Determinación de la factibilidad de la detección de estrategias de operación en el mercado de divisas colombiano utilizando la información del libro de órdenes.

Andrea Marcela Cruz Moreno

Universidad Nacional de Colombia
Facultad de Ingeniería
Bogotá, Colombia
2016

UNIVERSIDAD NACIONAL DE COLOMBIA

Determining feasibility of trading strategies detection using order book information from the Colombian currency market

Andrea Marcela Cruz Moreno

Universidad Nacional de Colombia
Engineering School
Bogotá, Colombia
2016

Determining feasibility of trading strategies detection using order book information from the Colombian currency market

Andrea Marcela Cruz Moreno

Submitted to the Engineering School of the Universidad Nacional de Colombia, in partial fulfillment of the requirements for the degree of:

Master of Science in Systems and Computer Engineering

Under the guidance of:
Germán Jairo Hernández Pérez, Ph.D.
Associate Professor
Engineering School

ALGOS Research Group
Universidad Nacional de Colombia
Engineering School
Bogotá, Colombia
2016

To Laura, Uccelito and Abel Cruz.

Acknowledgements

This thesis would not have been possible without the help, support, guidance and trust of my thesis advisor, Professor Germán Hernández.

Also, I am grateful to my amazing parents and siblings: Laura, Tom, and Diego who have always given me their unconditional love, support, encouragement and advice. Thanks to Benjamin Walz for his loving support in the final stage of this work.

Abstract

The detection of effective, i.e. profitable and efficient, trading strategies requires the identification of predictive patterns in the information provided by the market. Usually the market information is presented as a time series collection of prices (open, close, high and low) and volume that are available with a particular time granularity. This depends on the information provider, e.g. every transaction, every minute, every day, etc. depending on the access level.

In this work we have access to an information tool that is not commonly available: The Limit Order Book information for the Colombian Bulk currency market.

Order book data provides a valuable source of information in financial markets. For this reason, it is an excellent candidate for the construction of new trading tools and models. Order book representation is an still active study branch in quantitative finance.

This work addresses the problem of information visualization of financial data from Colombian Bulk Currency exchange using two approaches: a heatmap representation, and a Haar Wavelet based representation in order to filter high frequency noise. This requires dealing with a massive amount of data coming from the Colombian Forex Market Limit Order Book, a register with all the buy and sell intentions of the market's participants.

The experimental evaluation shows that the proposed strategies are able to identify frequent patterns within the presented visualizations tools. Furthermore, and more important, it is possible to associate some of those frequent patterns with a trend with a probability greater than 0.5. This result is useful in order to generate buy and sell signals for a trader.

Keywords: Order book, Haar Wavelet, Scientific Visualization, Financial Engineering, Machine Learning, Heatmap, Information Representation.

Resumen

La detección de estrategias comerciales efectivas, es decir, rentables y eficientes, requiere la identificación de patrones predictivos en la información proporcionada por el mercado. Normalmente el mercado la información se presenta como una colección de series temporales de precios (apertura, cierre, máximo y mínimo) y volumen que están disponibles con una granularidad de tiempo particular. Esto depende del proveedor de información. Por ejemplo, cada transacción, cada minuto, cada día, etc. dependiendo del niver de acceso.

En este trabajo tenemos acceso a una herramienta de información que comúnmente no está disponible: El Límite Información del Libro de órdenes para el mercado de divisas intradiario colombiano.

Los datos del libro de pedidos proporcionan una valiosa fuente de información en los mercados financieros. Por esta razón, es un excelente candidato para la construcción de nuevas herramientas y modelos comerciales. La representación del libro de pedidos es una rama de estudio todavía activa en las finanzas cuantitativas.

Este trabajo aborda el problema de la visualización de información de datos financieros del cambio de divisas a granel de Colombia utilizando dos enfoques: una representación de mapa de calor y un Haar. Representación basada en wavelet para filtrar ruido de alta frecuencia. Esto requiere tratar con una gran cantidad de datos provenientes del Libro de órdenes Límite del Mercado Forex de Colombia, un registro con todas las intenciones de compra y venta de los participantes del mercado.

La evaluación experimental muestra que las estrategias propuestas son capaces de identificar frecuentes patrones dentro de las herramientas de visualización presentadas.

Keywords: Libro de órdenes, Haar Wavelet, Visualización científica, Ingeniería financiera, Aprendizaje de Máquina, Mapa de Calor, Representación de información, Mercado de Divisas, Bolsa de palabras.

Declaración

Me permito afirmar que he realizado la presente tesis de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en la propia tesis. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los he reconocido en el presente trabajo. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de tesis.

Suttgart, 29.08.2022

Andrea Marcela Cruz Moreno

Content

List of Figures

List of Tables

1. Introduction

The Limit Order Book (LOB) for a financial instrument gathers two lists: one orders form buyers and other for sellers' orders (bids and offers) registering the price and volume that they are willing to trade [65]. Figure **1-1** shows an example of a LOB for the Colombian Forex Market. The buy orders list corresponds to the bid side and the sell orders side is equivalent to the ask side. These price and volume changing orders lists are placed by market agents who produce book movements. The best quote is the first order in each list: this is, the maximum price for the bid side and the minimum price for the ask side.

Figure **1-1** is a 2018 snapshot of the visible segment of the order book. The last registered transaction was at 3018.00 COP. At that moment, it was the market price. LOB contains many more entries, but only those close to the best quote are shown. In this book representation, orders at the same price are not aggregated. Using same color implies orders at the same price. Nevertheless, orders can be accumulated at the same price level, as shown in this work.

The so called matched markets, have an engine that pairs and executes orders according to a matching criteria. Other markets, called unmatched, require the intervention of a trader to execute an order. Examples of order execution can be found in figure **1-2**. Events (transactions) define discrete time points; this is not a continuous process, in this figure, lines parallel to the vertical axis represent events.

An example of bid side distribution can be seen in figure **1-3**. The difference between the best quote of the ask side and the one in the bid side is called "spread". Within a given price, several orders from different traders can be cummulated in order to shape the volume for that quote; the priority for the order's execution is assigned according to the order placement queue.

An order is a combination of price and volume which is placed within the order book at a given time point. Market actors can perform different kind of operations in the order book. Brokers can put or remove orders at the best quotes. Other orders, placed in positions other than the best quote can also be put and removed. Market participants can modify their orders changing volume or price as desired, they also have the possibility of execute an order

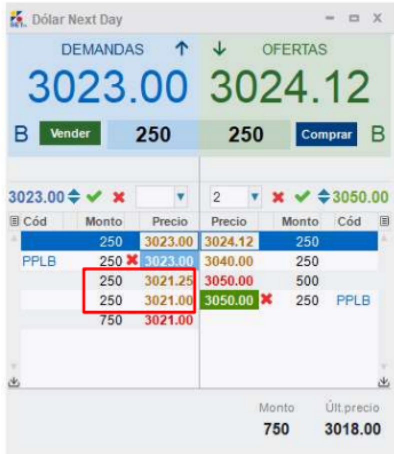


Figure 1-1.: Colombian Limit Order Book example [44].

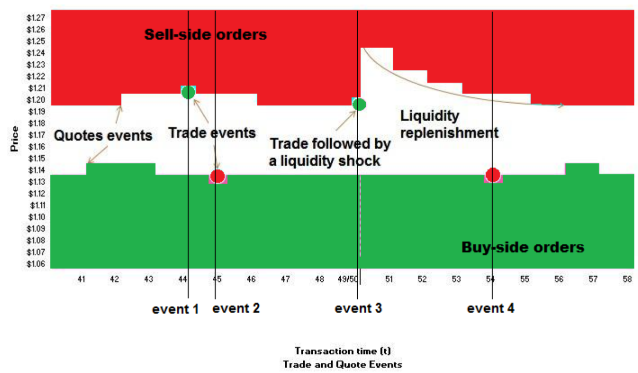


Figure 1-2.: Order execution example [66].

at the best quote or subsequent orders if the volume available at the best quote is not enough for the trader's demand, see figure **1-4**

The LOB evolves by discrete time events produced by buyers and sellers which insert, modify or delete limit orders and also, by execution of orders, either by the engine on matched markets or by liquidity takers on unmatched markets [17]. An illustration of some of these events is presented in figure **1-3**.

A Limit Order Book can be considered a variable length list where transactions occur in non-uniform time intervals. At the instant when transactions occur, the spread is zero. As shown in figure **1-3**, for a given time slot, information of prices and the cumulative volume for each price are provided.

Figure 1-3.: Bid-ask distribution example [22].

Recently, there has been a strong interest in modeling and analysis of LOBs dynamics because these provide richer information about the market trends than the one available using closing prices [27], [46], [47], [48], [50], [59] and [60]. Closing prices are in general publicly available today while LOB information is usually only available to brokers or professional investors willing to pay for this extra information.

Huang et al. [27] model order book high frequency dynamics in the London Stock Exchange using second level data (best quotes and volume at different prices) capturing the arrival of orders of different sizes. Ahn et al. [2], conclude that the further from the best a quote is, the less information it provides.

Palguna et al. [47] describe changes on LOB, specifically, forecasting price changes and separately the direction of such change. This approximation does not produce statistically significant results.

There are multiple models that describe the dynamics of the order book. In terms of characterizing the LOB of multiple markets, see for example,[50]. Vvedenskaya et al. [59] have analyzed the order book shape and transactions aggressiveness of liquidity seekers and liquidity providers. Moreover, they have moved forward to describe the LOB evolution assuming markovian properties and describing the book using parametric models.

There have been found contradictory evidence about LOB information content. [48] and [27] state that the whole book (not only best quotes) helps to determine price direction.

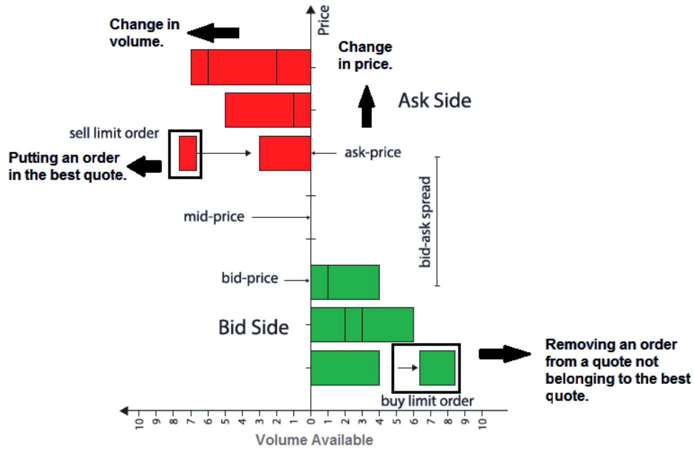


Figure 1-4.: Operations in the order book, modification of [22] for illustration purposes.

However, it seems to be an indirect relationship between information content and distance from best quotes [2].

Some authors present evidence of increased price predictability when LOB information is gathered and used [52]. However, there are other authors who describe not statistically significant results [47]. It seems that final results are highly dependent on the type of market that is been analyzed.

Other studies have tried more complicated representations for the LOB. Jiang et al. [32] propose a representation of four parameters by snapshot. A Kalman filter is used to estimate a linear dynamic system state and provide a liability measure, being used for prediction and filtering. Cheng et al. [9] used order book's slope to outline its informative content.

Due to the massive amount of information generated in electronic markets [41], efficient methods and hash functions to handle and operate with this data are helpful [62]. This work provides a methodological approach to manage this kind of data in order to extract information useful for profit generation.

Limit Order Book has become a valuable source of knowledge for traders, taking an important role in research and as support tool for making financial decisions [1] [4] [9] [23] [65].

One of the main issues in research is the size of the order book, due to the number of orders placed and cancelled every minute. For example, for London Exchange Market, for a single stock, several orders were placed within one millisecond in July 2009 [28]. This provides vast repositories of valuable data, which is hard to process and manage duly.

Due to this situation, the development of strategies for extracting information from these data is required, producing an increasing interest in this research field lately (See Chapter 2). Likewise, the demand of such strategies has increased in recent years.

Foreign exchange markets are essential for the correct operation of the world economy. By the time of writing of this thesis (See chapter 2), there is no evidence of publications of work related to the Colombian Forex Market Limit Order Book, a fact that motivates this work. The data used for this thesis is not publicly available, it was provided by *Algocodex*[1] in cooperation with *Acciones y Valores*[2] which gathered and recorded the information. Dollar buy and sell transactions are spot contracts, i.e. the transaction and its fulfillment are made on the same day. A trader buying dollars through SET-FX, receives and pays the product the same day. This market operates on workdays between 8:00 a.m. and 1:00 p.m.

1.0.1. Colombia

Colombia is an emerging market [3] belonging to the group of the so called CIVETS, a group of non-industrialized countries that despite of that fact are promissory economies due to the fast growing of their Gross Domestic Product (GDP) and the characteristics of their populations. for instance, youthfulness.

Colombia is also part of the Integrated Latin-American Market (MILA for its acronym in Spanish), being part of the biggest Stock Exchange in the Region, outperforming BOVESPA the Brazilian Stock Exchange. MILA is composed of the Stock Exchanges from Chile, Colombia, Mexico and Peru [29].

Colombian Stock Exchange doesn't record information about Limit Order Books or Stock transactions. Not even the Bolsa de Valores de Colombia (Colombia Stock Exchange) keeps track of this information.

This work addresses the problem of information extraction from financial datasets. The main goal of this research was to study trading strategies using order book information from the Colombian Forex Market and its potential in the construction of predicting models. This work additionally presents a visualization tool in order to facilitate the trader's

[1] Intelligent Algorithmic Trading: Algorithmic trading strategists specialized in algorithmic trading. Development of strategies ranging from classical algorithmic strategies to advanced computational game theory and computational intelligence based strategies.

[2] Colombian Stockbroker

[3] City Diary (2010-07-12). "Geoghegan digests and delivers new acronym". London: Telegraph.co.uk. Retrieved 2012-06-28.

understanding of large amounts of Limit Order Book data.

1.0.2. Strategy

Our definition of strategy is the one given by Hayes on [25]: *A forex trading strategy is a technique used by a forex trader to determine whether to buy or sell a currency pair at any given time.* [25] also provides the components to an effective forex trading strategy:

- Market selection.
- Position size.
- When to enter a position.
- When to exit a position.
- Rules for how to buy and sell.

There first two components are predetermined in this case: The market is the Colombian Forex Market and Position size is fixed to one unit. The last component, selecting the right execution technologies, is out of the scope of this work. The system will provide information about when to buy or sell.

1.1. Thesis goals

The main goal of this research was to study trading strategies using order book information from the Colombian Forex Market and its potential in the construction of predicting models. The following is the description of this research specific targets:

- Providing a survey of the methods published to date for detecting trading strategies using order book information, via a systematic literature review.
- Selecting or designing a methodology able to represent in a summarized and efficient way the order book information.
- Establishing a time window in order to preserve relevant order book information.
- Selecting or designing a methodology that allows representing properly the Colombian Forex Market Order Book information dynamics.
- Selecting or designing a trading strategies detection system for the Colombian Forex Market using Order Book information.
- Evaluating the performance and feasibility of the proposed system, in supporting the financial decision making process in the Colombian Forex Market.

1.2. Main contributions

The following is the summary of the main contributions of this work:

1.2.1. Modelling and Visualization of LOB

A Colombian Forex Market Order Book Visualization is presented. This visualization provides the trader with a framework which allows the interpretation of large sections of the limit order book at a glance. It shows relationships between price, volume and time directly.

This work was published as a contributed talk named «Order Book Microstructure Visualization: The case of Colombian High- Frequency Foreign Market. XIII Latin American Congress of Probability and Mathematical Statistics CLAPEM. September, 2014.»

1.2.2. Efficient trend predictive LOB patterns dictionary building

Algorithms for association between frequent patterns and a specific trend are depicted. These algorithms allow calculating the probability of each pattern of being associated with a bearish trend, a bullish trend or with no trend, labeling each pattern accordingly. The use of these algorithms allowed to detect patterns seasonality in the Colombian Forex Market Order Book.

This work was presented under the title «Market Trend Visual Bag of Words Informative Patterns in Limit Order Books» in the 6th Annual Stevens Conference on High Frequency Finance and Analytics (HF2015) that was held on October 29th-31st, 2015 at Stevens Institute of Technology, Hoboken, NJ, USA. It was published under the same title in the International Conference on Computer Science Proceedings. San Diego, California, U.S.A. (ICCS2016).

1.2.3. Bags of trend predictive LOB patterns using wavelets

Algorithms for Frequent Patterns Exploration are presented. These algorithms have reduced the amount of time required for mining a dataset up to two orders of magnitude depending on the pattern size, thanks to the use of a pattern summary function. The use of the Haar Transform, in some time windows, can reduce the initial dataset without loss of accuracy for the classifier, so it reduces the amount of non valuable information.

1.2.4. An algorithmic trading strategy for the Colombian US dollar inter-bank bulk market SET-FX based on an evolutionary TPOT AutoML predictive model

Presents a competitive algorithmic trading strategy for the Colombian US dollar inter-bank bulk order-driven market SET-FX based on an evolutionary predictive model built using the Tree-based Pipeline Optimization Tool (TPOT), an strongly typed genetic programming based automated machine learning tool that uses a mutliobjective evolutionary algorithm the Non-dominated Sorting Genetic Algorithm II (NSGA-II) to search for machine learning models with maximum accuracy and minimum complexity. ICAIF 22 - Women in AI and Finance

1.3. Methodological Framework

Following the approach suggested by Arksey and O'Malley [3] a five step Methodological Framework was selected for the development of this work:

- Specify the research question: Is it posible define a method to exploit the Colombian Limit Order Book's potential in the construction of predicting models for the Colombian Forex Market?
- Identify relevant literature: A survey of the methods published for detecting trading strategies using order book information, via a systematic literature review was conducted.
- Select studies: The state of the art using the relevant literature from previous step was written. It can be found on appendix E.
- Extract, map, and chart the data: The dataset was provided by Acciones y Valores, a broker in the Colombian Stock Market. It was mapped to a heatmap and then the haar wavelet transform was applied up to 4 times.
- Summarize, synthesize, and report results: These components are presented in chapters 3 to 7 and in 4 International publications (See previous subsections).

1.4. Thesis organization

The organization of this thesis is as follows:

- In chapter 1, an introduction and the main contributions of this work will be presented.
- The Modelling and Visualization of LOB are introduced in chapter 2.

- The Efficient trend predictive LOB patterns dictionary building is depicted in chapter 3.
- Bags of trend predictive LOB patterns using wavelets is detailed in chapter 4. A similar approach using clustering is published in chapter 5.
- An Effective trading strategy based on bags of trend predictive LOB patterns using clustering is presented in chapter 6.
- Finally, in chapter 7, conclusions and future work are described.

2. Modelling and Visualization of LOB

This chapter describes in detail the information's nature, its source, the way in which visualizations are built and how to interpret each image and the meaning of the visual components.

Experiments were conducted using real tick data of foreign exchange rate USDCOP from March to May of 2012. LOB provides information about time, price and volume for every request in the market; this information was summarized every minute, in a price range of 120 COP in the best quotes, using a 20 cents mark up. Volumes were quantized in levels of USD 250,000, which is the minimum trading volume for this market. The maximum volume observed for a particular order during the analyzed period was 43.5 USD millions. The maximum price observed was 1,862.6 COP and the minimum was 1,742.2 COP.

2.1. Heatmap based approach

This section presents a heatmap of the discretized volumes and prices of the selected currency as a visualization tool for the market microstructure. Heatmap's construction and interpretation are explained in this section.

2.1.1. First approach

The first approximation to a heatmap based representation, was gathering in a single matrix order price, volume price and its evolution over time. In the x axis, the book's depth (price) was set. In the y axis, time is represented and, finally, lightness indicates orders' cumulative volume. In this way figure **2-1** was assembled.

Thereafter, different thresholds were applied to images produced with the previously described procedure. In this way, images from figure 2.1.1 were obtained. In this figure local time evolving trends in price can be observed.

Figure 2-1.: First approach to the order book heatmap visualization.

