoff values, but the
smaller cutoff values yield negative results. The total profit starts out negative then shows a sharp increase to positive values, which then steadily decreases while remaining positive. The total profit decline agrees with the Island results due to the selective nature of the higher cutoff values.

Throughout this chapter, one constant theme has been the need for multiple price changes to gain a profit instead of just one. The next strategy will again buy and sell based on the significance of the probabilities generated from the model. However, this time the stock will be “held” until two price changes occur and will then be unwound. This strategy does not change based on whether the price changes were in the same or opposite directions or on whether the predicted price change directions were correct or incorrect.

Figure 5.11 shows the Sharpe ratios and total profit values for the strategy of holding for two price changes on Island. The Sharpe ratios fluctuate between -0.5
Figure 5.10: Sharpe ratios and total profits using the unlimited strategy with unwinding on ArcaBook

and 0, while the total profit increases from larger to smaller negative values as the cutoff value increases. However, no cutoff value yields a non-negative profit value. ArcaBook yields even worse results, with Sharpe ratios between -2 and -0.5 for all positive cutoff values and a similar pattern of negative total profit.

Why does this strategy always yield negative returns although the predictive model is highly accurate? The problem arises with the variability of the second price change. If the model is correct with probability $p$ and a price change is equally likely to be an increase or decrease, then a profit of one cent occurs when the first change is correctly predicted, and the second one follows the same direction as the first. This series of events has probability $p^2$. Equally likely is the event that the first change is correctly predicted while the second goes in the opposite direction, which brings the price to their original values and results in a one cent loss. An incorrect prediction followed by a change in the opposite direction has probability...
Figure 5.11: Sharpe ratios and total profits when holding for 2 epochs on Island

\[ \frac{1-p}{2} \] and yields a one cent loss, and an incorrect prediction followed by a change in the same direction has the same probability but yields a three cent loss. The expected value of the resulting profit \( R \) is

\[ \mathbb{E}[R] = \left(1 - \frac{p}{2}\right) - 1 \cdot \frac{1-p}{2} - 3 \cdot \frac{1-p}{2} = -2(1 - p) = 2p - 2. \]

Even if the model had 100% at predicting price changes, the uncertainty and randomness of the second price change makes the expected profit equal to 0. These models have less than 100% accuracy, so the expected profit will always be negative. If the trader wanted to hold the stock for a fixed number of epochs greater than 2, the expected profit would remain unchanged, as no prediction is being made about the future change and each subsequent change is equally likely to be an increase or decrease.

Next, the orders bought and sold will be held for a fixed amount of time instead of a fixed number of price changes before unwinding is performed. Figure 5.12
shows the plot of Sharpe ratios versus holding time (in minutes) for Island when the cutoff value is 0.05 (black), 0.1, (red), and 0.15 (blue). Holding the stock for short amounts of time yield negative returns, but holding the stock for several minutes gives positive profit. The most successful cutoff value was 0.1, which shows a profit for every time duration greater than 2 minutes and reaches a maximum Sharpe ratio of around 0.5 at holding time 6.5 minutes. When applying this strategy to ArcaBook, the results are much less successful. The resulting Sharpe ratios are mostly negative, and those that are positive are close to 0.

Figure 5.12: Sharpe ratios when holding for a fixed amount of time on Island

For Island, cutoff value 0.1 yielded the best positive results when holding for a fixed amount of time between two and ten minutes. When trading on ArcaBook, as shown in Figure 5.13, only cutoff value 0.05 (black) returns positive results for all holding times between two and ten minutes. Using cutoff value 0.10 (red) or 0.15 (blue) yields inconsistent results when holding for various times between two
and ten minutes. The models use information in the recent past to predict events in the immediate future, so using these models to determine when to buy or sell and then holding the shares for a fixed amount of time gives inconsistent results.

![Sharpe ratios when holding orders for a fixed time on ArcaBook](image)

Figure 5.13: Sharpe ratios when holding for a fixed amount of time on ArcaBook

The strategy of holding a stock for a fixed amount of time yielded much different results for Island as for ArcaBook. This difference in results indicates the strategy is unreliable and may have resulted from lucky price movements on Island. In fact, 4 out of 6 intervals yielded negative returns for Island, but the other 2 gave much larger positive results that brought the overall return to a positive number. The inconsistent results from this strategy, combined with the results from the previous strategies, proves the need for using the model to not only decide when to buy and sell but also when to unwind the resulting position.
5.4.1 Limited Position

Returning to the strategy of unwinding when the model returns a significant probability in the opposite direction, a limit is now imposed on the size of the position that can be held at any time. If the probability of price increase is significant but the trader is already holding \( n \) shares, then the trader ignores the model and does not buy another. Likewise, the trader will not sell if a significant probability of price decrease occurs while at a position of \(-n\). The plots of Sharpe ratios and total profit for Island and ArcaBook are shown in Figure 5.14 and Figure 5.15. These plots use a maximum position \( n \) equal to 1 (brown), 3 (purple), and 5 (green).

