Evolutionary Intelligence manuscript No.
(will be inserted by the editor)

Characterising Order Book Evolution Using Self-Organising Maps

Anthony Brabazon · Piotr Lipinski · Philip Hamill

Received: date / Accepted: date

Abstract Trading on major financial markets is typically conducted via electronic order books whose state is visible to market participants in real-time. A significant research literature has emerged concerning order book evolution, focussing on characteristics of the order book such as the time series of trade prices, movements in the bid-ask spread and changes in the depth of the order book at each price point. The latter two items can be characterised as order book *shape* where the book is viewed as a histogram with the size of the bar at each price point corresponding to the volume of shares demanded or offered for sale at that price. Order book shape is of interest to market participants as it provides insight as to current, and potentially future, market liquidity. Questions such as what shapes are commonly observed in order books and whether order books transition between certain shape patterns over time are of evident interest from both a theoretical and practical standpoint. In this study, using high-frequency equity data from the London Stock Exchange, we apply an unsupervised clustering methodology to determine clusters of common order book shapes, and also attempt to assess the transition probabilities between these clusters.

A. Brabazon
Natural Computing Research and Applications Group,
Smurfit School of Business,
University College Dublin, Dublin, Ireland
Tel.: +353-1-7168879
E-mail: anthony.brabazon@ucd.ie

P. Lipinski
Computational Intelligence Research Group,
Institute of Computer Science,
University of Wroclaw, Wroclaw, Poland
Tel.: +48-71-3757808
E-mail: lipinski@ii.uni.wroc.pl

P. Hamill
Emirates Institute for Banking and Financial Studies,
Dubai, United Arab Emirates
Tel.: +971-4-6070444
E-mail: philiph@eibfs.com

Findings indicate that order books for individual stocks display intraday seasonality, exhibit some common patterns, and that transitions between order book patterns over sequential time periods is not random.

Keywords Order book patterns · Ultra high-frequency financial data · Self-organising map · Unsupervised clustering

1 Introduction

Natural computing (NC) methodologies [6] have been applied extensively across a multiplicity of fields including finance [4,5]. Self-organising maps (SOMs) [22,23, 25], a subfield of NC, are a powerful clustering methodology and there has been a call for a greater awareness and use of SOM mapping techniques in the finance domain [34]. Technically, the SOM methodology implements an unsupervised clustering algorithm to uncover previously unknown relationships in complex, high-dimensional, datasets. This can provide the tangible benefit of an intuitively simple topological representation of the original data on a low-dimensional grid which is easy to interpret visually. There have been a number of applications of SOM methodologies in the finance literature including corporate failure prediction, financial stability surveillance and crisis prediction, financial trading, real-estate investment, stock selection, and mutual fund classification [13,29,36]. With the increased availability of high-frequency financial market data there has been a corresponding growth in the market microstructure literature [15,16,30,31]. However, within the specific context of market microstructure research there have been hardly any applications of a SOM methodology apart from [2].

In this paper we analyse London Stock Exchange rebuild order-book data with the aim of contributing to our understanding of two key issues. First, we investigate the intra-daily characteristics of liquidity including the size of the bid-ask spread and order-book depth for a number of large-cap securities. Then the SOM methodology is used to investigate whether the order book displays well-defined clusters of order book shapes, and whether there are predictable transitions within, and between, clusters. Enhanced understanding of order book shape dynamics could assist in active asset management and trading [15,17].

The rest of this paper is structured as follows. Section 2 provides a short introduction to limit order markets and to some of the factors which impact on order book shape. In Section 3 the dataset is discussed and some characteristics of the data are examined. Section 4 introduces the SOM methodology, followed by Section 5 which describes how this was applied to the order book data. Results are provided in Section 6 and the paper is concluded in Section 7 with some suggestions for future work.

2 Limit Order Markets

Most large equity and derivative markets operate an electronic double auction limit-order book which is visible to all market participants. In a limit order market, traders can either submit a limit or a market order. A market order, to buy or to sell a specified number of shares, guarantees immediate execution but provides no control over its execution price: a buy (sell) market order is executed at the best ask (bid) price. In contrast, a limit order is an order to buy or to sell a specified number of shares at a specified price. It provides control over execution price but does not guarantee execution.

The order book consists of two queues which store buy and sell limit orders, respectively. Buy limit orders are called *bids* and sell limit orders are called *offers* or *asks*. The highest bid price on the order book is called *best bid*, and the lowest ask price on the order book is called *best ask*. The difference between best bid and best ask is called *bid-ask spread*. Prices on the order book are quoted in discrete quanta called *ticks* (one cent or one penny in the case of the US and the UK).

Figure 1 displays a sample limit order book. The volume of shares available for sale / purchase is represented by the height of the bar at each price point (x-axis). The order book can be described at varying 'depths', referring to the number of price points (and associated share volumes) illustrated.