2.1.2. Towards a better information understanding

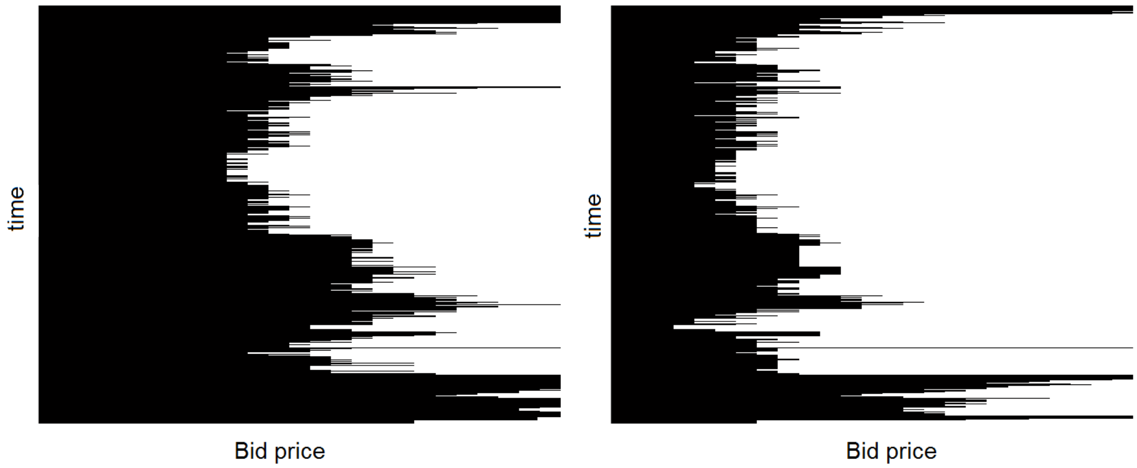
In order to improve information understanding, figure **2-3** was produced. In this figure, x axis corresponds to price information, y axis come into time and gray level is used for representing orders' volume (not cumulated volume) gathered for price-time intervals. The white meander in the middle of the image corresponds to the spread, it means that in this representation ask and bid order books were combined (ask order book above the meander and bid order book below the meander).

This chapter proposes a tool to facilitate a human trader locating visual patterns. The proposed visualization can be observed in figure **2-4** is built as follows:

1. Book event's discretization: Events which happen every minute are aggregated. Increases along the x coordinate indicate progress over time.
2. Trading volume quantization: Volumes within a range of prices in an instant of time are cumulated; higher volumes are indicated with a lower level of lightness.
3. Price quantization for each time unit: An increase along the y-axis in the picture indicates a higher price.

A similar representation named *BookMap X-ray*[1] is provided by VeloxPro. This represen-

[1] http://www.bookmap.com/



a) Low threshold.

b) High threshold.

Figure 2-2.: Different gray level thresholds for the previous image.

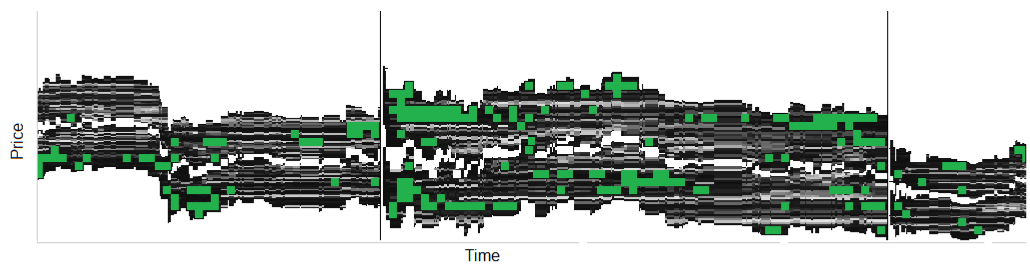


Figure 2-3.: Example of position of frequent patterns in a heatmap order book image for 2 days of trading.

tation is also volume inspired, however is built in gray scale, it does not take advantage of chromatic differences or relationships and it takes into account just the five best quotes.

2.2. BookMap by VeloxPro

A tool for exploring Order Book Dynamics based on the heatmap principle can be found in [6], this tool is named BookMap by VeloxPro. This section presents an outline of the aforementioned tool.

Figure **2-5** informs about the type of user of BookMap according to VeloxPro.
The main advantages for using Book according to VeloxPro, are:

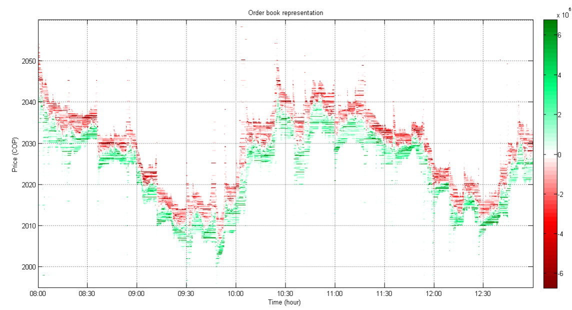


Figure 2-4.: Example of the proposed order book visualization.

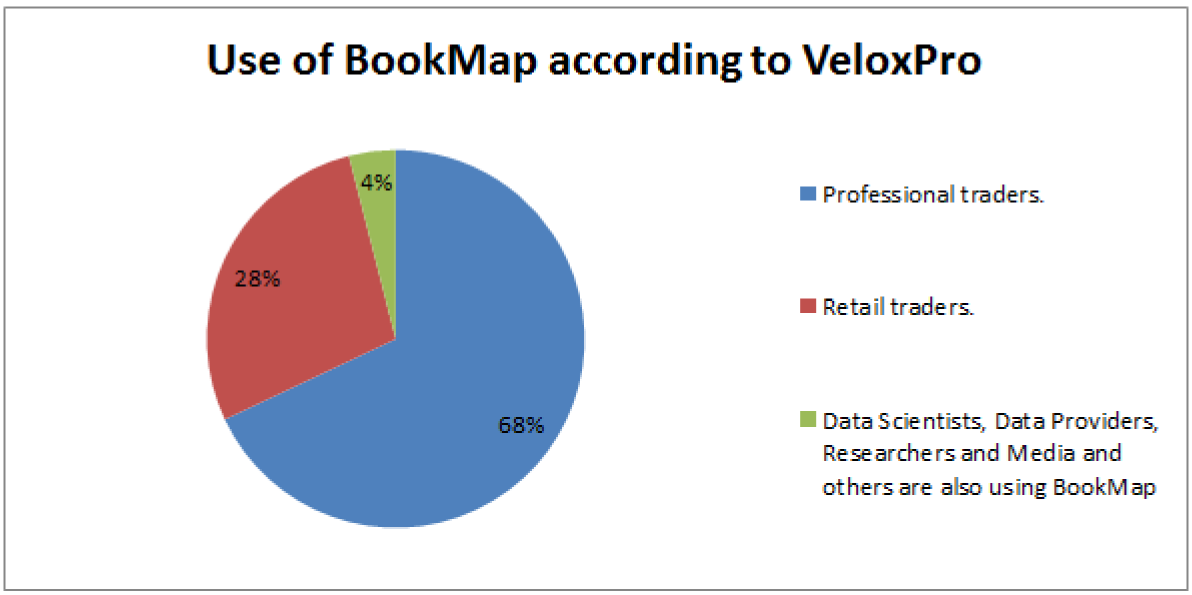


Figure 2-5.: Use of BookMap according to VeloxPro.

- Client BookMap xRay delivers a configurable heat map display that visualizes both real-time order flow and live trade analysis information combined with historical depth-of-market (DOM) data.
- Faster and deeper insight into live market dynamics and short-term price action due to its java-based Complex Event Processing (CEP) engine that handles millions of market data updates in real time, showing how the limit order book evolves over time.
- The user customizes the data provider. Traders of instruments with complete depth data can benefit from the insights of BookMap X-ray vision into the DOM.

Some examples of insights provided by VeloxPro demos are:

- Observing the strategy of Market Makers: sell orders are closer to current price than buy orders. This may indicate that the market maker has a net positive position and prefers to sell rather than buy to minimize risk. This in turn may affect the price.

- A significant imbalance in the order book depth pushes price downwards. A consistantly large buy order may serve as at least temporary support.
- Knowing the historical evolution of market depth at each price level can help evaluating the strength of a price level and distinguish between real liquidity and short term manipulating orders. They provide one video example of large buying orders consistently kept in place regardless of price change. This may indicate potential for real support and at least a short term price bounce.
- Video example of typical spoofing to acheive short term price action and be able to take the other side at favorable prices. Another example of spoofing type pushing price action: Large orders push the bid upwards to get filled on sell order(s) at more favorable terms. In that moment the pace of the cancel-place-and-replace buy orders is fast. This is probably an algorith trying to force a price.

Figure **2-6** is an example of the type of images provided by BookMap. The center of red circles represents the time and VWAP of trades. The radius of circles represents total executed volume. Another feature is separation of volume (the right column) according to trade initiator. Blue means the buyer initiated the trade.

Figure 2-6.: BookMap snapshot.

3. Efficient trend predictive LOB patterns dictionary building

This chapter presents an approximation to the extraction of informative patterns (bullish and bearish indicators) in real tick data from the Colombian Bulk currency Limit Order Book. This pattern exploration is performed on two scenarios: one composed of discretized events and other consisting of equally spaced samples from the discretized representation. Both scenarios are compared via accuracy in order to select the most supportive representation for investing.

Due to the massive amount of information generated in electronic markets, efficient methods and hash functions to handle and operate with this data are helpful. This work provides a methodological approach to manage this kind of data in order to extract information useful to profit generation.

The remainder of this chapter is structured as follows: In Section 1 the methodology is described. Section 2 presents the experimental setup. Finally, in section 3, results and discussion are provided.

3.1. Methodology

In this section a method for handling large amounts of data in Colombian Forex Markets is proposed. Real tick data of foreign exchange rate dollar-peso from March to May of 2012 were used in the experiments. LOB provides information about the time, the price and the volume of every request in the market.

This paper is based under the assumption that frequent Price-time-volume structures within the dataset are informative. In that vein, is reasonable to count every appearance of each pattern and store the number of times that the pattern is associated with a specific trend (in this case bullish or bearish), in order to calculate the probability of the pattern of being related with a trend. This process allows labeling frequent patterns in bullish or bearish patterns with the purpose of building a classifier (See Section 3).

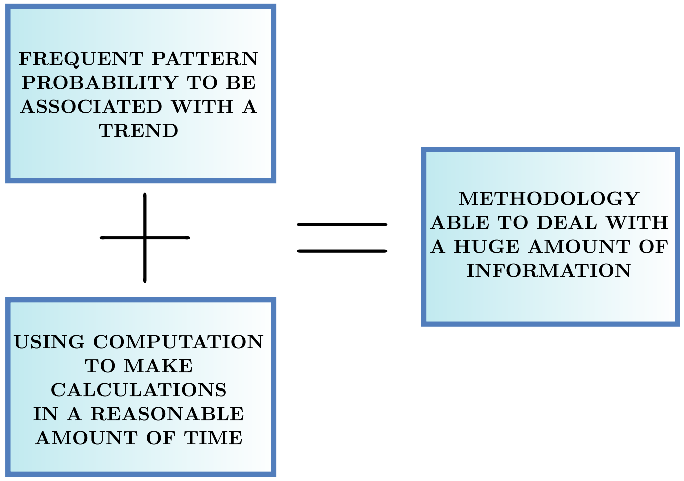


Figure 3-1.: Visual representation of the described methodology

Based on the idea that too frequent patterns are noisy patterns and that no frequent patterns doesn't provide useful information, a threshold was defined by both sides:

3.1.1. Dataset description

Two datasets with dollar-Colombian peso exchange rate from March to May of 2012 were available: one of them contained an orders summary minute by minute and the other one every ten minutes. For each one of those datasets, the following procedure was adopted:

- Each dataset was split into four subsets.
- Every subset was divided into two parts: one containing 30 % of the samples for training and 70 % for validation. This rare partitioning is due to the yearning of finding frequent patterns even in sets considered small. Furthermore, it works as a mechanism to avoid over fitting.
- The data was discretized in intervals of a) 1 and 10 minutes in time, b) 250000 dollars in volume and c) 5 cents in price.

The dataset arise from real data from the Colombian Forex Market, it was refined by Javier Sandoval[1] and it's a discretization of the order book. There are three matrices named Mv10min, Mv1min and Mv5min. These matrices were renamed as mejoresPuntas10min, mejoresPuntas1min and mejoresPuntas5min which contain the information of the best quotes for Mv10min, Mv1min and Mv5min respectively:

1. Each row in the matrix represents a level of quantization for the price.

[1] Research Professor in Universidad Externado de Colombia, Founder and Project Manager in AlgoCodex

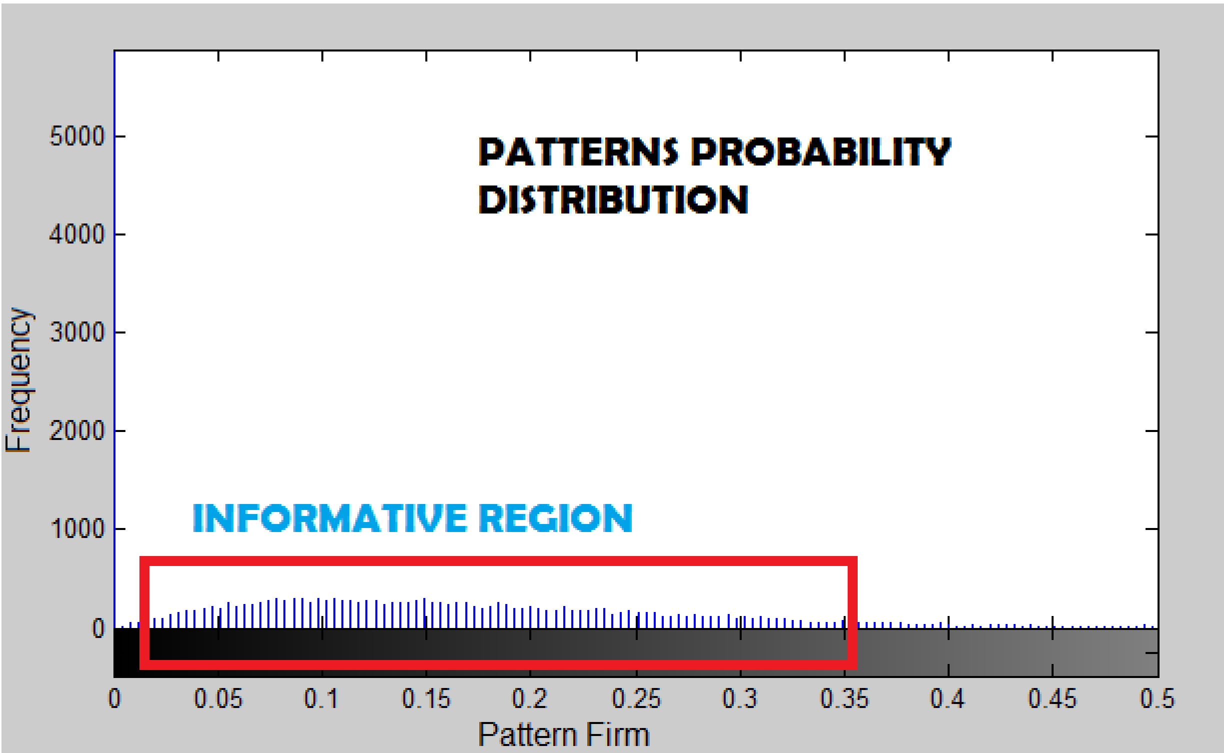


Figure 3-2.: Informative Region

2. Each column in the matrix represents a level of quantization for the time.

3. The value of each element represents the order's volume.

The input will be a humongous matrix and it's necessary to chop it into pieces of size(5015,5015), so the function mapPatterns(explorePatterns(choppedMatrix,width,deepness)) can find the patterns and assigns them to a certain firm (the sum).

Then, the function findObservedTrend(choppedVectorWithBestQuotes,patternWidth) will be applied in order to assign each pattern with a trend (bullish, bearish or no trend).

Later, the function countPatterns counts the number of different pattern and relates them with their most probable trend. It also saves the number of times that the pattern appears in the choppedMatrix (this is used in order to later determine a threshold for considering a pattern as frequent). Finally, findFrequentPatterns(frequency, threshold) select the patterns whose frequency is greater or equal to the given threshold.

Using a threshold of 10 and a probability of being associated with a certain trend greater than 0.55, 28 useful patterns were found, 19 were associated with a bearish trend and 9 were associated with a bullish trend (within the next 5 minutes). So, the frequency threshold was reduced to 5.

5 minutes timeslots experimental setup

There are available 4524 samples correponding to approximately 47 days. Cross validation will be made, using 4 folds:

- Each fold will have 1131 samples, 792 (70 %) for Training and 339(30 %) for validation.

With the best one the rest of the dataset will be tested to measure performance.

After that, other 4 dataset will be generated with the same size randomly to compare performance. This 5 minutes experimental setup was discarded later due to its performance.

3.1.2. Pattern exploration

In this section the pattern extraction basic process will be described, to do so, first a definition of pattern will be provided:

Definition 3.1 *A pattern in this scenario is a submatrix of size m by n, where m is the number of rows and n is the number of columns.*

In order to build a pattern, the value of consecutive samples was added. Clusters of $1 \times 10, 1 \times 20, 1 \times 40, 2 \times 10, 2 \times 20$ and 2×40 were summarized in a hash function defined as:

$$H = \sum_{i=a}^{a+n} \sum_{j=b}^{b+m} A_{ij} \tag{3-1}$$

Where A is the matrix which contains the volumes, a and b are the initial positions of each tile and, $n \in [1, 2]$ and $m \in S = \{10, 20, 40\}$.

The first step in order to find relevant patterns in the dataset is to make an exhaustive search of all the existing patterns in the dataset.

Given the dataset nature, the exhaustive search could be unfeasible with the available resources (one computer with OS: Windows 7 ultimate 32 bits, processor: Intel(R) Core(TM) i3 CPU M 330 @ 2.13GHz 2.13GHZ, RAM: 4 GB (2,93 available)). Three scenarios are considered:

1. In the whole dataset, there are not repeated patterns.
2. In the dataset there are repeated patterns, but they are not associated strongly with a trend.

3. In the dataset there are repeated patterns and they are associated strongly with a defined trend.

This work is conveyed under the hypothesis that the dataset is on the third scenario. The brute force algorithm proposed for building the patterns catalogue is as follows:

The dataset should be cut off in order to eliminate those frequent patterns that are not informative (remove NaNs or zeros).

```
ALGORITHM 1 BasicPatternExtraction(dataMatrix,numberOfRowsPattern,
                                   numberOfColsPattern)
Input: Matrix with the order book information dataMatrix.
Output: Matrix with the replacement of every pattern with
        an assigned number matrixLabels.

begin
 i=1, j=1, countOfPatterns=0;, PatternLabel=null,
  PatternsCatalogue=null.
 for every pattern in dataMatrix
 if(dataMatrix(i:i+(n-1),j:j+(m-1)) is not in
          PatternsCatalogue)
begin
add dataMatrix(i:i+(n-1),j:j+(m-1) in
             PatternsCatalogue
countOfPatterns= countOfPatterns+1
PatternLabel(i,j)= countOfPatterns
end
else
PatternLabel(i,j)= index of PatternsCatalogue
         where dataMatrix(i:i+(n-1),j:j+(m-1) appears.
end
 return PatternLabel, PatternsCatalogue and countOfPatterns.
end
```

Another requirement in order to extract relevant patterns is to associate them with trends, so it is possible to modify the previous algorithm for calculating trends simultaneously:

```
ALGORITHM 2 BasicPatternExtractionWithTrends(dataMatrix,numberOfRowsPattern,
            numberOfColsPattern, tradingPricesVector)
Input: Matrix with the order book information (price, time and volume).
Output: Matrix with the replacement of every pattern with an assigned number
```

```
        matrixLabels.

begin
 i=1, j=1, countOfPatterns=0, PatternLabel=null, PatternsCatalogue=null.
 for every pattern in dataMatrix
Trend= comparation between tradingPricesVector(j:j+(m-1)) and
       tradingPricesVector(j:j+(2m-1))

(Trend=-1 if tradingPricesVector(j:j+(m-1))
is greater, Trend=1 if
tradingPricesVector(j:j+(m-1))
is lower, and zero otherwise).

 if(dataMatrix(i:i+(n-1),j:j+(m-1)) is not in PatternsCatalogue)
begin
add dataMatrix(i:i+(n-1),j:j+(m-1) in PatternsCatalogue
countOfPatterns= countOfPatterns+1
PatternLabel(i,j)= countOfPatterns
end
else
PatternLabel(i,j)= index of PatternsCatalogue where
          dataMatrix(i:i+(n-1),j:j+(m-1) apperars.
end
 return PatternLabel, PatternsCatalogue and countOfPatterns.
end
```

Given the amount of possible patterns (circa $6{,}5 \times 10^7$), it is necessary to find efficient ways to convey this mining task. The idea of frequent itemset was proposed by [Agarwal et al. 1993]. In 1994 [Agrawal R, Srikant R] proposed the «A priori» method for mining patterns, using the downward closure property, which states that «A k-itemset is frequent only if all of its sub-itemsets are frequent». Several improvements and generalizations of this algorithm have been done [Savasere 1995, Toivonen 1996, Brin 1997, Cheung 96, Park 95, Agrawal schafer 96, Geerts 2001]. Nevertheless, the A priori algorithm can generate a huge number of candidate sets or scan several times the database looking for patterns.