![Sharpe ratios for limited orders per epoch on Island](image)

![Total profit for limited orders per epoch on Island](image)

Figure 5.14: Sharpe ratios and total profits for limited position on Island

The Sharpe ratio and total profit results for the unlimited and limited models are similar to each other for the larger cutoff values. This similarity is due to the lower quantity of intervals deemed significant when using higher cutoff values, so the limit is not often reached. However, for the lower cutoff values, the total profits...
Figure 5.15: Sharpe ratios and total profits for limited position on ArcaBook are lower but the Sharpe ratios are higher when placing limits on the size of the position taken. The lower cutoff values result in much more trading when no limits are placed on the amount of trading to be done, but with these larger profits comes greater variance. The smaller the limit that is placed on the position, the smaller the total profits. Placing limits on the position reduces the potential profit but also limits potential risk while yielding more consistent returns.

Thus far, the profit strategies have used only the fixed-time method’s results to decide the appropriate times to buy, sell, and unwind. However, these same strategies can be implemented using the epoch method and the asymmetry method, as described in the previous chapter. The epoch method requires similar setup to the fixed-time method in requiring a portion of the time at the beginning of the day to establish the coefficients before transactions based on the models can occur. The asymmetry method does not require this parameterization and can be used for
transactions at an earlier time.

Returning to the strategy that allows for unlimited buying and selling while not unwinding until the end of the day, the Sharpe ratios and total profits for different cutoff values are shown in Figure 5.16 for the epoch method and Figure 5.17 for the asymmetry method. Notice the cutoff values in the x-coordinate are larger than for the fixed-time method. The fixed-time method yielded probabilities of price changes in the next interval, which rarely went higher than 0.2 for price increase and price decrease. However, the epoch method gives probabilities of the next change, so larger cutoff values are chosen. Likewise, the measurement of asymmetry is not a probability, although it is scaled by the sum of the inner buy and inner sell shares. This scaling results in mostly small (less than 1) values, as a value greater than 1 would indicate an interval where the change in shares was greater than the sum of the inner shares.

![Sharpe Ratio using Epoch method](image1)

![Total profit using Epoch method](image2)

Figure 5.16: Sharpe ratios and total profits for unlimited buying and selling via the Epoch method.
Figure 5.17: Sharpe ratios and total profits for unlimited buying and selling via the asymmetry method.

The fixed-time method yielded negative returns and negative Sharpe ratios for all cutoff values. The epoch method also yields negative returns and Sharpe ratios that are even worse results than for the fixed-time method. The asymmetry method yields negative Sharpe ratios for the lower cutoff values, but the strategy improves as the cutoff value increases, showing slight positive results for a few cutoff values. Overall, this strategy is inconsistent and requires the modifications about unwinding that were presented earlier.

Adding the condition of unwinding whenever the opposite change is predicted drastically improved the size and consistency of the resulting profits when using the fixed-time method. As seen in graph Figure 5.18, the epoch method yields negative returns for cutoff values below 0.75 and positive returns for cutoff values above 0.75 and below 0.9. The asymmetry method is more successful, as seen in Figure 5.19, yielding positive returns for cutoff values between 0.2 and 0.9. The Sharpe ratios
are between 0.5 and 1 for many of the values, indicating the profits were not only positive but consistent as well.

Figure 5.18: Sharpe ratios and total profits for unlimited position with unwinding using the Epoch method

Placing limits on the size of the positions when trading using the fixed-time method resulted in lower overall profits but higher Sharpe ratios due to improved consistency of returns. Implementing these same limits on the epoch method has a similar impact, raising the Sharpe ratios from around 0.2 to around 0.5 for cutoff values between 0.75 and 0.9 while raising the total profit. However, placing limits on the position for the asymmetry method does not change the Sharpe ratios or the total profit values very much. This lack of change reveals that these limits were not often reached.

Some profit strategies presented here did not use the models for deciding when to unwind and instead held onto the stock for a fixed number of intervals or a fixed amount of time. Holding onto the stock for a fixed number of intervals always
Figure 5.19: Sharpe ratios and total profits for unlimited position with unwinding using the Asymmetry method proves to have a negative expected profit for any model with under 100% predictive accuracy. When trading based on significant probabilities from the epoch method and holding for two epochs, all of the returns are negative. The same is true when using the asymmetry method, due to the expected negative returns resulting from the uncertainty of the second price change.