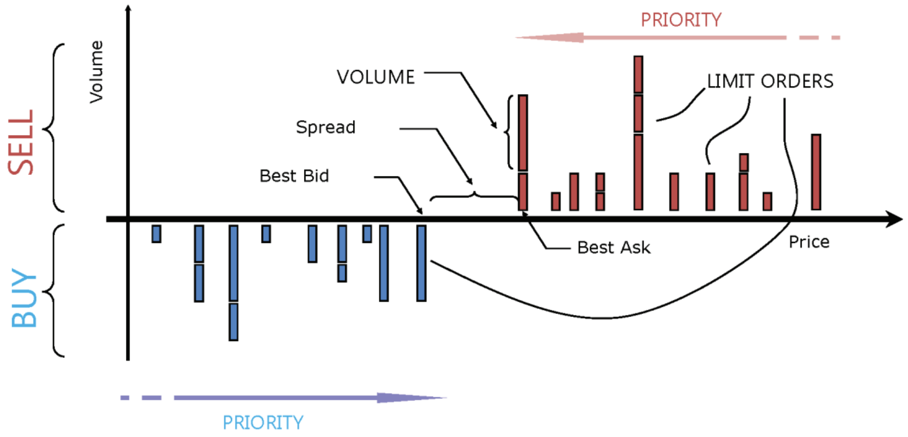


Fig. 1 Illustration of limit order book with stylised representation of best bid, best ask, spread and volume of limit orders at each price point

In a limit order market, orders arrive randomly in time. The price limit of a newly arrived order is compared to those of orders already held in the system to ascertain if there is a match. If so, the trade occurs at the price set by the first order. The set of unexecuted limit orders held by the system constitutes the dynamic order book, where limit orders can be cancelled or modified at any time prior to their execution. Limit orders on the order book are typically (depending on market rules) executed strictly according to (1) price priority and (2) time priority. Bid (ask) orders with higher (lower) prices get executed first with time of placement being used to break ties. A buy (sell) market order is executed at the best ask (bid) price. The limit order book is highly dynamic because market participants constantly submit and withdraw orders to buy or sell quantities of a financial security. As such, they provide (imperfect) insight into the instantaneous demand and supply curves facing market participants.

2.1 Market Depth

Market depth is indicated by the quantity of a stock on offer at each tick level in the order book. Hence, depth could be measured at the best bid / ask level, or more comprehensively by considering multiple order book levels. In considering market depth, a closely related item is the 'shape' of the order book. There are two aspects to this including the volume offered at each tick on the bid / ask side of the book and the degree of symmetry (or 'imbalance') between both sides of the book.

2.2 Order Type and Order Book Shape

There are a large number of studies analysing factors which impact on a trader's use of limit and market orders, both of which directly impact on the shape of the order book. One vibrant strand of this work concerns analysis of the relationship between order aggressiveness (i.e., how closely a trader places a limit order to the current best bid or ask, with closer placement being termed 'more aggressive') and the current state of the order book. As this research seeks to explain the willingness of traders to use limit orders, and their placement of limit orders on the book (i.e., close to or further away from the current bid-ask), it is relevant in the context of our understanding of the evolution of order book depth.

It has been reported that limit orders submitted at the best quote or inside the spread have lower trading costs than market orders [21]. Similarly, Griffiths, Turnbull and White (2000) [18] recommend placing buy (sell) limit orders at the best bid (ask) as an optimal strategy for minimising the implementation shortfall. As far as the probability of a limit order execution is concerned there is evidence that it is primarily determined by the distance of the limit price from the best quote. Orders closer to the best quote have a higher execution probability and a shorter time to execution [11, 27]. Lo and Sapp (2010) [28] find that traders are less likely to submit market orders, than limit orders at the best price and limit orders improving the best price.

It has been found that if one side of the order book is dominant, where the dominant side is the one with more depth, then limit orders on the dominant side take longer to execute [1] and have a higher risk of an adverse price movement leading to non-execution. Consequently, traders on the same side of the market as the dominant side of the book are more likely to submit market orders to achieve an immediate execution [9, 18, 32, 35, 37].

Traders are more willing to place market orders when the market depth on the same side of the order book is large. If the market depth on the opposite side is larger, traders prefer to submit limit orders [8, 14, 35, 39]. When the bid-ask spread widens, traders prefer to submit limit orders in order to avoid large bid-ask spread cost [3, 8, 14, 33, 35, 37, 39].

Other explanatory factors considered in previous literature include time of day effects as it is known that liquidity and price impact vary across the trading day [38].

In summary, trader decisions as to whether to place market or limit orders and in the case of limit orders, the degree of price aggressiveness to employ when placing the limit order, are impacted by multiple factors. Prior research indicates that the current state (shape) of the order book, by providing a proxy for market liquidity, influences these decisions. In turn, it may therefore have predictive relevance concerning future order book state.