In 2000, [Han et al] proposed a FP- growth method which doesn't use candidate generation. The algorithm orders patterns by frequency descending order and then, reduces it into a frequent pattern tree with the association information. The suffix pattern is concatenated with the conditional FP-tree patterns, producing the trees' growth. It is used to mine long patterns because reduces the search time. This algorithm was discarded because this work is not intended to find long patterns, so the cost of building the tree is not justified.

The CLASS Transformation (Eclat) algorithm proposed by [Zaki 2000], could be useful in this work. Every time slot can be considered as a transaction. The items bought in that transaction would be the patterns found in the column corresponding to that time slot.

When closed itemsets are used, the scalability and interpretability of the mining task is better, but is hard to verify if a pattern is closed.

Sequential pattern mining [Agrawal and Srikant (1995)] is the mining of ordered events; each itemset is a set of events which occurs in the same timeslot. Generalized Sequential Patterns [Agrawal and Srikant (1995)] include time constraints, a sliding time window, and user-defined taxonomies.

Afterwards, frequent patterns exploration was driven following two criteria: a frequency threshold and a probability threshold of being associated with a bullish or a bearish trend. Those patterns which satisfied both standards were labeled as bullish or bearish patterns, respectively. Based on those patterns, a classifier was built and its performance was measured in the remaining subset. This procedure was executed in both, 1 minute and 10 minutes subsets. This method is referred as heat map method in the results section.

With the aim of exploring different levels of resolution, the datasets were sampled using Haar Wavelet Transform four times for 1 minute patterns and two times for 10 minutes patterns. Both sets of coefficients (those from the average and those from the difference) were employed. Mallat describes this procedure as follows:

$$f(x) = f(x_1, x_2) : \left\{ \psi^k_{j,n}(x) = \frac{1}{2^j} \psi^k \left(\frac{x - 2^j n}{2^j} \right) \right\}_{j \in Z, n \in Z^2, 1 \leq k \leq 3} \tag{3-2}$$

Where:

$$\psi(t) = \begin{cases} 1 & if \quad 0 \leq t < \frac{1}{2}, \\ -1 & if \quad \frac{1}{2} \leq t < 1, \\ 0 & otherwise \end{cases} \tag{3-3}$$

$\psi(x)$ denotes the mother wavelet, 2^j corresponds to the scale, $2^j n$ the translation, k to the direction, $x1$ and $x2$ are the row and the column in the matrix.

After applying this preprocessing method, new frequent patterns were extracted and threshold criteria were applied. Finally, performance was evaluated.

3.1.3. Performance measurement

Performance was evaluated via accuracy. Since different frequent patterns pointing out contradictory trends could emerge several times in one sample or even, the same pattern could arise more than one time in that sample, a voting system was introduced. In this system, every pattern instance votes for the trend to which it is associated. Lastly, the prediction for every sample is decided by the result of the sum of the votes.

The experimental setup outlined above is used to determine the optimal parametrization of the model.

Limit Order Book is a variable length list where transactions occur in non-uniform time intervals. In the instant when transactions occur, the spread is zero. As shown in Figure **1-3**, for a given time slot, information of prices and the cumulative volume for each price are provided.

This paper proposes a tool to facilitate a human trader locating visual patterns. The proposed visualization is built as follows:

1. Book event's discretization: Events which happen every minute are aggregated. Increases in the x coordinate indicate progress over time.

2. Trading volume quantization: Volumes within a range of prices in an instant of time are cumulated; higher volumes are indicated with a lower level of Lightness.

3. Price quantization for each time unit. An increase in the y-axis in the picture indicates a higher price.

3.2. Experimental Setup

In this section a method for handling large amounts of data in Colombian Forex Markets is proposed. Real tick data of foreign exchange rate dollar-peso from March to May of 2012 were used in the experiments. LOB provides information about the time, the price and the volume of every request in the market; this information was summarized every minute in a price range of 120 pesos in the best quotes each 20 cents. Volumes were quantized in levels of USD 250,000, which is the minimum trading volume in this market. The maximum volume observed in an order in the analyzed period was 43500000 USD. The maximum price observed was 1862.6 COP and the minimum was 1742.2 COP.

This paper is based under the assumption that frequent Price-time-volume structures within the dataset are informative. Linares, Gonzalez et Hernandez [58] identified individual basic shapes on time series in order to build active trading strategies based on forecasting. In that vein, it is reasonable to count every appearance of each pattern and store the number of times that the pattern is associated with a specific trend (in this case bullish or bearish), in order to calculate the probability of the pattern of being related with a trend. This process allows labeling frequent patterns in bullish or bearish patterns with the purpose of building a classifier.

The first reference to the Bag of Words method appears in [24] when Harris states that *"it is possible to define a linguistic structure solely in terms of the "distributions" (= patterns of co-occurrences) of its elements. There is no parallel meaning-structure which can aid in describing formal structure. Meaning is partly a function of distribution."*. Later, in 2003, Sivic et al. [54] present an analogy between text and image retrieval for video retrieval where the construction of the visual vocabulary is made quantizing descriptors in clusters (using k-means) extracted from a fragment of the film. Lopez-Monroy, Gomez et al. [40] show the general process for generating a bag of visual words from a set of images.

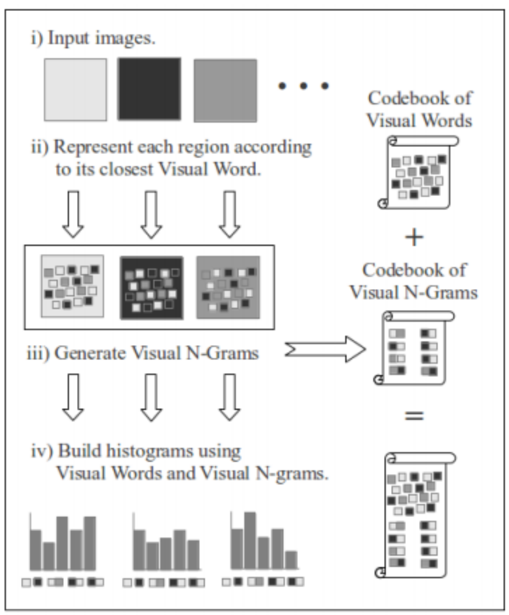


Figure 3-3.: Image Representation through Bag-of Visual-Ngrams, extracted from [40]

Definition 2 *In this context, a pattern is a matrix of market events, an aggregation of volumes spatially organized by price and date.*

3.2.1. Limit Order book

Cont et al. [11] define the order book as a grid of price ticks, where:

The ask price is defined as:

$$p_A(t) = inf\,\{p = 1, ..., n, X_p > 0\} \wedge (n + 1). \tag{3-4}$$

The bid price is defined as:

$$p_B(t) = sup\,\{p = 1, ..., n, X_p < 0\} \quad \vee \quad 0. \tag{3-5}$$

The distance from the best price is called depth. For the bid side the volume at a certain distance i is given by:

$$Q_i^B(t) = \quad \begin{cases} X_{P_A(t)-i}(t) & 0 < i < P_A(t) \\ 0 & \quad P_A(t) \leq i < n \end{cases} \tag{3-6}$$

For the ask side:

$$Q_i^A(t) = \quad \begin{cases} X_{P_B(t)+i}(t) & 0 < i < n - P_B(t) \\ 0 & \quad n - P_B(t) \leq i < n \end{cases} \tag{3-7}$$

The spread is the difference between the best price in the ask side and the best price in the bid side:

$$p_s(t) \equiv p_A(t) - p_B(t) \tag{3-8}$$

Finally, a trader is the agent which produce movements in the order book.

3.2.2. Market behaviors

Definition 2 *A trend is the sign of the difference between the current and the previous observation.*

In this case the observed phenomenon is the currency price. Two possible market behavior are specified for this technique:

$$t\,(t) = \begin{cases} 1 \quad if \quad p_o(t) - p_o(t-1) > 0 \\ -1 \quad if \quad p_o(t) - p_o(t-1) < 0 \\ 0 \quad otherwise \end{cases} \tag{3-9}$$

A trend is said to be bearish when $t(t) = -1$ and bullish if $t(t) = 1$.

3.2.3. Pattern exploration

In [49], Rajaraman et al. define a Hash function as *a function h which takes a hash-key value as an argument and produces a bucket number as a result.* In order to build a pattern, the value of consecutive samples was added. Clusters of $1 \times 10, 1 \times 20, 1 \times 40, 2 \times 10, 2 \times 20$ and 2×40 were summarized in a hash function H.

This function is used before the pattern extraction process to compress a whole pattern in a single value. In this way, it is possible to deal with smaller matrices. H is defined as:

$$H = \sum_{i=a}^{a+n} \sum_{j=b}^{b+m} A_{ij} \tag{3-10}$$

Where A is the matrix which contains the volumes, a and b are the initial positions of each tile and, $n \in [1, 2]$ and $m \in S = \{10, 20, 40\}$.

Afterwards, frequent patterns exploration was driven following two criteria: a frequency threshold and a probability threshold of being associated with a bullish or a bearish trend. Those patterns which satisfied both standards were labeled as bullish or bearish patterns, respectively. Based on those patterns, a classifier was built and its performance was measured in the remaining subset. This procedure was executed in both, 1 minute and 10 minutes subsets. This method is referred as heat map method in the results section.

3.2.4. Performance measurement

Performance was evaluated via accuracy. Since different frequent patterns pointing out contradictory trends could emerge several times in one sample or even, the same pattern could arise more than one time in that sample, a voting system was introduced. In this system, every pattern instance votes for the trend to which it is associated. Lastly, the prediction for every sample is decided by the result of the sum of the votes.

The experimental setup outlined above is used to determine the optimal parametrization of the model.

3.3. Results and Discussion

3.3.1. Results

The results of the parameter search for the proposed approach are shown in Figures I and II.

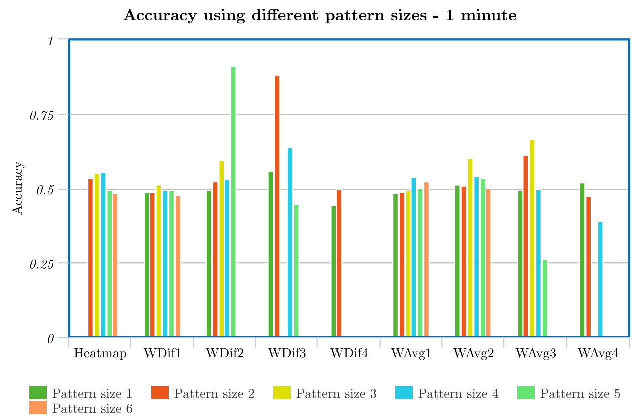


Figure 3-4.: Predictor's accuracy using different 1 minute pattern sizes.

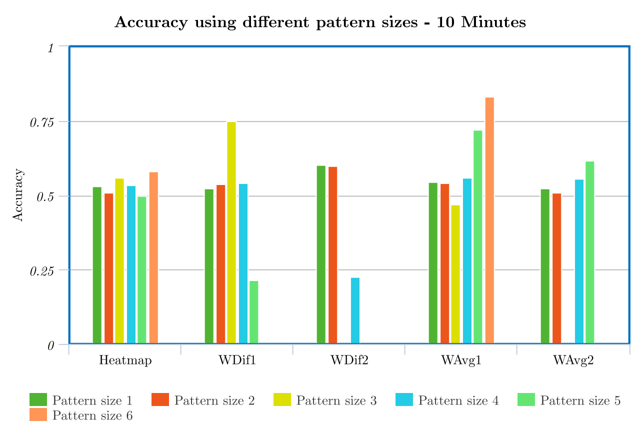


Figure 3-5.: Predictor's accuracy using different 10 minutes pattern sizes.

In some cases, during experimentation, no frequent patterns were found in accordance with the defined thresholds for pattern frequency and probability of being associated with a specific trend. Those patterns selected as frequent should, simultaneously, be associated with a trend with a probability higher than 0.55 and should appear more than five times in the corresponding subset. In those cases in which no frequent patterns were found, the accuracy was infinity, hereby, those accuracy values were discarded. For that reason, some bars are missing in Figures I and II to avoid misinterpretation.

The average number of frequent patterns by subset and pattern size was 70. It was found that frequent patterns lost its ability to predict trend over time, supported on this result, the future development of an adaptive method to overcome this difficulty is suggested.

The best performance for this method was provided by 1 minute frequent patterns, using as pre-processing method the application of the Haar Wavelet Transform twice, using those coefficients corresponding to the difference between consecutive values in the volumes matrix. In this case, pattern size 5 (2x20 items tile) were used. The accuracy achieved was 0.91 in the best case.

For 10 minutes frequent patterns, best performance was achieved using Haar wavelet transform once with the coefficients corresponding to the average between consecutive values in the volumes matrix and, pattern size 6 (2x40 items tile). The accuracy accomplished in this way was 0.8333 in the best case.

Overall, the pattern size which provided the highest average accuracy was pattern size 3 (1x40 items tile) for 1 minute patterns and pattern size 6 (2x40 items tile) for 10 minutes patterns. The pre-processing method which lead to the best average accuracy was: for 1 minute patterns, three iterations of wavelet transform using difference coefficients, in this scenario the accuracy was 0.632175 in the best case. For 10 minutes frequent patterns, the best performance was obtained by using Haar wavelet transform once, selecting those coefficients corresponding to the average. This pre-processing method produced an accuracy of 0.61168 in the best case.

Figures **3-4** and **3-5** compare best performance by pattern size among the different pre-processing choices: heat map, Haar wavelet transform differences and average coefficients until 4 iterations for Figure **3-4** and until 2 iterations for Figure **3-5**, every color represents a different pattern size.

3.3.2. Discussion

With an optimal parametrization, the performance is better compared with a random guess (acc. 0.91 vs. 0.5). However the performance does not keep that level from one subset to another, suggesting the existence of seasonal patterns. This phenomenon is described by Jiang et al. in [32]. To test this hypothesis, two subsets with global opposite trends were identified using moving average and the performance of 400 different pattern sizes was evaluated using only the hash function. In this scenario, 35 pattern sizes achieved accuracy higher than 0.6 in the first subset, when tested in the opposite trend, only 4 kept their accuracy. Results are shown in figure **3-6**.

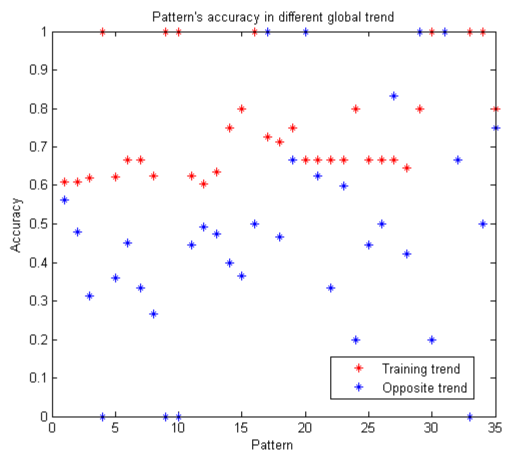


Figure 3-6.: Patterns accuracy within and outside a global trend.

In order to assure this method's scalability, testing in larger datasets is required. Figures **3-4** and **3-5** show the result of applying this methodology to obtain a strong classifier gathering several weak classifiers using a simple voting system.

3.4. Adaptive Method

There is a need of identifying when a labeled pattern starts to lose the ability to predict the trend. For this reason we presented an online method for informative frequent patterns identification.

1. Make an initial training stage in which a general market trend is identified.

2. Every time that a new sample arrives for classification, recalculate the probabilities associated to each trend for the patterns found in the sample.

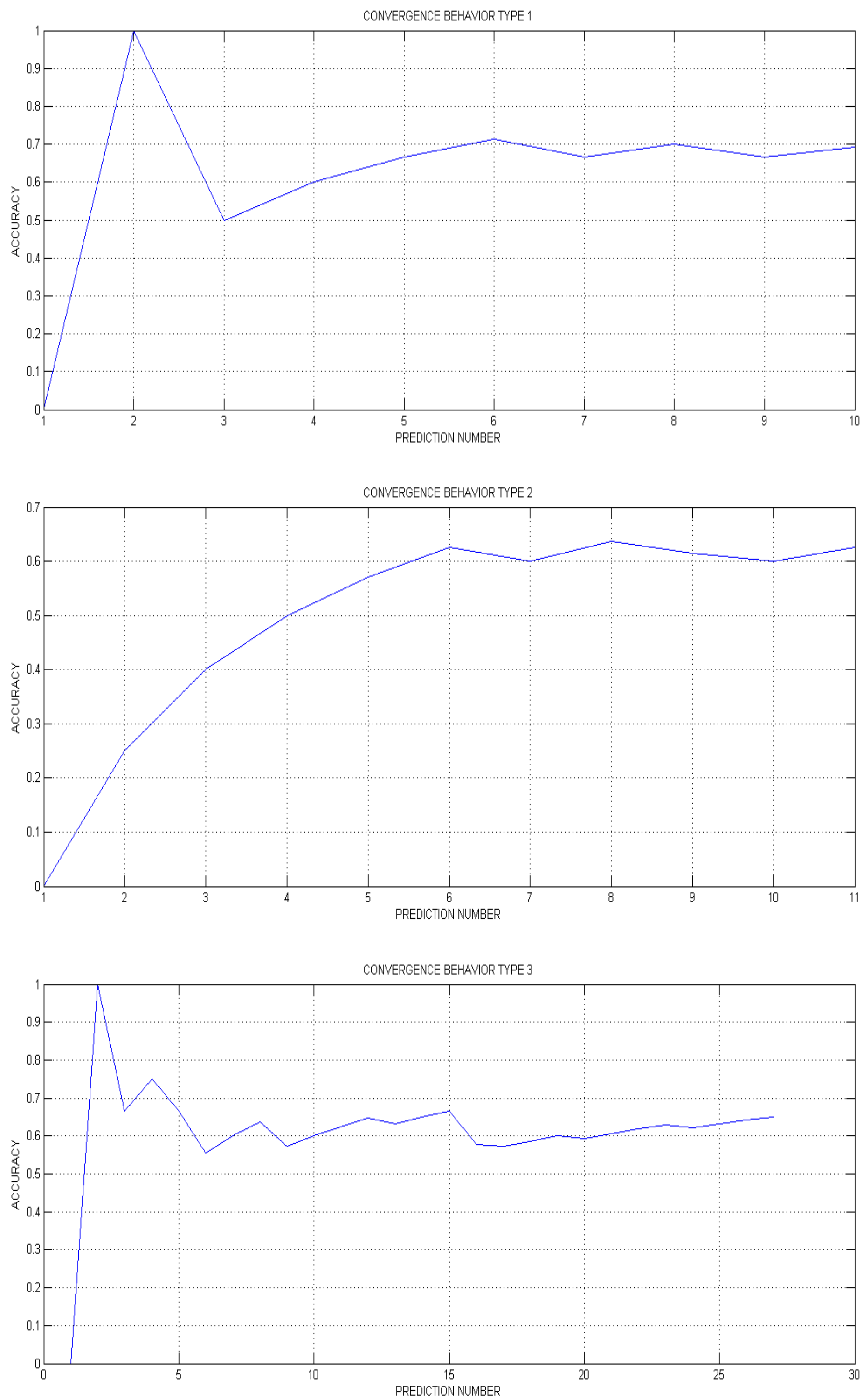


Figure 3-7.: Different convergence behaviors found for the adaptive method.

3. When the count for a new pattern apparitions reaches an specific amount and its probability of being associated with a determined trend surpasses a threshold, add that pattern to the informative frequent patterns' dictionary.

4. When the count of misclassifications for a pattern from the informative frequent patterns' dictionary reaches an specific number or when its probability of being associated with a determined trend falls to a determined threshold, remove that pattern from the informative frequent patterns' dictionary.

The previous procedure was tested for 300 different patterns' sizes, producing 29 configurations where patterns highly associated with a trend were found and whose accuracy was above 0.6. From those 29 configurations, 10 produce 5 or more predictions on the dataset. Three different convergence behaviors were found as it could be observed in figure **3-7**.

4. Bags of trend predictive LOB patterns using wavelets

4.1. Use of Wavelets in the representation

Wavelets allow a multi-resolution study of the images and a frequency analysis of them. Haar wavelets were selected due to ease of computation and to the speed at which the coefficients can be calculated. Figure **4-1** shows the result of applying four times wavelet transform over an image of the order book, keeping only averages. The level of compression without shape loss can be noted in this image.

Figure 4-1.: Example of image produced by four levels of compression using haar wavelet transform.

In figures **4-2** and **4-3**, the effect of applying three times Wavelet Transform, keeping coefficients corresponding to the differences, can be perceived. It can be noted how the image high frequency component is filtered (the two rightmost sub images in each image were amplified in order to preserve visibility). It is expected to find patterns in the high frequency noise which provide predictive information about price changes.