Figure 5.20 shows the results when holding the orders for a fixed amount of time when using the epoch method. The graph plots Sharpe ratio versus number of minutes held using cutoff values 0.65 (green), 0.75 (purple), and 0.85 (brown). The results follow the same pattern of improving until the holding time reaches around 6 minutes and then degrades. The results get better as the cutoff value increases, with cutoff value 0.85 showing a few positive returns while the others did not. Figure 5.21 shows the results for the asymmetry method using cutoff values 0.25 (green), 0.5 (purple), and 0.75 (brown). Unlike the fixed-time method, which
showed improved success when holding the stock for several minutes (between 5 and 10), the asymmetry method yields positive returns for only the cutoff value 0.75 when holding the stock for under 5 minutes. For the other two cutoff values, 0.25 and 0.5, the patterns were similar to the fixed-time method in that they improved as the holding time was longer, but none of the results were positive.

Figure 5.20: Sharpe ratios using the Epoch method when holding orders for a fixed time

5.4.2 Accounting for Fees

Thus far, none of the strategies have taken transaction fees into account when calculating profitability. Island charges $0.003 (three-tenths of a cent) per share to the trader who adds the order that removes the shares from the order book. The trader whose order was already in the order book gets a rebate of $0.002 per share when the trade occurs, which the ECNs do to encourage people to add orders. ECNs earn $0.001 per share for every transaction, which may not sound like a huge
total, but over 13.5 million shares are traded on Island for Microsoft for this day. The fees paid to Island totaled over $13,500 for only Microsoft, one of thousands of stocks traded on ECNs.

To test the impact of fees on the profitability of the strategies, the unlimited strategy with unwinding will be repeated, but with the $0.003 per share charge applied to every order. Figure 5.22 shows the Sharpe ratios and total profits for this strategy to the fixed-time model without applying fees (black) and when applying the trading fee (red). Applying the fees obviously lowers the total profits and Sharpe ratios, particularly for the lower cutoff values that trade more frequently. Although the profits are lowered when applying the fees, the strategy still yields positive profits for all cutoff values lower than 0.15. At 0.15, the profits become small enough that the fees overcome the profits and negative returns result.
5.4.3 Profit from Other Stocks/Days

Thus far, all of the strategies were evaluated using data from the Microsoft stock, which is the stock used by the PLAT competitions to evaluate the automated strategies devised by the entrants. The strategies presented here that buy and sell based on the models developed in the previous chapter and unwind at appropriate times turned profits for Microsoft. However, applying similar strategies to other stocks must return similarly profitable results before drawing broad conclusions about the success of the strategies.

Using the fixed-time method and the strategy that buys and sells when the probability of price increase or decrease is above a certain cutoff value and unwinds when the opposite probability is above the same cutoff value, the total profits yielded for Cisco and Yahoo are shown in Figure 5.23. The profit from Cisco is in red, the profit from Yahoo is in blue, and the profit from Microsoft is shown in
Cisco follows a similar pattern to Microsoft by showing large profits for lower cutoff values and showing decreasing profits as the cutoff value increases. Although the pattern is similar, Cisco returns smaller profits than Microsoft. Yahoo, which showed consistently worse results from the ROC curves and PSIC values, shows negative profits for cutoff values below 0.8 and positive profits for cutoff values above 0.8. However, Yahoo shows higher positive profits than either Cisco or Microsoft for many of the larger cutoff values (between 0.1 and 0.2). The different results yielded by cutoff values for the different stocks results from the proportion of intervals that contain price changes. Yahoo intervals contain a higher percentage of price changes in one-second intervals (around 7%) than Microsoft (around 2.5%), so it makes sense that the Yahoo models will output higher predictive probabilities of price changes. Thus, the higher cutoff values result in better accuracy and better
profitability for Yahoo than for the other stocks, while the lower cutoff values give better results for Microsoft and Cisco.