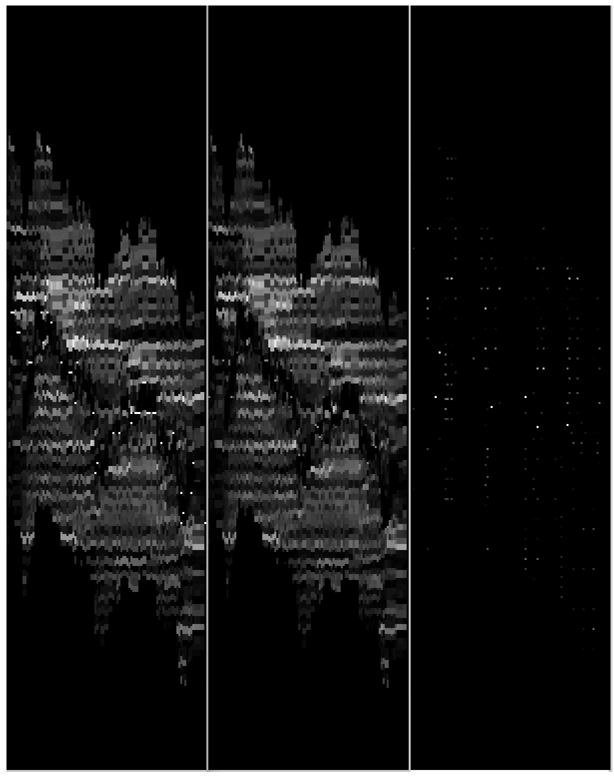


Figure 4-2.: Example of order book visualization using wavelets based approach (filtering).

This representation allowed observing graphically long trading periods at a glance. For example, figure **4-4** provides the information of prices in time of three months of trading. The image above, in the figure, keeps high frequency information and, the image below shows the average of the orders' volumes. Recall that x axis represents time and y axis, price.

4.2. Wavelet transform

With the aim of exploring different levels of resolution, the datasets were sampled using Haar Wavelet Transform four times for 1 minute patterns and two times for 10 minutes patterns. Wavelet Transform is used for sampling and compressing the dataset to work with even smaller matrices.

For ease, every matrix is treated like an image. Mallat [41] defines wavelets for images as:

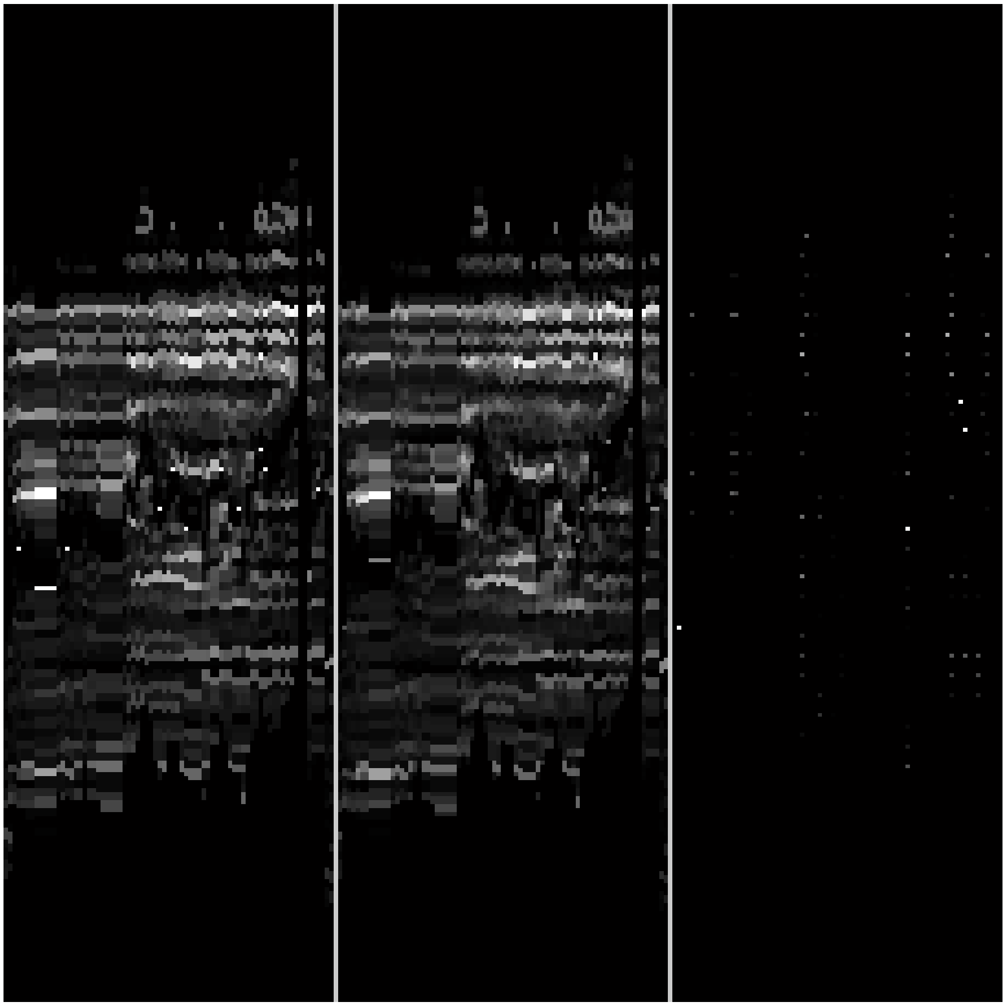


Figure 4-3.: Example of order book visualization using wavelets based approach (filtering).

$$f(x) = f(x_1, x_2) : \left\{ \psi^k_{j,n}(x) = \frac{1}{2^j}\psi^k\left(\frac{x - 2^j n}{2^j}\right) \right\}_C \tag{4-1}$$

Where:

$$C = j \in Z, n \in Z^2, 1 \leq k \leq 3 \tag{4-2}$$

$$\psi(t) = \begin{cases} 1 & if \quad 0 \leq t < \frac{1}{2}, \\ -1 & if \quad \frac{1}{2} \leq t < 1, \\ 0 & otherwise \end{cases} \tag{4-3}$$

$\psi(x)$ denotes the mother wavelet, 2^j corresponds to the scale, $2^j n$ the translation, k to the direction, $x1$ and $x2$ are the row and the column in the matrix.

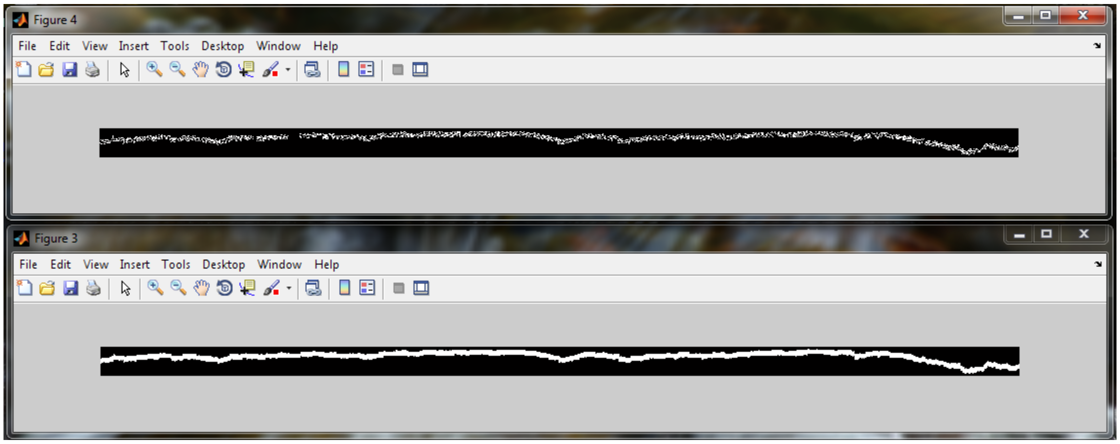


Figure 4-4.: Three months of trading using one minute resolution as basis for the image construction.

Both sets of coefficients (those from the average and those from the difference) were employed. After applying this pre-processing method, new frequent patterns were extracted and threshold criteria were applied. Finally, performance was evaluated.

5. Bags of trend predictive LOB patterns using clustering

5.1. Experimental Setup

This paper is based under the assumption that frequent price-time-volume structures within the dataset are informative. Linares, Gonzalez et Hernandez [58] identified individual basic shapes in time series in order to build active trading strategies based on forecasting. In that vein, is reasonable to count every appearance of each pattern and store the number of times that the pattern is associated with a specific trend (in this case bullish or bearish), in order to calculate the probability of the pattern of being related with a trend. This process allows labeling frequent patterns in bullish or bearish formations with the purpose of building a classifier.

5.2. Cluster classification

In order to introduce the idea of similarity between patterns, k-means was applied over the patterns matrix built over the order book. The procedure conveyed is as follows:

- Time-price-volume matrix was chopped and adjusted for working in regions near to the spread.
- The matrix was traversed using tiles of 30x30 elements.
- Every pattern was assigned to a cluster using k-means.
- In a new array, the size of each cluster was registered in order to keep track of the frequency at which patterns were assigned to every cluster.
- Finally, each cluster was labeled with the observed trend when the probability of being associated to it was higher than 0.55.

5.3. Clusters approach performance

The uppermost picture in figure **5-2** shows the centroid patterns assigned to each cluster, and the bottommost picture presents a matrix with example patterns assigned to each cluster. A sample of these kind of patterns can be observed more closely in figure **5-1**. Figure **5-3**provides the histogram that represents the size of each cluster.

Figure **5-3** introduces the performance of each cluster: in red the probability of the cluster patterns of being associated with a bearish trend, and in green, the probability of that cluster patterns of being associated with a bullish trend.

Figure 5-1.: Example of patches associated with clusters.

Figure **5-5** shows the cumulated return (COP amount by every invested dollar) for 6 months of trading using the cluster approach.

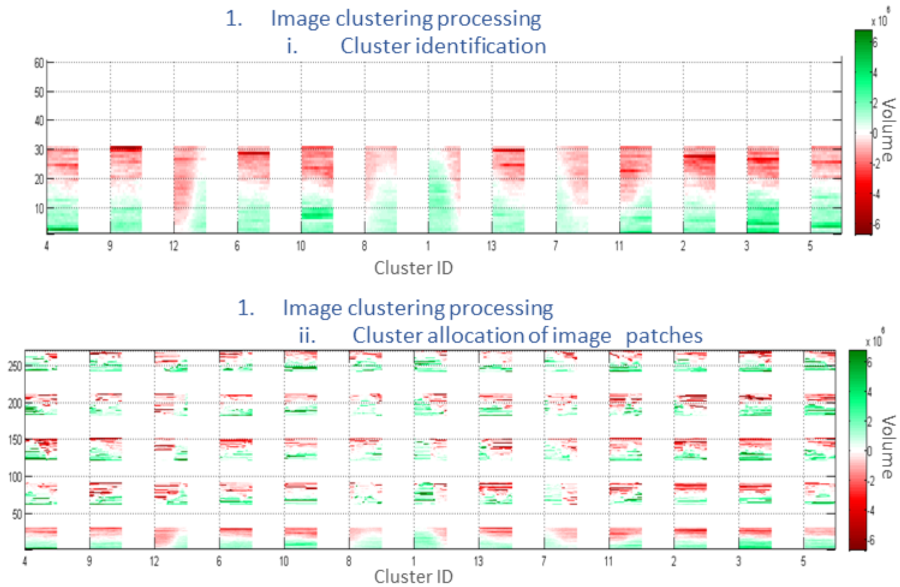


Figure 5-2.: Clusters matrix and centroids.

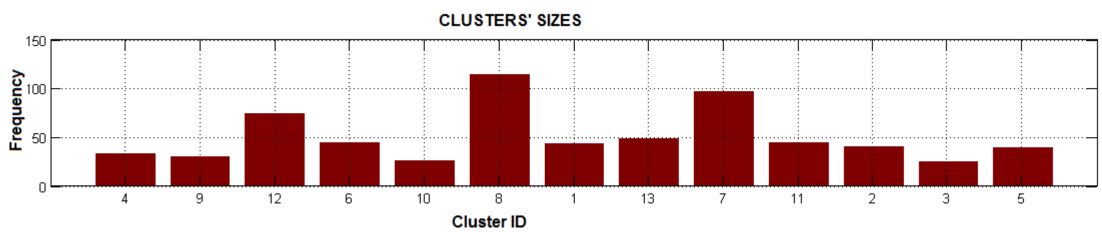


Figure 5-3.: Frequency at which patterns are associated with a certain cluster.

Figure 5-4.: Clusters and their performance.

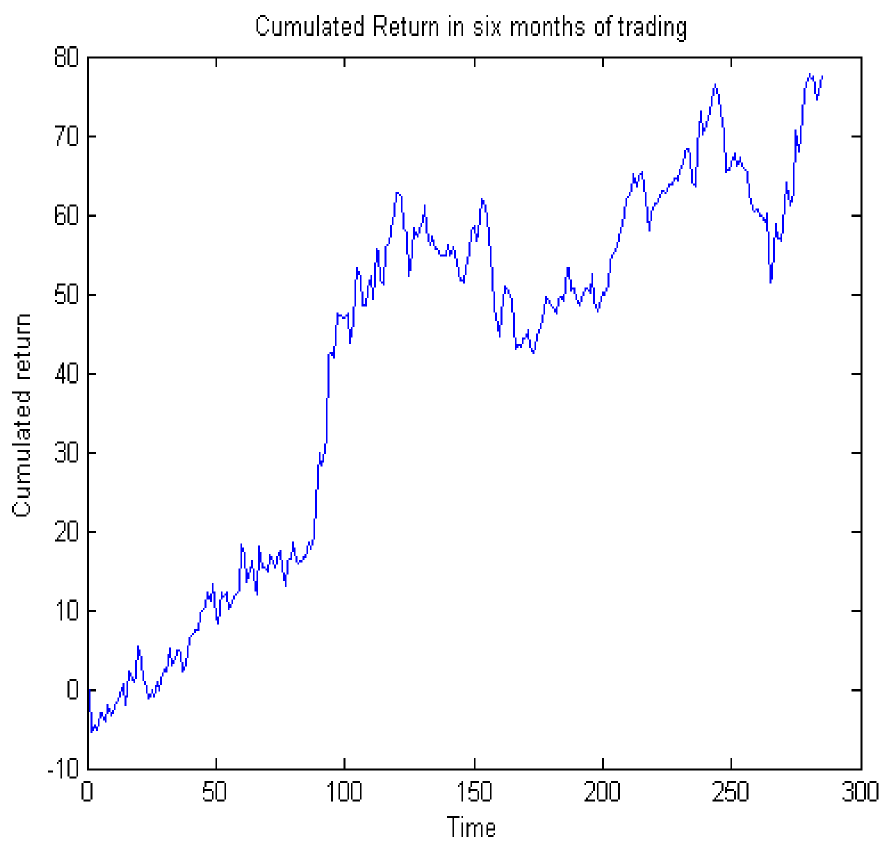


Figure 5-5.: Cummulated retrurn usd cop in six months.

6. Effective trading strategy based on bags of trend predictive LOB patterns using clustering

The Limit Order Book is composed of two registers which list market agents buy and sell intentions concerning a specific stock [65]. Every buy or sell intention is known as an order. These prices lists are sorted as follows: the buy list is in descending order and the sell list is in ascending order. The first price in each list is named “best quote” and the difference between them is called “spread”. There are several types of orders, which will be described later in this chapter. There are variations, depending on the market and order types, but in general, when buy and sell intentions agree, the orders which represent them are executed and disappear from the book. Market agents are also able to modify or cancel orders whenever within the trading operation period.

Colombian Order Book is a non-automatic, non-match market. In other words, orders are executed only by human traders’ agreement in an electronic platform. There are several kinds of operators and not all of them are allowed to modify the order book. The ones with permission to do so, access to the platform and register their orders.

The order book centralizes the market information; making trading, control and surveillance easier. The key point is to raise the transparency up in order to have a more efficient market. More about market transparency and efficiency in [39]. Data is generated by those market agents which are allowed to trade. They set a price and a volume for the stock they are willing to trade in a given time point. Also the orders’ modifications and cancellations made by those agents produce changes in the order book.

The order book provides the user with insights about the market state in a time point [27], [46], [47], [48], [50], [59] and [60]. It informs about volume, price and liquidity and it’s useful to anyone wishing to trade a stock in a market. Order book users are mainly brokers, stock brokerage firms and banks and not all of them are allowed to operate in it. The order book is important for traders as it contains complete market information and its analysis helps in the decision making process for systematic profit generation. It’s an attractive data source because it “contains the maximal amount of information about the state of the market that is available to traders" [16].

The idea behind using this tool (the order book) is to identify liquidity patterns (sub images in the book) which probably provide some level of certainty about the future price in a given time horizon.

In accordance with the aforementioned, one question arises: What should be considered as a liquidity pattern? Following Bookmaps' convention, a particular gray-scale squares combination provides information about the trend direction. Assuming this, the next question is: How useful liquidity patterns can be identified? The systematic way of answering this issue is by means of parameter exploration for providing evidence about the proper pattern size. This is, how many squares should shape a pattern? Which is the suitable time window for that observation? The parameter exploration should also inform about the time horizon in which the prediction is valid. This work focuses on short term prediction (in the order of minutes) due to the dynamic nature of the Colombian USD/COP order book in that time scale.

However, how to achieve a systematic parameter exploration with varying pattern size? In signal processing this problem is addressed by means of multiresolution analysis (MRA). There are several tools used for MRA. For instance: register banks, sampling, transformations in order to study the problem in other domains, etc. In particular, this work faces this question using two approaches: the first one is varying window size to identify informative configurations; the second one is the use of the Haar Wavelet Transform (HWT) for images [41]. HWT conducts a multiresolution analysis keeping spatial relationships and examining high frequency phenomena in a straightforward and easy to compute way. This tool is frequently used in the visual computation field [15],[41]. It's commonly used in order to identify wavelets, but in this case, classification isn't made using the waves provided by a Wavelet Transform but using the Bag of Words technique. Later, that approach is refined to cluster patterns with similar characteristics with an specific trend direction. To do so, the k-means algorithm was employed. Generally speaking, there were two kinds of exploration: first, using adjacent windows because, despite of passing over temporal information, it is a computationally less intensive alternative; the other exploration was made using sliding windows because it provides more comprehensive temporal information regardless of being more intensive computationally.

6.1. Dataset description

This study is based on the trading data of the manual market of Colombian Forex Market from May 2nd until October 10th during the year 2014. For being more informative, the order book section closer to the spread was chosen for the exploration[48]. This section accumulates most of the volume in the order book. In particular, USD-COP market was selected

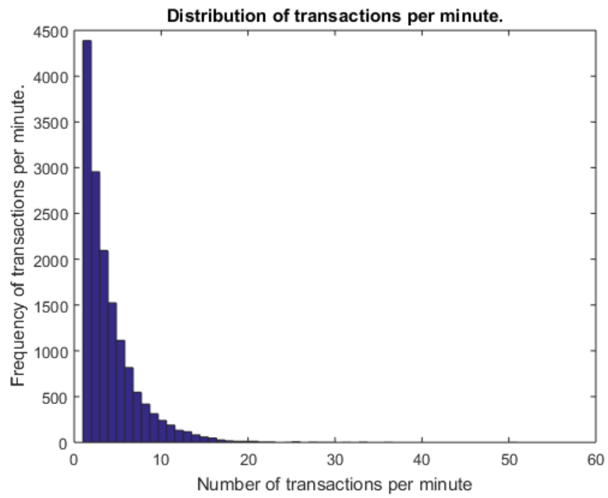


Figure 6-1.: Six months' transactions frequency distribution.

because it is one of the most traded and liquid products available in the Colombian Stock Exchange [19]. The data was preprocessed as a quantized price-time-volume matrix, similar to the one which allowed the heatmap representation in [10]. For further information about the implementation details, see [12].

Compared with the New York Stock Exchange (NYSE) [35], [47] and the London Stock Exchange (LSE) [27], the Colombian Bulk Currency Market is slow. To illustrate this point, figure **6-1** presents the distribution of transactions per minute in six months of trading in the USD-COP market. The average number of transactions per minute is 3.6809 with a variance of 11.0419. Figure **6-2** depicts the distribution of the number of modifications per minute in the same period. The average number of modifications per minute is 36.9665 with a variance of 332.7447. NYSE an LSE surpass those amounts per minute in less than a second [45],[56].

6.2. Experimental Setup

The main goal of this experimental setup is to find trend predictors that when trading, generate non-negative returns with a relatively small variance and that are simultaneously frequent in both, validation and testing sets. The performance using a subset exhibiting a clear global trend is expected to be better than the performance produced by using the whole dataset with mixed local trends [7],[13].

6.2.1. Tools

The original spread centered data set was split into rows or columns every n elements and then reorganized in $n \times n$ elements rows (where $n = 30$). Vectorization was used in order to

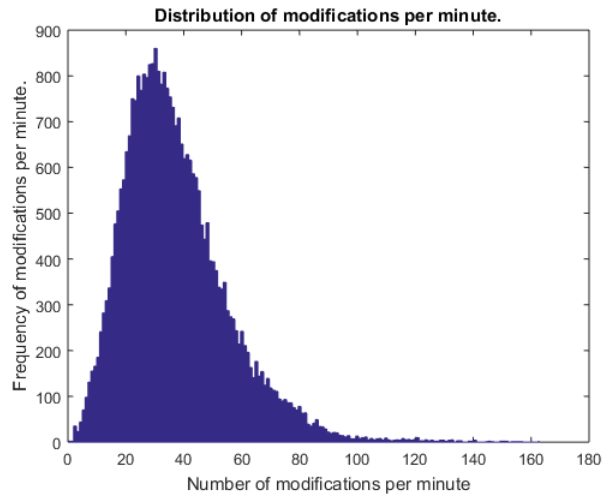


Figure 6-2.: Six months' modifications frequency distribution.

improve time performance when generating the corresponding row and column based datasets. When using adjacent window, a dataset reduction using only n elements surrounding the spread was made, and the new set was reorganized in $n \times n$ tiles. For sliding window, a function of the Matlab image processing toolbox for window sliding was applied and then the same reduction used for adjacent window was made.

The following tools were used for the described tasks: First, the Matlab k-means implementation was used for centroids and clusters' membership generation. Secondly, a Haar Wavelet Image Transform implementation was coded as defined by Mallat [41], subtracting and averaging adjacent pairs keeping both results separately. Finally, for pattern trend classification, the Bag of Visual Words method was employed for both cases: training using the whole dataset and training using a subset exhibiting a global trend.

6.2.2. Experiments description

After a data preprocessing stage, several configurations were generated from the original dataset according to the following guidelines: Vertical arrangement, Horizontal arrangement, Haar Wavelet Transform arrangement, sliding window arrangement and adjacent window arrangement. Subsequently, every generated configuration was restructured as a matrix where each row were an input sample (feature vector) for the k-means classifier.

At that point, each generated matrix was split into three subsets: training set, validation set, and testing set. Afterwards, every training set was used as input for the matlab k-means algorithm in order to obtain the clusters centroids and a label for every sample. Then, each sample's distance to every centroid was computed and every sample was labeled according to its closest centroid. The same process was conducted for each sample in every testing set of the generated sets.

Thereafter, corresponding observations to different time horizons (2, 4, 6, 10, 16 and 30 minutes) were calculated for each dataset. Then every sample (represented by its label) was matched with its respective observation.

Thus, the classification rules were computed: as a first step, the number of times that every pattern was associated to each trend (bullish or bearish) was accumulated. Secondly, every pattern with a probability higher than a certain threshold (in this case 0.6) of being associated with a trend was labeled as a predictor of that trend. That label is translated into a selling or buying signal as appropriate.

Validation and testing sets were passed through for trading signals generation according to the label assigned to each cluster. Finally, return was calculated for each pattern selected as trend predictor in every generated subset. Returns mean and variance for every subset were recorded.

6.2.3. Performance measurement

In order to achieve a comparison between methods it is necessary to find an appropriate way to present the results. There are two important factors to be measured: first, the potential profit generated for each method and second, the risk level associated to each method. The mean of the returns allows to win insights about each method's potential profit. With the purpose of providing a reliability measure, the returns variance was normalized: it is 1 when the variance is zero and tends to zero as the variance go to infinite:

$$Reliability = \frac{1}{variance + 1}. \tag{6-1}$$

Hence the higher this number, the better the returns' reliability. The classifier precision for each subset was also registered.

6.3. Results and Discussion

6.3.1. Cluster classification

In order to reduce the liquidity patterns search universe, a cluster algorithm over the samples was used as follows: a time-price-volume matrix was built with the 30 closest-to-the-spread quotes, then every sample was assigned to a cluster using k-means. After this, a cluster size register was stored and finally, each cluster was labeled with the observed trend if the probability of being associated to it was higher than 0.6.

6.3.2. Bag of Words

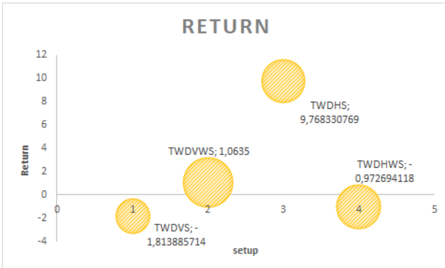
Relying in a representation similar to the one provided by Bookmap, this work aims to automate the trading decision making process supported on the identification of behaviors which provide insights about price movement. If a human trader is able to identify that some changes in the image affect price (through a visual information analysis), then a computer can automatize this task. The Bag of Words technique was selected for this price movements automatic identification process drawn from the premise that text identification and liquidity patterns trend classification share the same principle: identifying certain words it's possible to determine the text's topic with some probability. Similarly, identifying certain image configuration or structure it's possible to figure out the associated trend with some probability.

This simplified representation has been used for a long time in document classification, and more recently, in computer vision [40]. It was selected for this work in order to classify patterns according to a trend. In this case, the document is the order book piece in which the forecasting should be made, and the built patterns are the words.

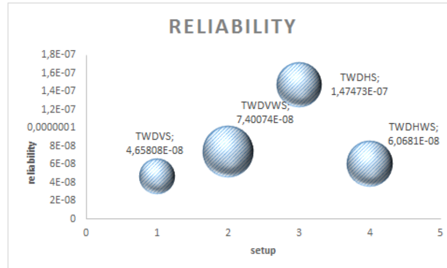
6.3.3. Results

After preparing the data according to the aforementioned configurations (Horizontal Sliding Window, Vertical Sliding Window, etc.), every pattern (or sub image) was assigned to a cluster among 13 clusters obtained for the corresponding configuration using two subsets: in the first one, the whole data set was used; in the second one, a subset of the previous one displaying a clear downwards trend. Extreme values were registered for every configuration and also those patterns indicating a trend (bullish or bearish) with a high probability. With the mentioned patterns dictionary the cumulated return per pattern over the validation set was measured and also the returns mean and variance per configuration. The classifiers precision by configuration was recorded as well.

Figures **6-3** and **6-4** were designed to provide clarity about the obtained results. In these figures, the vertical axis indicates returns' means or reliability means, as appropriate. The horizontal axis presents a set of related arrangements. Lastly, the radius of every circle represents the precision among the classifiers for every arrangement. For further details about the figure's information, please check annex C.



(a) Returns' means over the whole dataset.



(b) Reliabilities means over the whole dataset.

Figure 6-3.: Returns and Reliabilities Means over the whole dataset using vertical and horizontal window arrangement

Table 6-1.: Configurations presenting extreme values using the whole dataset.

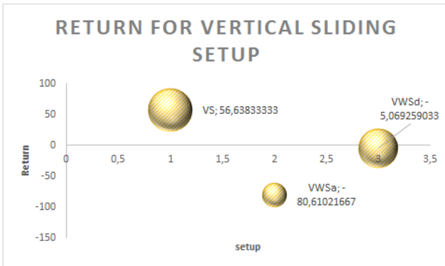
Extreme value	Corresponding arrangement
Highest return	Horizontal Sliding Window(TWDHS)
Lowest return	Vertical Sliding Window(TWDVS)
Highest precision	Vertical Wavelet Sliding Window(TWDVWS)
Lowest precision	Vertical Sliding Window(TWDVS)
Highest Reliability	Horizontal Sliding Window(TWDHS)
Lowest Reliability	Vertical Sliding Window(TWDVS)

In a subsequent task, those classifiers which presented extreme values for the selected metrics were identified in order to determine which configurations consistently provided high returns, high reliability or high precision. Results are available in tables **6-1** and **6-2**.

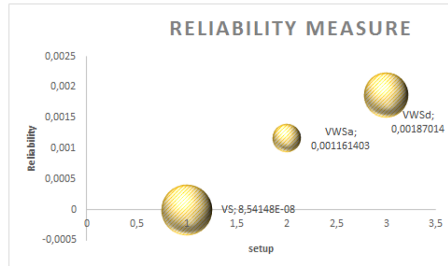
The time horizons with higher mean returns were 16 and 30 minutes. Sliding window arrangements provided higher mean returns than the ones using Adjacent window. In terms of configuration vs. time horizon, the most profitable pairs were: 6 and 16 minutes Vertical Wavelet Sliding Window using differences as well as 16 and 30 minutes using Horizontal Sliding. Examples of the liquidity patterns can be founded in figures **6-5**, **6-6** and **6-7**.

6.3.4. Liquidity Patterns

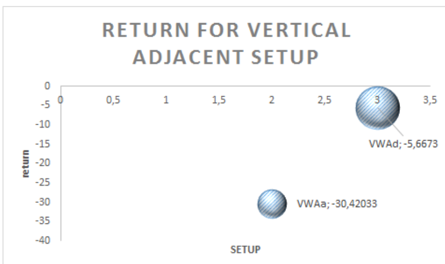
An example of the data in the Colombian Order Book after the pre-processing step is provided in figure **6-8**. A liquidity pattern is a subset or sub image extracted from the Order Book. Figure **6-9** shows examples of frequent patterns associated to a non-negative return from the same cluster.



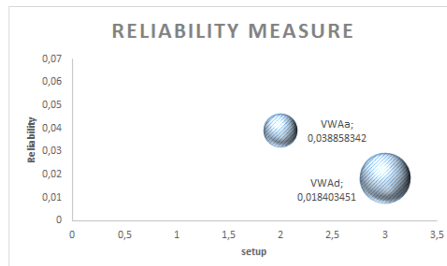
(a) Returns' means over a subset using Vertical Sliding Window arrangement.



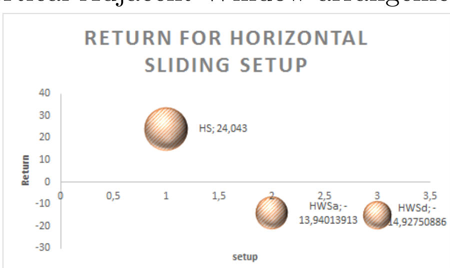
(b) Reliability means over a subset using Vertical Sliding Window arrangement.



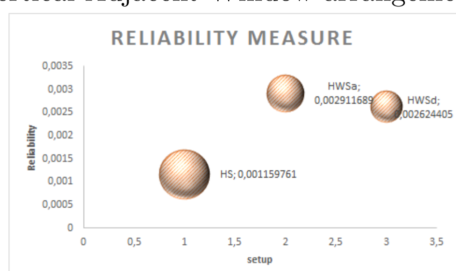
(c) Returns' means over a subset using Vertical Adjacent Window arrangement.



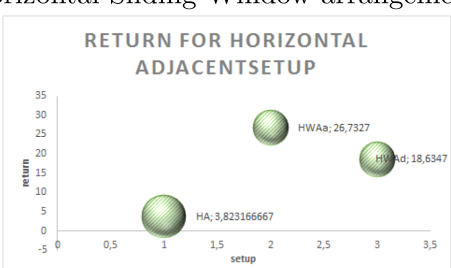
(d) Reliability means over a subset using Vertical Adjacent Window arrangement.



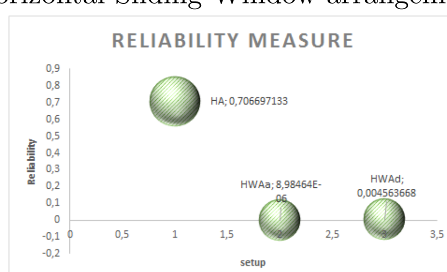
(e) Returns' means over a subset using Horizontal Sliding Window arrangement.



(f) Reliability means over a subset using Horizontal Sliding Window arrangement.



(g) Returns' means over a subset using Horizontal Adjacent Window arrangement.



(h) Reliability means over a subset using Horizontal Adjacent Window arrangement.

Figure 6-4.: Returns and Reliabilities means over the same subset displaying a clear trend using different Window arrangements.

Table 6-2.: Configurations presenting extreme values using the same subset displaying a clear trend.

Extreme value	Corresponding arrangement
Highest return	Vertical Sliding Window(VS)
Lowest return	Vertical Wavelet Sliding Window(VWSa)
Highest precision	Horizontal Adjacent Window(HA)
Lowest precision	Vertical Adjacent Window(VA)
Highest reliability	Horizontal Adjacent Window(HA)
Lowest reliability	Vertical Sliding Window(VS)

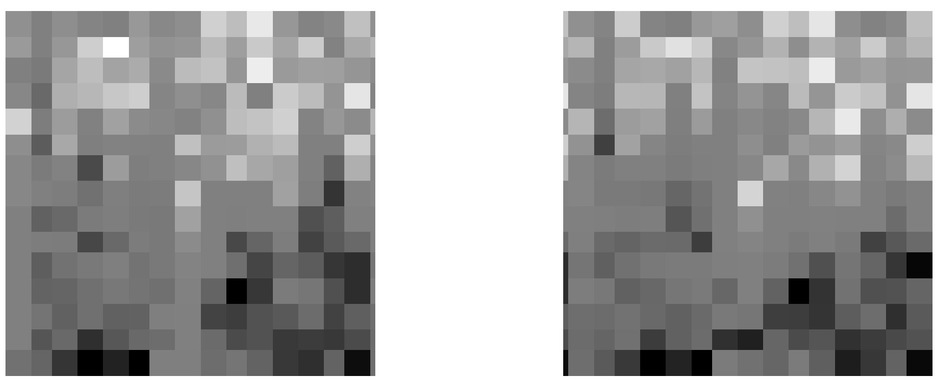


Figure 6-5.: Examples of liquidity patterns exhibiting non-negative returns: cluster 3, horizontal adjacent wavelet arrangement using average coefficients.

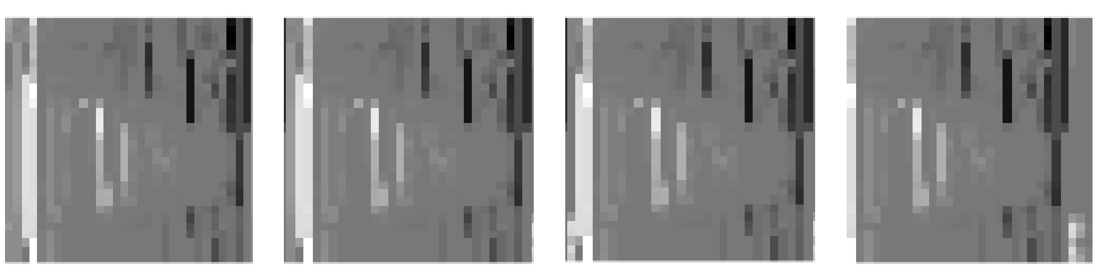


Figure 6-6.: Examples of liquidity patterns exhibiting non-negative returns: cluster 5, horizontal sliding arrangement.

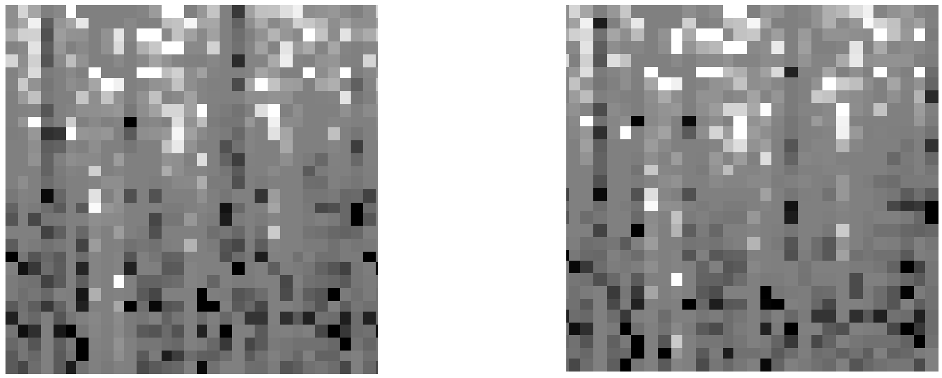


Figure 6-7.: Examples of liquidity patterns exhibiting non-negative returns: cluster 10, horizontal adjacent arrangement.

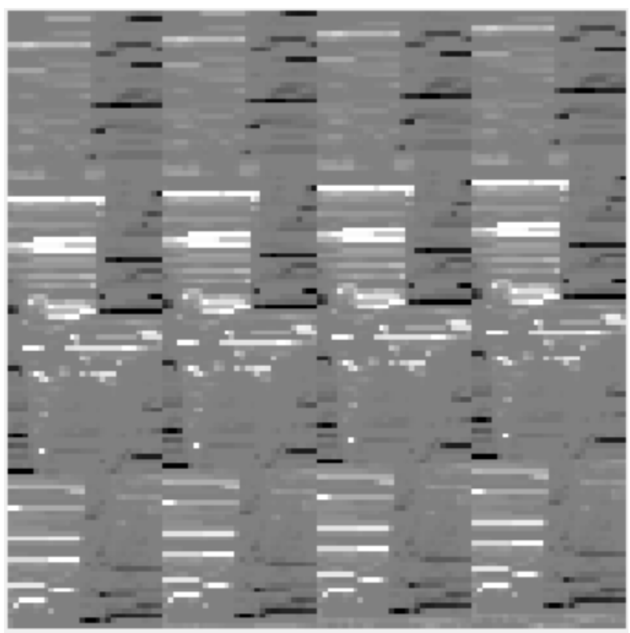


Figure 6-8.: Order Book Visualization example for Vertical Sliding Arrangement.

Figure 6-9.: Non-negative return liqudity patterns example from the same cluster in the Vertical Sliding Arrangement.

7. Conclusions and Future work

7.1. Conclusions

This work introduces a method for generating non-negative returns with a relatively small variance identifying liquidity patterns that are frequent in both, validation and testing sets over six months of real tick data from the Colombian Bulk Currencies Market (USD COP rate) using the Bag of Visual Words technique combined with a Clustering algorithm. The results suggest the presence of seasonal patterns in this market as stated in [7],[13].

This method was tested using five different window arrangements:Vertical arrangement, Horizontal arrangement, Haar Wavelet Transform arrangement, sliding window arrangement and adjacent window arrangement. The highest mean return was provided by the Vertical Sliding Window arrangement, see figure **6-4** (a), and the highest mean reliability was found in the Horizontal Adjacent Window arrangement, see figure **6-4** (h).

An overview of how the research goals were achieved can be found in **??**.

7.2. Discussion

In general, the mean return was higher when training in the region which exhibits a clear global trend. This suggests the presence of seasonal patterns. The classifiers recall also improves in this case. This is, the liquidity patterns founded and labeled in the training set arise in the validation set too. In figure **6-4** the absence of circles indicates that the patterns identified in the training set did not emerge in the corresponding arrangement. According to the results, a higher reliability can be obtained using horizontal configurations, and a Horizontal Adjacent arrangement provides higher returns in a systematic way. However, the arrangement with the highest mean return was Vertical Sliding Window.

As observed in the shape of patterns cluster in Figure **5-3**, the predictive information is gathered excluively around the spread for the granularity level selected in this work. Plausible explanations to this phenomenon are: 1. Orders placed far from the spread contain latent price information not discoverable for this time horizon prediction yet or 2. Indeed,

Objective	**Conclusion**
Providing a survey of the methods published to date for detecting trading strategies using order book information, via a systematic literature review.	A systematic literature review about the Order Book was presented, there is no evidence of previous work made based on the LOB for the Colombian Bulk Currencies Market until the writing date of the second chapter.
Selecting or designing a methodology able to represent in a summarized and efficient way the order book information.	Wavelet Heatmap visualization presents in a summarized and efficient way the order book information.
Establishing a time window in order to preserve relevant order book information.	Both, 1 Minute and 10 Minutes time horizons provide relevant information for trend pridiction in the LOB for the Colombian Bulk Currencies Market.
Selecting or designing a methodology that allows representing properly the Colombian Forex Market Order Book information dynamics.	A methodology which allows representing properly the Colombian Forex Market Order Book information dynamics is presented. The visualization tool presented in this work, can provide the user with a global understanding of a selected time interval in the Colombian Forex Market.
Selecting or designing a trading strategies detection system for the Colombian Forex Market using Order Book information.	Wavelet Heatmap visualization presents in a summarized and efficient way the order book information.Given the seasonality of the found patterns, the presented strategy was reformulated as an adaptive strategy which detects when a pattern is losing predictability in order to start a new training stage for detecting new informative patterns.
Evaluating the performance and feasibility of the proposed system, in supporting the financial decision making process in the Colombian Forex Market.	”The performance of the proposed system in supporting the financial decision making process in the Colombian Forex Market was evaluated.

Table 7-1.: Research goals connected with results.

the only relevant information for prediction is the one around the spread. Future work should explore including this method as part of a multimodal prediction system that operates with alternative information sources such as news, indices, sentiment analysis and fundamental analysis.

Given a high volatility scenario caused by a global order changing event such as war, blackout, pandemic, alien invasion or legislation changes, the following would be expected, given the adaptive nature of the proposed algorithm: Low loss risk. A possible outcome could be, that the dictionary ends empty as the limit for wrong predictions is reached by all patterns stored in it and the algorithm will stop making predictions. On the other hand, it is also possible that the algorithm learns new high volatility patterns which will replace the old ones, according to the new market dynamic.