Earlier, the PSIC values for predicting price changes on May 23, 2007 using the complete data from May 2, 2005 were higher than the PSIC values for predicting price changes on May 23, 2007 using the incomplete data from earlier on May 23, 2007. Figure 5.24 shows the total profits for those same models, with the May 2, 2005 model in black and the May 23, 2007 model in red. Just like before, the model that uses the larger but older data yields better results than the smaller but newer data. The model that uses May 2, 2005 to predict May 23, 2007 returns positive profit for most of the cutoff values between 0.05 and 0.2. For those same cutoff values, the model that uses half of the limited May 23, 2007 data to predict price changes in the other part returns profits close to 0 and shows a slight positive profit for only a few cutoff values. Again, the more complete data has shown superior predictability and profitability.
Figure 5.24: Total profits for using different days
Chapter 6

Summary and Conclusions

6.1 Summary of Results

This thesis examined the ECN order book in a way that has never been done before. Exploring the dynamics of the movement of shares in and out of the order book at different prices enabled the formulation of models to predict probabilities of the different price outcomes. Then these models laid the groundwork for market strategies that determined the appropriate times to buy, sell, and unwind. By starting at the order, the most basic unit on the ECN, this thesis presented the discovery of dynamics for prices and ways to profit from the dynamics.

The order book activity does not remain consistent throughout the day. Pre-market and after hours trading exhibits a significantly lower level of activity than the regular trading hours, as the access to ECN during these hours are limited. Even within the regular trading hours, the early and late hours show more adding and cancelling of orders than the middle hours. Orders arrive to the order book and leave the order book at a non-homogeneous rate, and the process does not follow a Poisson distribution.

Island and ArcaBook exhibit similar dynamics but also display some key differences. The level of activity on Island is higher than ArcaBook throughout the trading day. The distribution of prices for the added orders is centered around the inner prices for both ECNs, but the ArcaBook orders have a larger variance
of prices. The inner prices follow each other very closely, and only a few arbitrage opportunities exist for gaining profit by taking advantage of price differences across ECNs. These opportunities generally occur at the beginning and end of the days, the hours during which trading is limited and greater chances for abnormalities exist. Greater profit can be obtained by successful modeling than by waiting for these possible instances.

The order book shows primary peaks at the inner prices and secondary peaks centered around four cents away from the inner prices. The shares activity at these prices follow the movement of the inner prices very closely. During periods of time when prices are increasing, the primary peak on the sell order book generally decreases while the primary peak on the buy order book generally increases. Likewise, the secondary peak on the buy side increases and the secondary peak on the sell side decrease during intervals of price decrease. On the other hand, during periods of time when prices are decreasing, the primary peak on the sell order book generally increases while the primary peak on the buy order book generally decreases. Likewise, when prices are decreasing, the secondary peak on the sell side decreases while the secondary peak on the buy side decreases. The price prediction models use the information contained in these orders as the independent variables, and these price patterns are represented in the coefficients of the models.

The fixed-time method splits the day into intervals of one second and uses information about changes in shares in the previous second to generate probabilities of the different types of price changes in the next second. The time scale of one
second was determined to be optimal, as a shorter time scale would result in too many intervals with no shares activity, and a longer time scale would result in individual events getting lost when combined with others. To formulate and test the models, the day is split at 12:00 noon, with the data before noon used to calculate the coefficients for the model and the data after noon used to test the predictive accuracy of the models.

Data from Island, ArcaBook, or both were used to predict the price changes on either Island or ArcaBook. Generally, the models using both ECNs were most successful, followed by the one that used its own ECN, and the worst prediction came from using the opposite ECN. The fixed-time means model used means of orders added to and removed from the order book in the previous second to predict probabilities of price changes for the next second. The fixed-time means method yielded worse predictive accuracy than the fixed-time method that used shares differences, but the results still predicted well.

The epoch method split the day into epochs, defined as intervals of time between consecutive price changes. This method does not use changes in shares in the previous second, but instead uses changes in shares since the beginning of the current epoch. The output no longer predicts the probability of changes in the next second, but instead gives probabilities for the direction of the next price change. That next price change may or may not occur within the second. The epoch method did not perform as well as the fixed-time method, as too many aberrations early in epochs caused poor predictions later on in the epochs.
The asymmetry method uses differences in orders added to and deleted from the order book to evaluate the movement of the order book. Instead of formulating a model with coefficients, the asymmetry method calculates a single value that can directly predict the likelihood of a price change in the next second. The asymmetry method yielded less successful results than the fixed-time method, but the asymmetry method does not require calculation of parameters. While the fixed-time method uses time at the beginning of the day to formulate its coefficients, the asymmetry method is ready to be used right away, allowing for greater trading time.

To evaluate the predictive effectiveness of the models, several diagnostics measured the success of the model output compared to what really happened. The first diagnostic compares the distributions of the probabilities of price increase, decrease, and no changes during intervals where prices increase and decrease. During intervals where price increases occur, the distribution of the probabilities of price increases is significantly higher than the distributions of the probabilities of no price change or of a price decrease. Likewise, during intervals of decreases in prices, the distribution of probabilities of price decreases is higher than the distributions for the probabilities of no price change or price increases.