An additional advantage of the proposed system is the opportunity to adjust the parameters according to a given risk profile or a particular application via tuning the accuracy threshold and the limit for allowed wrong predictions per pattern. This allows using this system as support tool for risk reduction in international business operations or as a prediction support tool for brokers and stockbrokers.

7.3. Future work

The fact that the proposed strategies provide useful results with a relatively small dataset from one single currency, throws as a natural consequence the need of testing them in broader datasets and new assets, even for portfolio selection.

On the other hand, an important feature of the Wavelet based approach is that is highly parallelizable, allowing easy implementation in distributed systems such as GPUs. In order to reduce latency, it would be useful to implement the presented algorithms directly on hardware, for instance in a FPGA.

A. Appendix: Heatmap approach results

Pattern Size 1	1x10
Pattern Size 2	1x20
Pattern Size 3	1x40
Pattern Size 4	2X10
Pattern Size 5	2X20
Pattern Size 6	2X40

Table A-1.: Pattern Sizes

DataSet	1	2	3	4
Pattern Size 1	NaN	NaN	NaN	NaN
Pattern Size 2	0.5352	NaN	0.4747	NaN
Pattern Size 3	0.4615	0.4837	NaN	0.5513
Pattern Size 4	0.4793	0.5087	0.5108	0.5553
Pattern Size 5	0.4950	0.4878	0.4807	0.4670
Pattern Size 6	0.4716	0.4855	0.4843	0.4632

Table A-2.: Experimental Setup 1 (Raw data) 1 minute

DataSet	1	2	3	4
Pattern Size 1	0.4700	0.4703	0.4877	0.4583
Pattern Size 2	0.4700	0.4677	0.4877	0.4570
Pattern Size 3	0.4775	0.4677	0.4876	0.4533
Pattern Size 4	0.4535	0.5081	0.4915	0.4268
Pattern Size 5	0.4649	0.5236	0.4894	0.4504
Pattern Size 6	0.4606	0.5338	0.4691	0.4139

Table A-3.: Experimental Setup 1 (Raw data) 5 minutes

DataSet	1	2	3	4
Pattern Size 1	0.4831	0.5126	0.4969	0.5318
Pattern Size 2	0.4517	0.4906	0.5079	0.4562
Pattern Size 3	0.4583	0.4925	0.5586	0.4586
Pattern Size 4	0.4691	0.5135	0.5325	0.4082
Pattern Size 5	0.4989	0.4640	0.4709	0.3782
Pattern Size 6	0.5807	0.5441	0.4580	0.4296

Table A-4.: Experimental Setup 1 (Raw data) 10 minutes

B. Appendix: Wavelet based approach results

B.1. Accuracy for Wavelet transform approach using only differences, one minute time slot.

DataSet	1	2	3	4
Pattern Size 1	0.4742	0.4683	0.4863	0.4748
Pattern Size 2	0.4880	0.4659	0.4870	0.4653
Pattern Size 3	0.4576	0.4646	0.5113	0.4950
Pattern Size 4	0.4733	0.4666	0.4955	0.4925
Pattern Size 5	0.4758	0.4678	0.4930	0.4417
Pattern Size 6	0.4520	NaN	0.4768	0.4776

Table B-1.: Accuracy for Wavelet transform approach using only differences, one minute time slot, first iteration.

DataSet	1	2	3	4
Pattern Size 1	0.4612	0.4784	0.4958	0.4569
Pattern Size 2	0.4818	0.4821	0.5225	0.5154
Pattern Size 3	0.5758	0.3125	0.5957	0.5632
Pattern Size 4	0.5046	0.4811	0.5102	0.5300
Pattern Size 5	0.9102	0.7470	0.5777	0.5476
Pattern Size 6	NaN	NaN	NaN	NaN

Table B-2.: Accuracy for Wavelet transform approach using only differences, one minute time slot, second iteration.

DataSet	1	2	3	4
Pattern Size 1	0.5126	0.5048	0.5414	0.5584
Pattern Size 2	0.3333	0.8824	NaN	0.5000
Pattern Size 3	NaN	NaN	NaN	NaN
Pattern Size 4	0.4187	0.4026	0.6392	0.4162
Pattern Size 5	NaN	NaN	NaN	0.4487
Pattern Size 6	NaN	NaN	NaN	NaN

Table B-3.: Accuracy for Wavelet transform approach using only differences, one minute time slot, third iteration.

DataSet	1	2	3	4
Pattern Size 1	0.4444	NaN	0	NaN
Pattern Size 2	0.5000	NaN	NaN	NaN
Pattern Size 3	NaN	NaN	NaN	NaN
Pattern Size 4	NaN	NaN	NaN	NaN
Pattern Size 5	NaN	NaN	NaN	NaN
Pattern Size 6	NaN	NaN	NaN	NaN

Table B-4.: Accuracy for Wavelet transform approach using only differences, one minute time slot, fourth iteration.

B.2. Accuracy for Wavelet transform approach using only averages, one minute time slot.

DataSet	1	2	3	4
Pattern Size 1	0.4585	0.4642	0.4841	0.4617
Pattern Size 2	0.4698	0.4635	0.4882	0.4660
Pattern Size 3	0.4721	0.4714	0.4917	0.4947
Pattern Size 4	0.5094	0.5367	0.4691	0.5065
Pattern Size 5	0.4782	0.5009	0.4929	0.4741
Pattern Size 6	0.4993	0.4989	0.4702	0.5234

Table B-5.: Accuracy for Wavelet transform approach using only averages, one minute time slot, first iteration.

DataSet	1	2	3	4
Pattern Size 1	0.4718	0.4818	0.5125	0.4676
Pattern Size 2	0.4708	0.4653	0.5080	0.4614
Pattern Size 3	0.5013	0.4906	0.6017	0.4767
Pattern Size 4	0.5325	0.4981	0.4772	0.5398
Pattern Size 5	0.4878	0.5333	0.3618	0.5114
Pattern Size 6	NaN	NaN	NaN	0.5027

Table B-6.: Accuracy for Wavelet transform approach using only averages, one minute time slot, second iteration.

DataSet	1	2	3	4
Pattern Size 1	0.4652	0.4930	0.4865	0.4693
Pattern Size 2	0.5227	NaN	0.4573	0.6120
Pattern Size 3	NaN	NaN	NaN	0.6667
Pattern Size 4	0.4300	0.4126	0.4973	0.4003
Pattern Size 5	NaN	NaN	NaN	0.2629
Pattern Size 6	NaN	NaN	NaN	NaN

Table B-7.: Accuracy for Wavelet transform approach using only averages, one minute time slot, third iteration.

DataSet	1	2	3	4
Pattern Size 1	0.4722	NaN	NaN	0.5198
Pattern Size 2	NaN	NaN	NaN	0.4745
Pattern Size 3	NaN	NaN	NaN	NaN
Pattern Size 4	NaN	NaN	NaN	0.3892
Pattern Size 5	NaN	NaN	NaN	NaN
Pattern Size 6	NaN	NaN	NaN	NaN

Table B-8.: Accuracy for Wavelet transform approach using only averages, one minute time slot, fourth iteration.

B.3. Accuracy for Wavelet transform approach using only differences, ten minutes time slot.

DataSet	1	2	3	4
Pattern Size 1	0.4914	0.5220	0.4969	0.4775
Pattern Size 2	0.4603	0.4884	0.5362	0.4231
Pattern Size 3	NaN	0.7500	NaN	NaN
Pattern Size 4	0.4628	0.5430	0.4604	0.3223
Pattern Size 5	NaN	NaN	NaN	0.2164
Pattern Size 6	NaN	NaN	NaN	NaN

Table B-9.: Accuracy for Wavelet transform approach using only differences, ten minutes time slot, first iteration.

Table B-10.: Add caption

DataSet	1	2	3	4
Pattern Size 1	0.5429	0.4407	0.6027	0.4531
Pattern Size 2	NaN	NaN	NaN	0.6000
Pattern Size 3	NaN	NaN	NaN	NaN
Pattern Size 4	NaN	NaN	NaN	0.2269
Pattern Size 5	NaN	NaN	NaN	NaN
Pattern Size 6	NaN	NaN	NaN	NaN

Table B-11.: Accuracy for Wavelet transform approach using only differences, ten minutes time slot, second iteration.

B.4. Accuracy for Wavelet transform approach using only averages, ten minutes time slot.

DataSet	1	2	3	4
Pattern Size 1	0.4839	0.5026	0.5440	0.4588
Pattern Size 2	0.4917	0.4890	0.5412	0.4475
Pattern Size 3	0.4048	0.4091	0.4600	0.4706
Pattern Size 4	0.4709	0.5590	0.5067	0.3640
Pattern Size 5	0.5859	0.6090	0.7220	0.4704
Pattern Size 6	NaN	0.8333	NaN	0.3571

Table B-12.: Accuracy for Wavelet transform approach using only averages, ten minutes time slot, first iteration.

DataSet	1	2	3	4
Pattern Size 1	0.5000	0.5238	0.5155	0.4143
Pattern Size 2	0.5102	0.3929	0.4872	0.3636
Pattern Size 3	NaN	NaN	NaN	NaN
Pattern Size 4	0.4464	0.5383	0.5556	0.3384
Pattern Size 5	0.5097	0.2636	0.6164	0.3048
Pattern Size 6	NaN	NaN	NaN	NaN

Table B-13.: Accuracy for Wavelet transform approach using only averages, ten minutes time slot, second iteration.

C. Appendix: Clusters and Bag of Words results

C.1. Return measure

Table C-1.: Vertical Sliding Configurations.

	AvgVar	ret/var	AvgRet	useful/notUseful
VS	11707567,05	4,83775E-06	56,63833333	0,545454545
VWSa	860,0277083	-0,09372979	-80,61021667	0,166666667
VWSd	533,7192641	-0,009497988	-5,069259033	0,428571429

Table C-2.: Vertical Adjacent Configurations.

	AvgVar	ret/var	AvgRet	useful/notUseful
VA	17,3969	-0,868054079	-15,10145	0
VWAa	24,73450	-1,229874602	-30,42033	0,333333333
VWAd	53,3376	-0,10625237	-5,6673	0,75

Table C-3.: Horizontal Sliding Configurations.

	AvgVar	ret/var	AvgRet	useful/notUseful
HS	861,2468	0,027916504	24,043	0,64
HWSa	342,4432826	-0,040707877	-13,94013913	0,347826087
HWSd	380,0386887	-0,039278919	-14,92750886	0,263157895

C.2. Reliability measure

Table C-4.: Horizontal Adjacent Configurations.

	AvgVar	ret/var	AvgRet	useful/notUseful
HA	0,415033333	9,211709903	3,823166667	1
HWAa	111.300,0843	0,000240186	26,7327	0,666666667
HWAd	218,1220	0,085432312	18,6347	0,666666667

Table C-5.: Vertical Sliding Configurations.

	Reliability	useful/notUseful
VS	8,54148E-08	0,545454545
VWSa	0,001161403	0,166666667
VWSd	0,00187014	0,428571429

Table C-6.: Vertical Adjacent Configurations.

	Reliability	uAeful/notUAeful
VA	0,054356984	0
VWAa	0,038858342	0,333333333
VWAd	0,018403451	0,75

Table C-7.: Horizontal Sliding Configurations.

	Reliability	useful/notUseful
HS	0,001159761	0,64
HWSa	0,002911689	0,347826087
HWSd	0,002624405	0,263157895

Table C-8.: Horizontal Adjacent Configurations.

	Reliability	uAeful/notUAeful
HA	0,706697133	1
HWAa	8,98464E-06	0,666666667
HWAd	0,004563668	0,666666667

D. Appendix: Glossary Basic Notions and Definitions

D.1. Basic Financial Concepts

This section was created in order to provide a guide for those readers which are not familiar with the notions related to financial markets or with concepts belonging to linear algebra and statistical learning.

D.1.1. Forex Markets

«*The foreign exchange market is a global decentralized market for the trading of currencies. In terms of volume of trading, it is by far the largest market in the world.*»[51]

D.1.2. Order book

Cont et al. [11] model the order book as a grid of price ticks, where:

Ask

The ask price is defined as [11]:

$$p_A(t) = inf\,\{p = 1, ..., n, X_p > 0\} \wedge (n + 1). \tag{D-1}$$

Bid

And the bid price is defined as [11]:

$$p_B(t) = sup\,\{p = 1, ..., n, X_p < 0\} \quad \vee \quad 0. \tag{D-2}$$

Depth

Distance from the best price. For the bid side the volume at a certain distance i is given by [11]:

$$Q_i^B(t) = \begin{cases} X_{P_A(t)-i}(t) & 0 < i < P_A(t) \\ 0 & P_A(t) \leq i < n \end{cases} \quad \text{(D-3)}$$

For the ask side:

$$Q_i^A(t) = \begin{cases} X_{P_B(t)+i}(t) & 0 < i < n - P_B(t) \\ 0 & n - P_B(t) \leq i < n \end{cases} \quad \text{(D-4)}$$

D.1.3. Spread

Is the difference between the best price in the ask side and the best price in the bid side [11]:

$$p_s(t) \equiv p_A(t) - p_B(t) \quad \text{(D-5)}$$

D.1.4. Trader

Is the agent which produce movements in the order book.

D.2. Scientific Visualization

Scientific Visualization is the mapping of scientific data and information to imagery to gain understanding or insight.[1]

In 1995 Moorhead and Zhu [43] encouraged scientific community to cooperate in the scope of the scientific visualization due to the massive quantity of information. They describe the visualization process of geometric objects in 3D space and point out the fact that sometimes data are directly mapped into images, skipping the geometric description. In order to represent the data, attention is drawn to the following image components: shape, color and opacity.

This thesis attemps to extract knowledge from scientific visualization according to the definition of Chen et al. [8], provided in table **D-1**.

D.3. Heatmap

In 2014 Todd et al. [57] depict an order book heat map to facilitate the analysis of three dimensional data. They suggest the use of heat maps in order to auditor long periods, for

[1]Moorhead, R.J.; Zhifan Zhu, "Signal processing aspects of scientific visualization,"Signal Processing Magazine, IEEE , vol.12, no.5, pp.20,41, Sep 1995. DOI: 10.1109/79.410438

Table D-1.: Todd et al. [57] definitions of data, information and knowledge in computational space.

Category	Definition
Data	Computerized representations of models and attributes of real or simulated entities.
Information	Data that represents the results of a computational process, such as statistical analysis, for assigning meanings to the data, or the transcripts of some meanings assigned by human beings.
Knowledge	Data that represents the results of a computer-simulated cognitive process, such as perception, learning, association, and reasoning, or the transcripts of some knowledge acquired by human beings.

instance, years. They use color to map buy and sell orders' depth as shown in **D-1**. They suggest using order book for market surveillance.

The underlying hypothesis for using this technique for the visualization of the order book is that the order placement produces an effect in the price behavior and therefore it helps in order to identify price trends.

The dataset is explored through several time window sizes in order to find a set of frequent blocks or a set of frequent groups of adjacent blocks strongly associated with a price trend.

The number of times that a pattern in the validation set matches the price trend assigned during the training will be this visualization evaluation's method.

D.4. Bag of Words

This work is based under the assumption that frequent Price-time-volume structures within the dataset are informative. Linares, Gonzalez et Hernandez [58] indentified individual basic shapes on time series in order to build active trading strategies based on forecasting. In that vein, is reasonable to count every appearance of each pattern and store the number of times that the pattern is associated with a specific trend (in this case bullish or bearish), in order to calculate the probability of the pattern of being related with a trend. This process allows labeling frequent patterns in bullish or bearish patterns with the purpose of building a classifier.

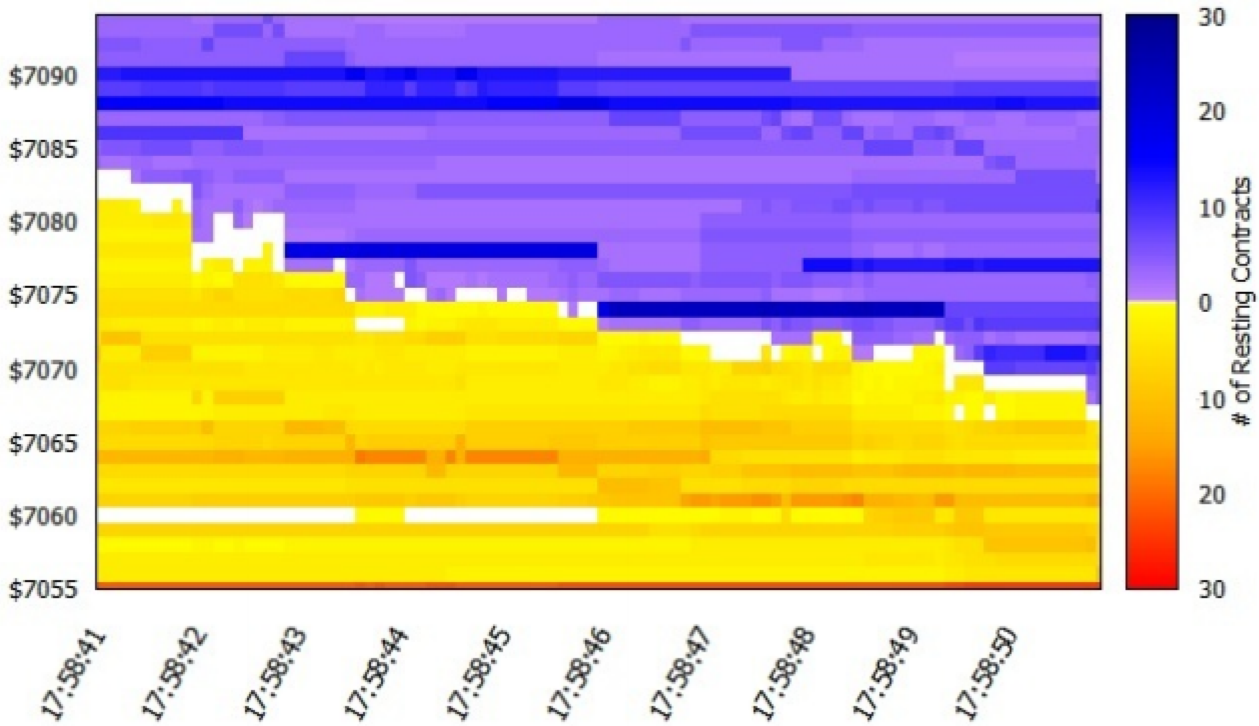


Figure D-1.: Example of Heat Map of Order Book Depth, Todd et al. [57])

The first reference to the Bag of Words method appears in [24] when Harris states that «*it is possible to define a linguistic structure solely in terms of the "distributions" (= patterns of co-occurrences) of its elements. There is no parallel meaning-structure which can aid in describing formal structure. Meaning is partly a function of distribution.*». Later, in 2003, Sivic et al. [54] present an analogy between text and image retrieval for video retrieval where the construction of the visual vocabulary is made quantizing descriptors in clusters (using k-means) extracted from a fragment of the film. Lopez-Monroy, Gomez et al [40] show the general process for generating a bag of visual words from a set of images.

D.5. Wavelts

The idea behind the use of this tool for visualizing the order book is that the reduction of redundant information will make easier the patterns visual detection task.

The dataset is explored through several time window sizes in order to find a set of frequent blocks or a set of frequent groups of adjacent blocks strongly associated with a price trend, as before.

The number of times that a pattern in the validation set matches the price trend assigned during the training will be this visualization evaluation method too.

From this point this section will be an outline from the first chapter of the book «A Wavelet Tour of Signal Processing: The Sparse Way» by Stephane Mallat [41][2]. Meyer and Mallat provide a systematic theory through the elaboration of multiresolution signal aproximations.

D.5.1. Wavelet bases

Wavelet bases (WB) reveal the signal regularity through the amplitude of coefficientes, and their structure leads to a fast computational algorithm. A WB defines a sparse representation of piecewise regular signals, which may include transients and singularities. In images, large wavelet coefficients are located in the neighborhood of edges and irregular textures.

D.5.2. Quick glossary

- Haar Wavelet: piecewise constant function.

[2]Mallat Stephane.A Wavelet Tour of Signal Processing: The Sparse Way. Elsevier. Third Edition. 2009. (pp. 3-16.)

- Orthonormality: two vectors in an inner product space are orthonormal if they are orthogonal and unit vectors. For Euclidean spaces two vectors are orthogonal if and only if their dot product is zero. Is an extension of the concept of perpendicularity amongst vectors. $* <, >$ represents the inner product.
- Strömberg Wavelet: A piecewise linear function ψ that also generates an orthonormal basis and gives better approximations of smooth functions.
- Mayer Wavelet: A family of orthonormal wavelet bases with infinitely continously differentiable functions.

D.5.3. Filter Bank

A filter bank is an array of band-pass filters that separates the input signal into multiple components, each one carrying a single frequency sub-band of the original signal.

Conjugate mirror filters

There is an equivalence between continuous time wavelet theory and discrete filter banks. A new interface between digital signal processing and harmonic analysis.

Usefulness of mixing continuous infinite analysis with discrete finite analysis

Mixing continuous infinite analysis with discrete finite analysis is useful because the asymptotic results provided for the first one are precise enough to understand the behavior of discrete algorithms but not sufficient for elaborating discrete signal-processing algorithms. The restriction of the constructions to finite discrete signals adds complexity because of the border problems, but with the understanding of the properties of the bases, this issue can be addressed.

D.5.4. Wavelets for images

Wavelet Orthogonal Bases (WOB) of images can be constructed from WOBs of one-dimension signals. An algorithm for calculating fastly wavelet coefficients is provided in chapter 7 of [41]. «Like in one dimension, a wavelet coefficient has a small amplitude if the function which defines it is regular over the support of the mother wavelets. It has large amplitude near sharp transitions such as edges.»(k is the direction and 2^j is the scale and both correspond to a subimage).

Approximation and processing in bases

«Sparse representations that reduce the number of parameters can be obtained by thresholding coefficients in an appropiate orthogonal basis.»

D.5.5. Sampling

There are two kinds of approximation errors in sampling: Linear approximation error and non linear approximation error. An aproximation by thresholding is made by selecting the best vectors in the orthogonal basis of the whole analog signal space. This approximation is not linear. This is important due to the increase of the approximation resolution where the signal is irregular. Approximation support provides geometric information on the orthogonal projection, relative the to dictionary, that is a wavelet basis in the given example, so the error is smaller.

Sparsity with Regularity

When the image is not that sharp, the non linear wavelet approximation produces small errors. As an adaptation to this issue, more representations with curvelets and bandlets can be used.

D.5.6. Compression

When coding a sparse representation via transform codes, the coefficients are approximated by quantized values. Mallat states that «Compression is a sparse approximation problem.». Higher coefficients are associated with geometric properties such as edges.

D.5.7. Denoising

Donoho and Johnstone [15] state that «Simple thresholding in sparse representations can yield nearly optimal nonlinear (noise) estimator». Bayes risk «is the expected risk calculated with respect to the prior probabilbity distribution pi of the random signal model F».

Wald used deterministic models, where signals are elements of a set, without specifying their probability distribution in this set.

Thresholding Estimators

Donoho and Johnstone proved that in an orthonormal basis, a simple thesholding of noisy coefficients provides a sparse support of the orthogonal projection from the noisy data.«Minimax risk is the lower bound computed over all operators D.»

D.5.8. Time Frequency Dictionaries

Gabor [21] «proposed decomposing signals over dictionaries of elementary waveforms which he called time frequency atoms that have a minimal spread in a time-frequency plane.» He also states that «The key issue is to understand how to construct dictionaries with time-frequency atoms adapted to signal properties.»

D.5.9. Heisenberg Uncertainty

The uncertainty principle theorem proves that this rectangle has a minimum surface that limits the joint time-frequency resolution: $\sigma_(t,\gamma)$,$\sigma_(\omega,\gamma) >= 1/2$. See **D-2**

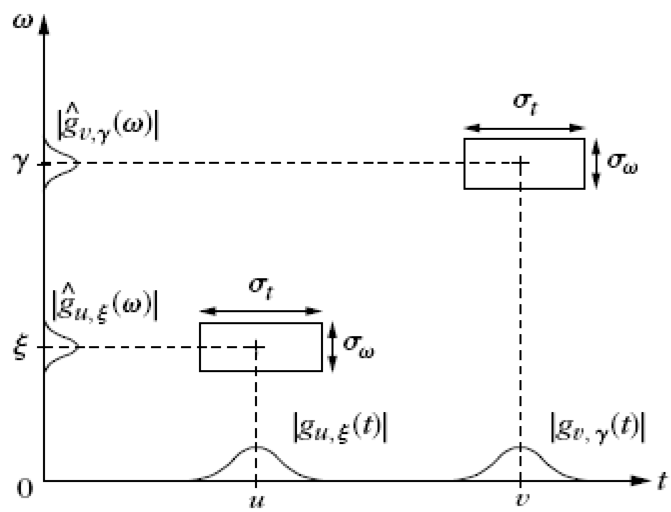


Figure D-2.: Time-frequency boxes representing the energy spread of two windowed Fourier atoms [41].

D.5.10. Wavelets for images

Mallat [41] defines wavelets for images as:

$$f(x) = f(x_1, x_2) : \left\{ \psi^k_{j,n}(x) = \frac{1}{2^j}\psi^k\left(\frac{x - 2^j n}{2^j}\right)\right\}_C \tag{D-6}$$

Where:

$$C = j \in Z, n \in Z^2, 1 \leq k \leq 3 \tag{D-7}$$

$$\psi(t) = \begin{cases} 1 & if \quad 0 \leq t < \frac{1}{2}, \\ -1 & if \quad \frac{1}{2} \leq t < 1, \\ 0 & otherwise \end{cases} \tag{D-8}$$

$\psi(x)$ denotes the mother wavelet, 2^j corresponds to the scale, $2^j n$ the translation, k to the direction, $x1$ and $x2$ are the row and the column in the matrix. Here ends the outline from Mallat's book

.

E. Systematic Literature Review / State of the art

This research studied the problem of order book information extraction and its application in the Colombian Forex Market. The main objective of this work is to determine if the information provided by the Colombian Forex Market Order Book is enough in order to recognize propitious trading scenarios.

Two main components have been investigated in the proposed strategy: a proper visualization and the frequent patterns exploration. A proper visualization facilitates the task of user's data understanding, according to the information needs. The frequent patterns exploration deals with the processing of a huge amount of available data in a reasonable time lapse. These components will be tested with real tick data from the Colombian Forex Market of the year 2012.

E.1. Previous work

This section presents a review of recent work done about the information content, dynamics, order placement and strategies in the order book during the last decade, approximately. The purpose is to provide to a non expert reader with a general understanding of the subject. This is achieved by explaining the order book operation, what is relevant in the order book information, how it operates, how is modeled and represented, and presenting some methods of using the order book to support investment strategies. Finally, an analysis and a discussion about future trends and possible enhancement methods is introduced.

The recent progress of the order book exploration in several markets worldwide and in simulated markets is presented in this chapter. It was made a classification of the literature in six categories: information content, dynamics, order placement, representation and modeling, trading strategies, and consequences and forecasting. This classification introduced a general view that serves as a basis for trend analysis.

Goods exchange has been one of the activities developed by humankind, necessary for civilizations to thrive. Nowadays, this activity has evolved towards automatic boosting profit methodologies in sophisticated markets, such as Forex market.

In order to perform a profitable investment, two questions should be solved: where to invest and when to trade. In this paper, the order book, a tool that provides information which supports the decision making process necessary to answer the second question, is explained.

The order book was employed under the assumption that there is market information which allows the discovery of price behavior pattern association with a future market trend, that would provide some probably optimal trading points in time which would support the traders' decision making task.

The main aim of this work is to provide elements to understand the informational potential of the order book by means of a systematic literature review, analyzing trends, defining study categories and determining the study state of the markets surveyed.

The rest of the chapter is organized as follows: subsection 2 presents a brief review of the main concepts and definitions necessary to understand the order book dynamics and, connects High Frequency Trading with Forex markets and order book analysis; subsection 3 introduces a discussion about the strategies based on the order book analysis; lastly, subsection 4 presents the conclusions and some suggestions for improving trading strategies.

E.1.1. Order book

With the purpose of registering the prices at which traders would buy or sell a financial instrument and at which volume, the order book was created. This tool allows accessing information in order to characterize the asset behavior, and based on this model, being able to generate trading strategies that would increase the investor wealth.

Order book information content

For the sake of understanding the order book information content impact, the way in which the order book is presented has been studied by [65] in order to determine whether it influences the order placement strategies. Yu shows that the limit order book information does have a different impact in the order submission in the bid side than in the ask side, using a generalization of a linear regression model which assumes a discrete dependent variable and a probit model.

Another study concerning the order book content information analysis [33], proposes a strategy formed on a combination of dynamic focus and naive price adjustment (NPA). The results presented by Jiaqi et al. outperform the use of NPA exclusively, in a higher level if the asset presents more liquidity. Order book dynamic volume adjustment points are shown to be useful for controlling an adverse price selection and information covering. It is showed that price and volume information in the book are complementary and essential to select informational features.

Limit order book were increased from three to five, the top price levels displayed in Chinese stock market. Li et al. [39], study the effects in price discovery. The results show that the there exists significant difference between the pre and post transparency rise. The new quotes added have little information content, but is useful for traders to improve the price discovery process.

Fletcher et al. [18] use SVM's in junction with multiple kernel learning (MLK) for analyzing LOB information. MLK was used to mitigate the error over wrong kernel selection with mixed results. Temporal window lengths were defined after experimentation. Price movement of the EUR/USD currency was tested under this framework but profitability was not achieved. MKL and simple kernel methods did not show the expected performance difference but MKL was able to identify the most informative feature subsets. Future work can be interesting using reduced feature spaces.

Order book dynamics

In the study of the order book dynamics, two aspects are considered: its depth and to which extent is informative. The helpfulness of the whole content of the book has been studied in [48], distinguishing between two kind of data: those belonging to the best quotes, and further. The conclusion was that the whole book helps to determine if the trader provides or consumes liquidity and that more patient traders use deeper order book information. The study presented by [27] models the order book high frequency dynamics in the London Stock Exchange using second level data (best quotes and volume at different prices), this model allows to capture the arrival of orders of different sizes.

Changes in order book have been studied in [47], where forecasting price change and the direction of such change are problems addressed separately. However, this approximation did not produce statistically significant results. Ranaldo [50] also studies changes in order book shape, analyzing transaction aggressiveness and the order book flow using Swiss Stock Exchange data. This study shows how traders using market orders and those using limit orders react in opposite ways to market changes. Also, market equilibrium is associated with weak aggressiveness in trading and imbalance between ask and bid is associated with higher aggressiveness, this provides information about how order book shape changes before any submission operation, forming up a thinner shape in that side.

Kercheval et al. [34] defined a set of classes (dynamic metrics) and feature spaces (raw data from books, «economical set» or refined raw data after entropy reduction). Support vector machines were used in order to predict the market based on his strength of optimal classification over linear separable samples. Those metrics were created focused on the profitability and the search of a possible competitive advantage (for example, a future bid versus actual ask). Multi-class SVM and kernel transformations were taken into account to improve reliability and over-adjustment. Results from both feature spaces did not show significant differences. On the other hand, SVM's as classification engine proved to be reliable (over 98.5 % on precision and recall) for the less risky scenarios (profit sense).

Orders placement

Finding an optimal point to buy or sell is another widely discussed problem in finance. Using the order book, in [36], price change point quickest detection is pursued employing a social learning model and a protocol for quickest detection It is evidenced that an optimal decision policy has several thresholds.

The order book time-varying composition expressed by each time stamp, is studied in [32]. Jiang et al. aim to model the limit price distribution in the ask and the bid side at each instant, using a gamma distribution to analyze its impact in volume and the existence of seasonal patterns. The results present strong parameters distribution seasonality, providing a model of how markets evolve in time.

Malik et al. [42], selected LOB data over bid-ask spread seeking a much deeper connexion between the data. Liquidity is defined as trading opportunity over large volumes and curve fitting inside a large time window is the most radical approach. Besides no profitability is found in the long run, the market has shown nonlinear, time-varying characteristics. Trade scheduling algorithms can be improved with this framework.

Order book representations and modeling

With the aim of harnessing order book information content, finding a representation which allows handling its volume, and a model that reflects its changes is required. For markets operated in discrete time periods, in [59], a Markov process is modeled providing a minimum number of constant parameters, considering a one-product-market where the prices have a finite and small number of possibilities. In order to provide a shorter order book representation, in [30], Jiang et al. reduce the representation to four parameters by snapshot. A Kalman filter is used to estimate a linear dynamic system state and provide a liability measure, being used for prediction and filtering. The results, obtained from the London Stock Exchange, show that jumps in the estimated parameters are always detected.

Another factor used to represent the order book information content is its slope [5], Cheng et al. study how order placement contributes to the price formation process, measuring how the quantity supplied changes as price does (elasticity). It is concluded that order book limits contain the current supply information and its relationship with the demand, allowing the creation of more precise investment strategies.

Ahn et al. [2], performs a quotes' price clustering over a maximum of five quotes to limit orders. Deeper quotes presented higher clustering than the best ones, pointing that the further from the best a quote is, the less information provides. This study uses data from the Hong Kong Stock Exchange.

Consequences and forecasting

The Triennial Central Bank Survey[53] presents a foreign exchange turnover report evidencing the growth in forex exchange trading, with a turnover of USD 5.3 trillion per day in April 2013 and a growth rate of 32.5 % compared with April 2010. This survey shows that the dollar is stil the dominant currency in FX deals with a 87 % of presence in April 2013. This information provides an insight on the importance of studying currencies behavior, specially the United States Dollar.

High frequency trading (HFT) is defined by the International Organization of Securities Commissions (IOSCO) as «a very quantitative trading form» with the following common features[1]:

- It involves the use of sophisticated technological tools for pursuing a number of different strategies, ranging from market making to arbitrage;
- It is a highly quantitative tool that employs algorithms along the whole investment chain: analysis of market data, deployment of appropriate trading strategies, minimisation of trading costs and execution of trades;
- It is characterized by a high daily portfolio turnover and order to trade ratio (i.e. a large number of orders are cancelled in comparison to trades executed);
- It usually involves flat or near flat positions at the end of the trading day, meaning that little or no risk is carried overnight, with obvious savings on the cost of capital

[1] Regulatory Issues Raised by the Impact of Technological Changes on Market Integrity and Efficiency Consultation Report. July 2010. Retrived on January 2014 from `http://www.iosco.org/library/pubdocs/pdf/IOSCOPD354.pdf`

associated with margined positions. Positions are often held for as little as seconds or even fractions of a second;

- It is mostly employed by proprietary trading firms or desks; and
- It is latency sensitive. The implementation and execution of successful HFT strategies depend crucially on the ability to be faster than competitors and to take advantage of services such as direct electronic access and co-location.

LOB analysis with reasonable execution time requires the use of HFT techniques. Consequences of the adoption of High Frequency Trading on an Electronic Market are discussed in [35], where the Flash Crash, «a brief period of extreme market volatility on May 6, 2010» is analyzed. Some of the highlights in [35] are:

- ”High Frequency Traders aggressively trade in the direction of price changes. This activity comprises a large percentage of total trading volume, but does not result in a significant accumulation of inventory.”
- ”High Frequency Traders are not willing to accumulate large positions or absorb large losses.”
- ”When rebalancing their positions, High Frequency Traders may compete for liquidity and amplify price volatility.”
- ”HFTs did not trigger the Flash Crash, but their responses to the unusually large selling pressure on that day exacerbated market volatility”

E.1.2. Strategies based on the order book analysis

In an attempt to generate trading strategies with order book information, Bates et al. [4], combine the use of technical indicators with order book information, using a reinforcement learning algorithm. The results show that this combination outperforms the use of technical signals solely. Two types of indicators are employed: stop loss orders and get profit orders.

Farmer et al. [17], address the problem of constrains over intelligent agents by market institutions. It was empirically proved that even the most simple models develop a similar behavior. This work has shown circumstances where the most intelligent strategies cannot be successful.

With the objective of characterizing the behavior of traders, independently of the circumstances, Yang et al. [64], used inverse reinforcement learning to infer a solution model to the choices made by the trader, using a reward function learnt from the previous observations, producing an autonomous process. This study presents a simulation model based on agents for the futures market. It provides evidence that is possible to identify reliably other algorithmic trading high frequency strategies, using inverse reinforcement learning. It is also shown that is possible to identify precisely a manipulative high frequency strategy among other high frequency strategies.

Informed traders adjust order submissions to the level of risk perceived, and remove mispriced limit orders in the book. Trading strategies of the informed and the liquidity traders diverge in time. Informed traders provide liquidity through limit order submission due the dynamic order behavior[5]. Cheng et al. [9] Using simulations support the evidence that traders with analysis of LOB information produce more accurate expectation of future asset price. A market lead by LOB analysis as a strategy would show high liquidity and low volatility.

Evolving trading strategies [63] presents two main behaviors which provides a reasonable profit: buy early and hold the stock, selling solely if the price decreases and, buy if the price is lower than the average price over a certain amount of time steps.

Pascual et al.[48], provides a model for aggressiveness which outperforms the ordered probit model for forecasting. This model is a two-stage sequential ordered probit (SOP) model which separates the decision of withdraw, provide or consume liquidity, from the decision of choosing a particular type of order, and allows separating patient traders' order choices from impatient traders' order choices, using information of the whole LOB. This study also provides a confirmation of other studies:

- Asymmetry increases patient buyers operations and impatient sellers operations.
- The higher the bid-ask spread, the higher the frequency of inside-the-quotes limit orders, and increases the frequency of the most aggressive market orders.
- Patient traders increase aggressiveness if the depth at the best quote increases, impatient traders decrease aggressiveness as the thickness of the opposite quote increases.
- Patient and impatient buyers (sellers) increase aggressiveness if the book above (below) the best ask (bid) gets deeper.
- The aggressiveness of incoming impatient (patient) traders is converse to the length of the opposite (same) side of the market.

E.1.3. Discussion

Figure **E-1** shows that the topic that has been more widely studied from the proposed categories, is the representation and modeling of the order book, followed by trading strategies

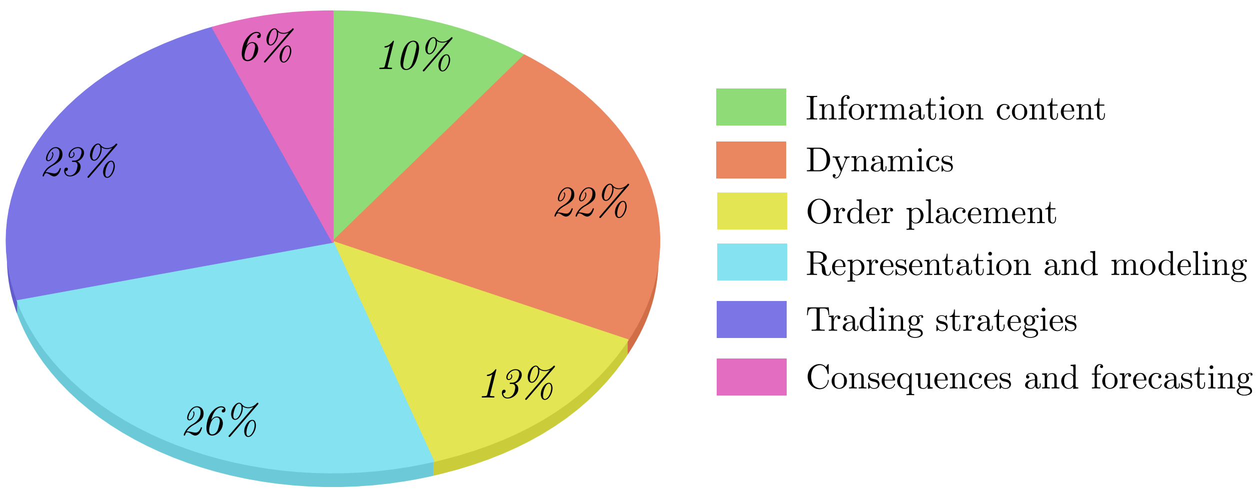


Figure E-1.: References distribution by category.

based on it. This suggests that guiding future research towards this direction could be rewarding because is a very active branch. Table 1 provides information about the references linked to the proposed categories.

Table E-1.: References classification by subject.

Order book classification	References related
Information content	[65], [33] and [39]
Dynamics	[27], [46],[47],[48], [50], [59] and [60]
Order placement	[26], [32], [36] and [55]
Representation and modeling	[2], [9],[20],[23], [27],[30],[31], [37] and [59]
Trading strategies	[1], [4], [5], [38] and [61],[63],[64]
Consequences and forecasting	[35] and [53]

In figure 2, a graphical summary of the main techniques used in each category is provided. Techniques used for order book representation and modeling are diverse, harnessing control theory, physics, econometrics, machine learning and statistical techniques. In future research, the use of more sophisticated probabilistic graphical models (PGM) is encouraged by the author of this paper, due to the successful application of such techniques in problems which involve handling large amount of data. PGM could improve trading strategies also, given their ability to characterize hidden behavior in data.

Figure **E-2** provides a publications' timeline for each category. From the selected literature, based on the variance in the number of publications by category, it can be observed that:

- The categories that presented growing in time were LOB's Representation and modeling, LOB's Order placement, and LOB's Trading strategies.