The distributions of probabilities provide a visual way to see whether the model yields successful results. However, ROC curves give a numerical and visual way to evaluate model success. ROC curves use sensitivity and specificity analyses to measure the true positive rate and true negative rate. For different probability
cutoff values, certain intervals are tagged as significant if the predicted probability of a price increase or the predicted probability of a price decrease is greater than the cutoff value. The predicted outcomes are compared to the actual outcomes, and the true positive and true negative rates for different cutoff values create a curve. The models generate ROC curves that lie well above the line of chance and yield a large area under the curve, which measures general success of ROC curves.

The last measurement of predictive success is the PSIC, which measures the percentage of intervals tagged as significant that correctly predict the direction of the next price change. The PSIC provides an indication of a model's profitability when applied to a market strategy, because correctly predicting the direction of the next price change creates opportunities for earning profit. Generally, PSIC values increase as the cutoff values increase, due to the greater selectivity. The different models yield high PSIC values, some as high as 0.90, which indicate accurate prediction and potential for profitability.

Several of the existing strategies that claimed positive returns and successful implementation on ECNs yielded poor results when tested on the data from Island and ArcaBook. These strategies may have succeeded at some time due to fortunate market conditions or were seen as successful compared to other, more unsuccessful strategies. Many papers presented several strategies, some of which used similar but slightly different techniques. When entering the PLAT competition, more strategies tested in competition means a better chance of one succeeding and performing well. One of the PLAT competitions featured two strategies in the final round that each
returned a negative Sharpe ratio after performing the best in the previous rounds. Such inconsistency causes concerns when implementing these strategies into real situations with real money involved.

Many of the existing strategies that yielded positive Sharpe ratios, such as the market-making strategies presented by Feng et al. [FYS04] and Sherstov and Stone [SS05], gained profit by capitalizing on price fluctuations throughout the day. A modified form of the SOBI strategy [SR05] uses similar criteria for buying and selling as the strategies based on the asymmetry method and returns positive results.

The strategies developed in this thesis rely on successful price modeling, accurate predictions about probabilities of price changes, and wise unwinding techniques to gain profit. The most profit can be earned when several successive price changes occur in the same direction, as the trader can develop a progressively larger position and unwind for a greater profit. Also, due to the nature of ECN prices, accurate prediction can only profit when more consecutive price changes occur in the same direction.

The strategies based on using the fixed-time, epoch, and asymmetry methods performed similarly, in that the strategies that worked well using one of them generally worked well on the others. Also, the strategies that performed badly on one performed badly on all. The strategies that performed poorly include the method of unlimited buying and selling without unwinding until the end, holding onto the stock for two (or another fixed number) of epochs, and holding onto the
stock for a fixed amount of time. The successful strategies were the ones that used the models to determine the appropriate time to buy and sell as well as when to unwind the position. Limits placed on the size of the position or on the number of orders added per epoch reduced the amount of money that could be gained or lost, resulting in smaller but more consistent returns. A trader could set his or her own limits depending on the amount of money invested and risk-seeking or risk-averse tendencies.

Not all of the strategies presented yielded positive results. However, those that were not successful failed due to a built-in flaw, and the removal of the flaw resulting in greater profitability. These flaws were put in there intentionally to verify that their inclusion would ruin an otherwise successful strategy. The most successful strategies that made buying and selling decisions based on the predictive probabilities from the models included conditions for unwinding throughout the day instead of at the end. The models were formulated using one second at a time, so logically the best strategies do not allow for holding a shares position for a long time if the model does not agree.

Although a majority of the results presented throughout this thesis used Microsoft stock, similarly successful results were achieved using other spread-minimal stocks, such as Cisco and Yahoo. Models developed using data from these other stocks resulted in successful prediction, and the trading strategies that showed positive profits for Microsoft also showed positive profits for these other stocks. The applicability of the models and strategies to multiple stocks reaffirms the techniques
developed throughout this thesis.

All of the information used in the formulation of the models and the development of the profitable strategies came directly from orders placed by traders to buy or sell a stock. No external information, such as historical price trends or market news on that day, was included throughout this process. Market news includes any information made available during the trading day that may influence trader decisions and stock prices. ECNs are user-driven markets, and the patterns of the prices can be determined based on the actions of the traders.

After formulating models and testing trading strategies for Microsoft, the most successful models and strategies were applied to Cisco and Yahoo. The other stocks yielded similar results, including the same significant variables and similar ROC and PSIC plots. Likewise, the strategies that showed positive profits for Microsoft also showed positive profits for the other spread-minimal stocks, although testing on Microsoft yielded the largest total profits.

The models that used both Island and ArcaBook were tested using data from May 2, 2005, because that was the only day for which data from both ECNs was available. However, models using only Island data were created for May 23, 2007, the most recent day for which data was available. Only a limited amount of data was available, resulting in an inability to formulate an accurate model. However, applying the coefficients formulated from the May 2, 2005 model to predict price changes on May 23, 2007 yielded better predictive success and showed positive profits when using the trading strategies.
6.2 Future Work

The models and strategies described throughout this thesis captured order book dynamics and yielded successful results, but future work can be done to build off the work presented here. Most of the results presented here used data from May 2, 2005, which was the only day for which ArcaBook data was available. Other more recent days were available for Island, but the data only included the early part of the trading day. Choosing a day on which a noteworthy event occurred, such as an introduction of a new product or a takeover of another company, may add an extra element of uncertainty. Including external information and comparing the results to the models that solely use shares information may reveal insight into whether the models capture everything necessary on “special” days.

The models required spread-minimal stocks, as only these stocks exhibit the qualities needed for the kinds of price prediction that these models can perform. Although the information used in the models presented here do not work on stocks that have larger spread, the order books for these stocks may contain different information that can be used for predicting price changes. The order books for stocks such as Google, which has a spread significantly larger than one-cent, may include different factors that could provide the information for price modeling. Further examination of the order books for these stocks is needed before performing price modeling on them.

Another modification on the strategies combines the models, just as Silaghi and Robu [SR05] combined existing strategies. Instead of buying and selling according
to the fixed-time model, epoch model, or asymmetry model, using combinations of these models may result in greater profitability. Using a $\frac{2}{3}$ rule or only acting when all models agree may result in greater consistency, as an aberration on one model may not affect the others and false positives may be reduced.
Appendix A
Parsing Data

This section provides a detailed description of the process by which the data used throughout the thesis was collected from the much larger data sets provided by Island and ArcaBook. The initial data for a given day on Island includes every added (A), deleted (X), traded (E), and hidden traded (P) order for every stock, ordered chronologically from the beginning of the day to the end of the day. The hidden trades were omitted throughout the work, so the hidden trades do not need to be parsed. For an added order, the data indicates the time (in milliseconds after midnight), an “A” for add, the order’s unique ID number, a “B” to indicate a buy or an “S” to indicate a sell, the number of shares, the stock, and the price. Traded and cancelled orders indicate only the time, “X” for cancellation or “E” for execution/trade, the ID number, and the number of shares, but not the name of the stock or the price.

In order to retrieve all of the adds, cancellations, and trades for a given stock on a given day, a computer program searches for all lines of data in the original file that correspond to that particular stock. Only the added orders contain the name of the stock, so in order to identify a trade or cancellation, all occurrences of the ID numbers that correspond to added orders for the stock are included. The price for a traded or deleted order is determined by looking back at the added order.

Once all of the lines of data for a particular stock’s added, traded, and deleted
orders are parsed into a file, the inner prices can be determined by calculating the maximum value of any buy order that has been added but not fully traded or cancelled and the minimum value of any sell order that has been added but not fully traded or cancelled. After obtaining the inner prices at a given time, standardized price for the added orders are calculated. Also, lifetimes for a given order can be determined by subtracting the time of the last trade or cancellation for an order's ID number minus the time when the order was added.

The data for ArcaBook is formatted differently from the Island data. The most notable difference is the lack of distinction between trades and cancellations in the ArcaBook data. The parsing procedure goes similarly for ArcaBook as for Island, but ArcaBook requires careful attention to certain details. The Island data maintained column alignment for added and traded/deleted orders. However, ArcaBook trades and cancellations have some of the information in different columns for added orders and removed orders. However, the data contains all of the same information, so as long as the different columns are taken into account, the parsing process on ArcaBook will proceed in similar steps as for Island.
Appendix B

Data and R Code

Below are five sample lines from the raw Microsoft data, then the same five lines after formatting, and then the corresponding five lines in the data set of the inner prices. The first two lines are added orders, each for selling 1000 shares at $27.30, occurring shortly after 10:00 AM (equivalent to 36,000,000 milliseconds after midnight). The next two lines indicate trades, for 1000 and 100 shares respectively, that occur on the buy side at $27.29. The raw data does not name price and whether it was a buy or sell, but these items can be determined in the R program (code provided later) by reverting back to when the order with that unique ID number was added to the order book. The numbers 793181 and 793182 in the raw data are counters of the numbers of trades for all stocks that occur during the day and do not provide relevant information for this thesis. The final line indicates a cancellation of a previous buy order for $27.29. Each of the five lines reflects an event at the inner buy or inner sell price, and the addition or removal of shares is shown in columns 2 and 5 of the inner price data.
Raw MSFT data (I):