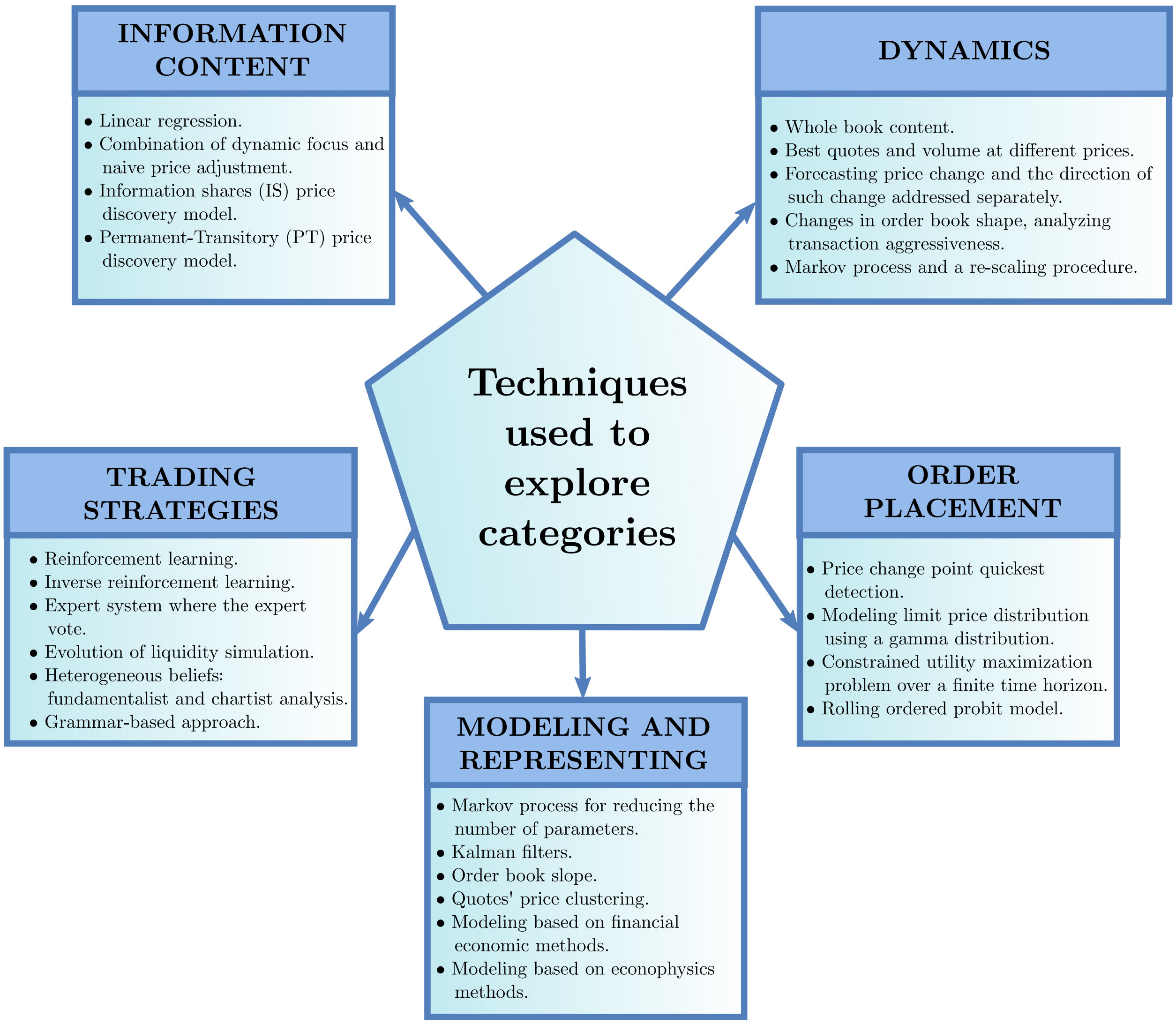


Figure E-2.: Main techniques used for exploring each category.

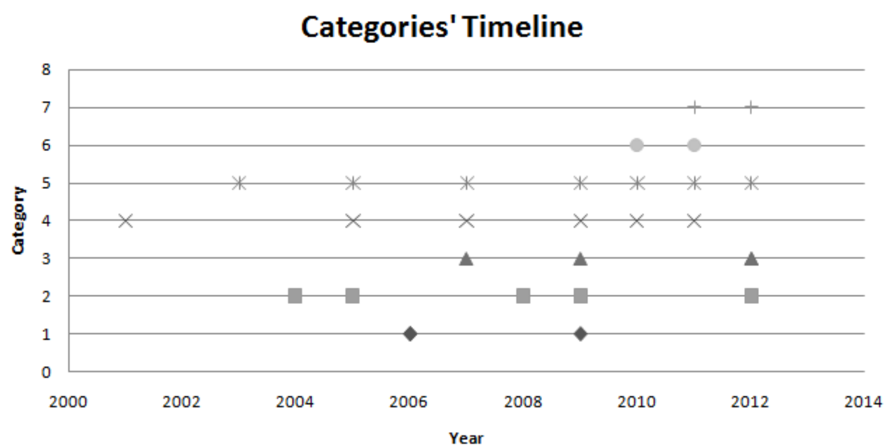


Figure E-3.: Publications' timeline for each category: line 1 presents information about LOB's Information content, line 2 presents information about LOB's Dynamics, line 3 presents information about LOB's Order placement, line 4 presents information about LOB's Representation and modeling, line 5 presents information about LOB's Trading strategies, line 6 presents information about LOB's Consequences and forecasting and line 7 presents information about LOB's Dynamics and Representation and modeling simultaneously.

- Publications about LOB's Information content and LOB's Dynamics, showed stagnation in time.

- Articles on LOB's Consequences and forecasting, and about LOB's Dynamics and Representation and modeling simultaneously have recently emerged.

Table 2 presents a market classification by geographic region and association with the order book categories explored on them. The literature related with the Asian and the European markets is ample, but no literature describing African or south American markets was found. Studies from European markets are focused in the understanding of the order book dynamics, while studies from Asian markets are producing results in all categories.

A promising field related to trading strategies is strategy detection. Yang et al.[64] provide a methodology (inverse reinforcement learning) successfully tested which allows specific trading strategies detection. This would provide tools to design specialized strategies for getting profit from an identified strategy.

Table E-2.: Markets classificated by geographic region and associated with the order book subjects explored on them (IC, D, OP, RM, TS, CF stands for information content, dynamics, order placement, representation and modeling, trading strategies, and consequences and forecasting, respectively).

Market	Subjects	References	Region
Australian Stock Exchange	RM, IC	[23], [33]	Oceania
HSBC	TS	[4]	Global
Hong Kong Stock Exchange	RM	[2], [37]	Asia
Korea Stock Exchange	TS	[38]	Asia
London Stock Exchange	D,RM,OP	[27], [30], [32]	Europe
New York Stock Exchange	CF, D	[35], [47]	America
Taiwan Stock Exchange	OP,D	[26], [60]	Asia
Shangai Stock Exchange	TS,IC	[61], [65]	Asia
Shenzhen Stock Exchange	IC	[39]	Asia
Spanish Stock Exchange	D	[48]	Europe
Switzerland Stock Exchange	D	[50]	Europe

E.1.4. Remarks

High Frequency Trading does not produce periods of extreme volatility by itself, but can exacerbate market volatility [35].

FX markets experienced a growth rate of 32.5 % in the last three years, with the United States Dollar as the most traded currency[53]. This information presents the USD behavior analysis as a still interesting research area.

Based on the variance in the number of publications by category surveyed in this paper, categories that presented growing in time were LOB's Representation and modeling, LOB's Order placement, and LOB's Trading strategies. The topic that has been more widely studied from the proposed categories, is the representation and modeling of the order book, followed by trading strategies based on it. Combining growing in time and volume of publications, suggests that guiding future research towards trading strategies based on LOB could be rewarding because is a very active branch.

The relationship between the progress in the proposed categories and the geographic market location were presented in a summary table providing evidence of which markets have been more deeply characterized in the literature.

References

[1] Ahmed, M., Chai, A., Ding, X., Jiang, Y., Sun, Y. (2009). Statistical Arbitrage in High Frequency Trading Based on Limit Order Book Dynamics, 1-26.

[2] Ahn, H.-J., Cai, J., Cheung, Y. L. (2005). Price clustering on the limit-order book: Evidence from the Stock Exchange of Hong Kong. Journal of Financial Markets, 8(4), 421-451.

[3] Hilary Arksey and Lisa O'Malley, Scoping studies: towards a methodological framework, International Journal of Social Research Methodology, Vol.8, Number 1, 2005, 19-32.

[4] Bates, R. G., Dempster, M. A. H., Romahi, Y. S. (2003). Evolutionary reinforcement learning in FX order book and order flow analysis. In 2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings. (pp. 355-362). IEEE.

[5] Bloomfield, R., O'Hara, M., Saar, G. (2005). The «make or take» decision in an electronic market: Evidence on the evolution of liquidity. Journal of Financial Economics, 75(1), 165-199.

[6] BookMap by VeloxPro. (2015, October 8). Retrieved from `http://www.bookmap.com/`.

[7] Breymann W., Dias A., Embrechts P.: Dependence structures for multivariate high-frequency data in finance. Quantitative Finance Vol. 3, Iss. 1. (2003)

[8] Chen, M.; Ebert, D.; Hagen, H.; Laramee, R.S.; van Liere, R.; Ma, K.-L.; Ribarsky, W.; Scheuermann, G.; Silver, D., "Data, Information, and Knowledge in Visualization,Çomputer Graphics and Applications, IEEE , vol.29, no.1, pp.12,19, Jan.-Feb. 2009

[9] Cheng, W., Liu, S., Jiao, H., Qiu, W. (2009). How Does Limit Order Book Information Affect Trading Strategy and Market Quality: Simulations of an Agent-Based Stock Market. In 2009 International Conference on Management and Service Science (pp. 1-4). IEEE.

[10] Christensen H. L., Turner R. E., Hill S. I., Godsill S.J.: Rebuilding The Limit Order Book: Bayesian Inference on Hidden States. Quantitative Finance. pp. 1779-1799J. (2013)

[11] Cont, Stoikov, and Talreja: A Stochastic Model for Order Book Dynamics. Operations Research Vol. 58, No. 3, May - June 2010, pp. 549 - 563 issn 0030 - 364X eissn 1526 - 5463 10 5803 0549 in.

[12] Cruz, A., Nino. J., Sandoval. J., Rincon, J., Hernandez, G.: Market Trend Visual Bag of Words Informative Patterns in Limit Order Books. International Conference on Computer Science Proceedings. San Diego, California, U.S.A. (2016)

[13] Danielsson J., Payne R.: Real trading patterns and prices in spot foreign exchange markets, Journal of International Money and Finance, Volume 21, Issue 2, April 2002, Pages 203-222, ISSN 0261-5606, `http://dx.doi.org/10.1016/S0261-5606(01)00043-2`. (2002)

[14] Lopez-Monroy A. P., Gomez, M. M., Escalante, H. J., Cruz-Roa, A. and Gonzalez, F. A. Bag-of-visual-ngrams for histopathology image classification, in Proc. of SPIE 8922, 2013, p. 89220P.

[15] Donoho, D. and Johnstone, I. Ideal spatial adaptation via wavelet shrinkage. Biometrika, 81(3):425-455, 1994.

[16] Eisler, Z., Kertesz, J., Lillo, F.: The Limit Order Book on Different Time Scales, arXiv.org, Quantitative Finance Papers 0705.4023, May 2007. [Online]. Available: `http://ideas.repec.org/p/arx/papers/0705.4023.html` (2007)

[17] Farmer, J. Doyne and Patelli, Paolo and Zovko, Ilija I., The Predictive Power of Zero Intelligence in Financial Markets (February 9, 2004). AFA 2004 San Diego Meetings.

[18] Fletcher, T., Hussain, Z., Shawe-Taylor, J. (2010). Multiple Kernel Learning on the Limit Order Book. In WAPA (pp. 167-174).

[19] Foreign Exchange Transaction Electronic System (Set-FX), `http://www.set-fx.com/index.html`

[20] Forni, M., Lippi, M. (2001). The generalized dynamic factor model: Representation theory. ECONOMETRIC THEORY, 17(6), 1113-1141.

[21] Gabor,D. Theory of communication. J. IEE, 93:429-457, 1946.

[22] Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J. and Howison, S. D. Limit order books. Quantitative Finance, Vol. 13, No. 11, 1709â€“1742. 2013.

[23] Hall, A. D., Hautsch, N. (2007). Modelling the buy and sell intensity in a limit order book market. Journal of Financial Markets, 10(3), 249-286.

[24] Harris, Zellig S. Distributional structure. Word, Vol 10, 1954, 146-162.

[25] Hayes, Adam. Forex Trading Strategy, 2021.https://www.investopedia.com/terms/forex/f/forex-trading-strategies.asp

[26] Hsin, P.-H., Wang, M.-C. (2007). Information Indicators of Limit Order Book and Optimal Dynamic Order Submission Strategy. In Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007) (pp. 197-197). IEEE.

[27] Huang, H., Kercheval, A. N. (2012). A generalized birth-death stochastic model for high-frequency order book dynamics. Quantitative Finance, 12(4), 547-557.

[28] Huang, R., Polak, T. (2011). LOBSTER: Limit Order Book Reconstruction System. Available at SSRN 1977207.

[29] Integrated Latin American Market (MILA), http://www.mercadomila.com

[30] Jian Jiang, Wing Lon Ng. (2010). Capturing order book dynamics with Kalman filters.

[31] Jiang, G., Wang, S., Dong, H. (2011). A Survey of Limit Order Book Modeling in Continuous Auction Market. In 2011 3rd International Workshop on Intelligent Systems and Applications (pp. 1-4). IEEE.

[32] Jiang, J., Ng, W. L. (2009a). Revealing Intraday Market Efficiency – Estimating Diurnal Price Densities in Limit Order Books. In 2009 International Conference on Information and Financial Engineering (pp. 8-12). IEEE.

[33] Jiaqi Wang, Zhang, C. (2006). Dynamic Focus Strategies for Electronic Trade Execution in Limit Order Markets. In The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEEâ€™06) (pp. 26-26). IEEE.

[34] Kercheval,Alec N. and Zhang,Yuan. Modelling high-frequency limit order book dynamics with support vector machines,Quantitative Finance, volume 15, number 8, pp.1315-1329. 2015.

[35] Kirilenko, A., Kyle, A. S. (2011). The Flash Crash : The Impact of High Frequency Trading on an Electronic Market.

[36] Krishnamurthy, V., Aryan, A. (2012). Quickest detection of market shocks in agent based models of the order book. In 2012 IEEE 51st IEEE Conference on Decision and Control (CDC) (pp. 1480-1485). IEEE.

[37] Lee, S.-Y., Poon, W.-Y., Song, X.-Y. (2007). Bayesian analysis of the factor model with finance applications. QUANTITATIVE FINANCE, 7(3), 343-356.

[38] Lee, W.-B., Choe, H. (n.d.-a). Short-term return predictability of information in the open limit order book. Asia-Pacific Journal of Financial Studies (2007) vol. 36, number 6, pp. 963-1007.

[39] Li, Y., Zhang, X. (2009). A Comparative Study of Information Content of Limit Order Book before and after Transparency Was Increased: Evidence from Shenzhen Stock Exchange. In 2009 International Conference on Management and Service Science (pp. 1-4). IEEE.

[40] Lopez-Monroy A. P., Gomez, M. M., Escalante, H. J., Cruz-Roa, A. and Gonzalez, F. A. Bag-of-visual-ngrams for histopathology image classification, in Proc. of SPIE 8922, 2013, p. 89220P.

[41] Mallat Stephane. A Wavelet Tour of Signal Processing: The Sparse Way. Elsevier. Third Edition. 2009.

[42] Malik, Azeem and Ng, Wing Lon, (2014), Intraday liquidity patterns in limit order books, Studies in Economics and Finance, 31, issue 1, p. 46-71.

[43] Moorhead, R.J.; Zhifan Zhu, "Signal processing aspects of scientific visualization,"Signal Processing Magazine, IEEE , vol.12, no.5, pp.20,41, Sep 1995. DOI: 10.1109/79.410438

[44] Narasimhan, Priya (Carnegie Mellon University). (2006). Fault-Tolerant Distributed Systems [Course Material]. Retrieved from https://www.ece.cmu.edu/~ece749/teams-06/team3/.

[45] NYSE Arcabook for Options Client Specification for NYSE Arca Options and Nyse Amex Options Exchanges. 2014 NYSE Euronext. Technical Report. (2014)

[46] Onorato, M., Altman, E. I. (2005). An integrated pricing model for defaultable loans and bonds. EUROPEAN JOURNAL OF OPERATIONAL RESEARCH, 163(1), 65-82.

[47] Palguna, D., Pollak, I. (2012). Non-parametric prediction of the mid-price dynamics in a limit order book. In 2012 IEEE Statistical Signal Processing Workshop (SSP) (pp. 896-899). IEEE.

[48] Pascual, R., Veredas, D. (2009). What pieces of limit order book information matter in explaining order choice by patient and impatient traders? Quantitative Finance, 9(5), 527-545.

[49] Rajaraman, Anand and Ullman, Jeffrey David. Mining of Massive Datasets. Cambridge University Press. New York, NY, USA. 2011

[50] Ranaldo, A. (2004a). Order aggressiveness in limit order book markets. Journal of Financial Markets, 7(1), 53-74.

[51] Record Neil, Currency overlay. Wiley Finance. England. 2003

[52] Russell,Jeffrey R. and Kim,Taejin .A New Model for Limit Order Book Dynamics,"Volatility and Time Series Econometrics : Essays in Honor of Robert F. Engle. Oxford ; New York: Oxford University Press, 2010.

[53] Settlements, I. (2010). Triennial Central Bank Survey Report on global foreign exchange market activity in 2010 (pp. 1-95).

[54] Sivic, J. and Zisserman., A. (2003). Video google: A text retrieval approach to object matching in videos. In Proceedings of the International Conference on Computer Vision, ICCV.

[55] Song, N., Ching, W.-K., Siu, T.-K., Yiu, C. (2012). Optimal Submission Problem in a Limit Order Boovk with VaR Constraints. In 2012 Fifth International Joint Conference on Computational Sciences and Optimization (pp. 266-270). IEEE.

[56] MERCADO Next day. Manual de usuario. SET ICAP FX. `https://set-icap.com/Manuales/Manual_de_Usuario_Next_Day.pdf` (2016)

[57] Todd, A.; Scherer, W.; Beling, P.; Paddrik, M.; Haynes, R., «Visualizations for sense-making in financial market regulation», Big Data (Big Data), 2014 IEEE International Conference on , vol., no., pp.730,735, 27-30 Oct. 2014

[58] Vasquez Linares, Mario. Gonzalez Osorio, Fabio Augusto and Hernandez Losada, Diego Fernando. Mining Candlesticks Patterns on Stock Series: A Fuzzy Logic Approach. Advanced Data Mining and Applications. Lecture Notes in Computer Science. Springer Berlin Heidelberg. 2009. pp. 661-670.

[59] Vvedenskaya, N., Suhov, Y., Belitsky, V. (2011). A non-linear model of limit order book dynamics. In 2011 IEEE International Symposium on Information Theory Proceedings (pp. 1260-1262). IEEE.

[60] Wang, M.-C., Zu, L.-P., Kuo, C.-J. (2008). The state of the electronic limit order book, order aggressiveness and price formation. Asia-Pacific Journal of Financial Studies, 37(2).

[61] Wang Yanhong, Liu Shancun. (2011). An empirical heterogeneous trading strategy model in the Shanghai stock market of China. In MSIE 2011 (pp. 227-230). IEEE.

[62] Weinberger, Kilian and Dasgupta, Anirban and Langford, John and Smola, Alex and Attenberg, Josh. Feature Hashing for Large Scale Multitask Learning. Proceedings of the 26th Annual International Conference on Machine Learning. ACM. Montreal, Quebec, Canada. 2009. pp. 1113-1120.

[63] Whigham, P. A., Withanawasam, R., Crack, T., Premachandra, I. M. (2010). Evolving trading strategies for a limit-order book generator. In IEEE Congress on Evolutionary Computation (pp. 1-8). IEEE.

[64] Yang, S., Paddrik, M., Hayes, R., Todd, A., Kirilenko, A., Beling, P., Scherer, W. (2012). Behavior based learning in identifying High Frequency Trading strategies. In 2012 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFEr) (pp. 1-8). IEEE.

[65] Yu, Y. (2006). The Limit Order Book Information and the Order Submission Strategy: A Model Explanation. In 2006 International Conference on Service Systems and Service Management (Vol. 1, pp. 687-691). IEEE.

[66] Algorithmic Trading Challenge. (2012, January 8). Retrieved from `https://www.kaggle.com/c/AlgorithmicTradingChallenge/details/Background/`.