36006862A 16333901S 1000MSFT 273000Y 
36006862A 16334005S 1000MSFT 273000Y 
36006880E 15346499 1000 793181 
36006880E 15346536 100 793182 
36006963X 16267614 300 

Formatted MSFT data (II):

"36006862" "A" "16333901" "27.3" "1000" "S"
"36006868" "A" "16334005" "27.3" "1000" "S"
"36006880" "E" "15346499" "27.29" "1000" "B"
"36006880" "E" "15346536" "27.29" "100" "B"
"36006963" "X" "16267614" "27.29" "300" "B"

MSFT inner prices data (III):

"36006862" "32308" "27.29" "27.3" "14345"
"36006868" "32308" "27.29" "27.3" "15345"
"36006880" "31308" "27.29" "27.3" "15345"
"36006880" "31208" "27.29" "27.3" "15345"
"36006963" "30908" "27.29" "27.3" "15345"

Code for transforming data from raw (I) to formatted (II)

```python
# ms.data is raw data
```
ms1 = ms.data[,1]  # time and type
ms2 = ms.data[,2]  # id number and buy/sell (A,P only)
ms3 = ms.data[,3]  # shares and stock
ms4 = ms.data[,4]  # price
stock.name = "MSFT"
time = as.numeric(substring(ms1, 1, 8))
type.of.event = substring(ms1, 9, 9)  # A,X,E,P
big.matrix = matrix(0, length(ms1), 6)  # create matrix for data in new format
big.matrix[,1] = time  # insert time into column 1
big.matrix[,2] = type.of.event  # insert type of event (A,X,E,P) into column 2
x.events = which(type.of.event == "X")  # cancellations
a.events = which(type.of.event == "A")  # adds
e.events = which(type.of.event == "E")  # trades
p.events = which(type.of.event == "P")  # hidden trades
new.ID = as.character(ms2)  # id number and buy/sell
big.matrix[x.events,3] = new.ID[x.events]  # insert ID number into column 3
big.matrix[a.events,3] = substring(new.ID[a.events], 1, nchar(new.ID[a.events]) - 1)
big.matrix[e.events,3] = new.ID[e.events]
big.matrix[p.events,3] = substring(new.ID[p.events], 1, nchar(new.ID[p.events]) - 1)
new.price = as.character(ms4)  # price (dollars x10000)
big.matrix[a.events,4] = as.numeric(substring(new.price[a.events], 1, nchar(new.price[a.events]) - 1)) / 10000  # price for added orders (column 4)
big.matrix[p.events,4] = as.numeric(substring(new.price[p.events], 1, nchar(new.price[p.events]))) / 10000  # price for hidden trades (column 4)
new.shares = as.character(ms3)  # number of shares
len.stock.name = nchar(stock.name)  # number of characters in stock name
big.matrix[x.events,5] = new.shares[x.events]  # insert shares into column 5
big.matrix[a.events,5] = substring(new.shares[a.events], 1, nchar(new.shares[a.events]) - len.stock.name)
big.matrix[e.events,5] = new.shares[e.events]
big.matrix[p.events,5] = substring(new.shares[p.events], 1, nchar(new.shares[p.events]) - len.stock.name)
big.matrix[a.events,6] = substring(new.ID[a.events], nchar(new.ID[a.events]))  # insert buy/sell into column 6
big.matrix[p.events,6] = substring(new.ID[p.events], nchar(new.ID[p.events]))
# fill in prices (column 4) and buy/sell (column 6) for trades (E) and cancellations (X)
remove.events = sort(c(e.events, x.events))  # find rows where type of event is E or X
for (i in 1:length(remove.events)) {
    temp.count=remove.events[i] # row number
    temp.id=big.matrix[temp.count,3] # ID number associated with that order
    which.id=which(big.matrix[1:temp.count,3]==temp.id) # find all rows with that ID number
    first.count=min(seq(start.index, end.index)[which.id]) # find 1st row with that ID number (added)
    big.matrix[temp.count,4]=big.matrix[first.count,4] # insert price from row where order was added
    big.matrix[temp.count,6]=big.matrix[first.count,6] # insert buy/sell from row where order was added
}

Code for using formatted data (II) to create inner price data (III)

# island.data: imported data
island.time=as.numeric(as.character(island.data[,1]))
island.type=island.data[,2]
island.id=as.numeric(as.character(island.data[,3]))
island.price=as.numeric(as.character(island.data[,4]))
island.shares=as.numeric(as.character(island.data[,5]))
island.buyorsell=island.data[,6]
island.buy.matrix=matrix(0,1000,2) # create matrix for buy order book
island.sell.matrix=matrix(0,1000,2) # create matrix for sell order book
inner.price.matrix=matrix(0,length(island.time),5) # create matrix for inner prices
# columns: (1) time,
# (2) inner buy shares, (3) inner buy price,
# (4) inner sell price, (5) inner sell shares
inner.price.matrix[,1]=island.time # insert time in column 1
# loop goes through each add/deletion/trade, updates order book, and calculates inner buy/sell price/shares
for (i in 1:length(island.time)) {
    if (island.buyorsell[i]=="B") { # buy order
        if (is.element(island.type[i],"A") { # added order
            if (is.element(island.price[i],island.buy.matrix[,1])==1) {
                island.buy.matrix[which(island.buy.matrix[,1]==island.price[i]),2]=
                island.buy.matrix[which(island.buy.matrix[,1]==island.price[i]),2]+island.shares[i] }
            if (is.element(island.price[i],island.buy.matrix[,1])==0) {
                island.buy.matrix[length(island.buy.matrix[,1])[island.buy.matrix[,1]!=0]) +1,1]=island.price[i]
                island.buy.matrix[length(island.buy.matrix[,1][island.buy.matrix[,1]!=0]),2]=island.shares[i]
            }
        }
    }
}
if (island.type[i] == "E") { # traded order
    temp.val = which(island.buy.matrix[,1] == island.price[i])
    island.buy.matrix[temp.val,2] = island.buy.matrix[temp.val,2] - island.shares[i]
}

if (island.type[i] == "X") { # cancelled order
    temp.val = which(island.buy.matrix[,1] == island.price[i])
    island.buy.matrix[temp.val,2] = island.buy.matrix[temp.val,2] - island.shares[i]
}

if (island.buyorsell[i] == "S") { # sell order
    if (island.type[i] == "A") { # added order
        if (is.element(island.price[i], island.sell.matrix[,1]) == 1) {
            island.sell.matrix[which(island.sell.matrix[,1] == island.price[i]),2] =
            island.sell.matrix[which(island.sell.matrix[,1] == island.price[i]),2] + island.shares[i]
        }
        if (is.element(island.price[i], island.sell.matrix[,1]) == 0) {
            island.sell.matrix[length(island.sell.matrix[,1][island.sell.matrix[,1] != 0]) + 1,1] = island.price[i]
            island.sell.matrix[length(island.sell.matrix[,1][island.sell.matrix[,1] != 0]),2] = island.shares[i]
        }
    }
    if (is.element(island.price[i], island.sell.matrix[,1]) == 0) {
        island.sell.matrix[length(island.sell.matrix[,1][island.sell.matrix[,1] != 0]) + 1,1] = island.price[i]
        island.sell.matrix[length(island.sell.matrix[,1][island.sell.matrix[,1] != 0]),2] = island.shares[i]
    }
    if (island.type[i] == "E") { # traded order
        temp.val = which(island.sell.matrix[,1] == island.price[i])
        island.sell.matrix[temp.val,2] = island.sell.matrix[temp.val,2] - island.shares[i]
    }
    if (island.type[i] == "X") { # deleted order
        temp.val = which(island.sell.matrix[,1] == island.price[i])
        island.sell.matrix[temp.val,2] = island.sell.matrix[temp.val,2] - island.shares[i]
    }
}

inner.buy = max(island.buy.matrix[,1][island.buy.matrix[,2] != 0]) # calculate inner buy
inner.sell = min(island.sell.matrix[,1][island.sell.matrix[,2] != 0]) # calculate inner sell
inner.price.matrix[i,3] = inner.buy # inner buy in column 3
inner.price.matrix[i,4] = inner.sell # inner sell in column 4
which.buy = which(island.buy.matrix[,1] == inner.buy)
which.sell = which(island.sell.matrix[,1] == inner.sell)
inner.price.matrix[i,2] = island.buy.matrix[which.buy,2] # inner buy shares in column 2
inner.price.matrix[i,5] = island.sell.matrix[which.sell,2] # inner sell shares in column 5
Bibliography


S. Kooths, T. Mitze, and E. Ringhut. Inflation forecasting-a comparison between econometric methods and a computational approach based on