3 Data

In this section we describe the data used in this study, the preprocessing steps required, and the results of a series of initial explorations of the dataset.

3.1 The Data Set

The London Stock Exchange (LSE) is the third largest exchange in the world and the largest in Europe in terms of market capitalisation. It operates an electronic order book platform, Stock Exchange Electronic Trading System (SETS). This study uses data from the LSE Rebuild Order Book (ROB) and this allows the reconstruction at a millisecond level of all order submissions, order cancellations, order modifications and executions on the LSE for the period of the data. The LSE provides three separate files to record trade and order information for each trading day, namely:

1. an ‘order details’ file, which contains details of every new order entering the electronic order book,
2. an ‘order history’ file, which contains information on changes to each order, including order partial/full matching, order deletion, order expiration and order modification, and
3. a ‘trade report’ file, which contains details of every trade execution.

Each ‘event’ such as a trade execution or the placement of an order is assigned a unique code label, a combination of numbers and letters, by the LSE. The order code maps the new orders listed on the ‘order details’ file and their order trajectories (full execution, partial execution, deletion, expiration and modification) recorded on the ‘order history’ file. Similarly, the trade code is used to track the trading details contained on the ‘trade report’ file for each order matching record on ‘order history’ file.

The ROB data is provided in a raw state and a considerable amount of preprocessing is required in order to reconstitute the daily order book sequences. Certain information must be inferred from the files including market orders and three missing events. Market orders can be inferred from the ‘order history’ file as it records information on the matching side of each transaction. The first group of missing events in ROB data are the execution of iceberg limit orders. LSE allows traders to place iceberg limit orders, part of which are hidden in the order book and not recorded in the ROB data. When the visible part of the iceberg limit order is matched by a market order with larger size, the hidden part will be executed against the rest of the market order. The traded hidden part of the iceberg limit order can be inferred from the records of the limit order whose transaction size is larger than its original size. The second set of missing events are crossing limit orders which are traded immediately after submission. The 'order details' file only records the unexecuted part of the crossing limit orders. The traded part can be found from the 'order history' file in that each crossing limit order is matched by one or more limit orders previously submitted to the market. The third group of missing events are old limit orders which were submitted to the market in prior trading days but executed today. The information on details of these orders needs to be recovered from older data files.

3.2 Descriptives From Dataset

Initially, we extract some intraday descriptives from the dataset which are relevant to liquidity. In this analysis we focus attention on four of the most actively-traded, large cap, stocks during the period of analysis (127 trading days from 1 April 2010 - 30 September 2010). The selected stocks are British American Tobacco, BP, Glaxo Smith Kline, and HSBC Holdings (Table 1). Two aspects of the data are examined for each company, the:

1. intraday bid-ask spread, and the
2. intraday market depth.

Ticker	**Name**	**Sector**	**Market Capitalisation**
BATS	British American Tobacco PLC	Consumer Goods	£61.80bn
BP	BP PLC	Oil and Gas	£83.92bn
GSK	Glaxo Smith Kline PLC	Healthcare	£70.95bn
HSBC	HSBC Holdings PLC	Financial	£98.81bn

Table 1 Company Information

In all cases the reported metrics are measured at 15-minute intervals on each trading day and are averaged for this time slot over the period of 127 trading days. We consider the intraday patterns for each stock separately.

3.2.1 Intraday Bid-ask Spread

The intraday proportional bid-ask spreads are shown in Figure 2. This metric is defined as the ratio of the prevailing bid-ask spread and the prevailing mid-quote price $(0.5 * (ASK + BID))$:

$$ProSpr = \frac{ASK - BID}{0.5 * (ASK + BID)}.$$

The 15-minute proportional quoted spread is calculated as the average quoted spread during the period.

The spread is high at the opening in all four stocks. It remains stable over the rest of the day in BATS and GSK, but with one peak around 10:00am in BP and HSBC. The figures indicate that individual stocks do not have identical patterns of bid-ask spread behaviour during the trading day.

Unlike the findings in [7], the spread does not rise towards the end of the trading day. Instead, it decreases at the market close in BP and HSBC. Chen and Cai (2008) [10] suggest that an L-shape pattern on bid-ask spread may arise due to the lack of participation by informed traders. They speculate that there is heavier trading by noise traders at the opening of the market due to information dissemination overnight, with the spread being widened to reflect potential adverse selection risk. At the end of the trading day, liquidity traders are replaced by informed traders. As the private information is absorbed into the market price, the spread becomes smaller.

3.2.2 Intraday Market Depth

We report the market depth over the first ten ticks on each side (bid / ask) of the order book.The near market depth ($NDep$) is measured as the average number of all shares on the bid and ask side as follows:

$$NDepShare = \frac{\sum_{i=1}^{10} Shares_{bid,i} + \sum_{i=1}^{10} Shares_{ask,i}}{2}$$

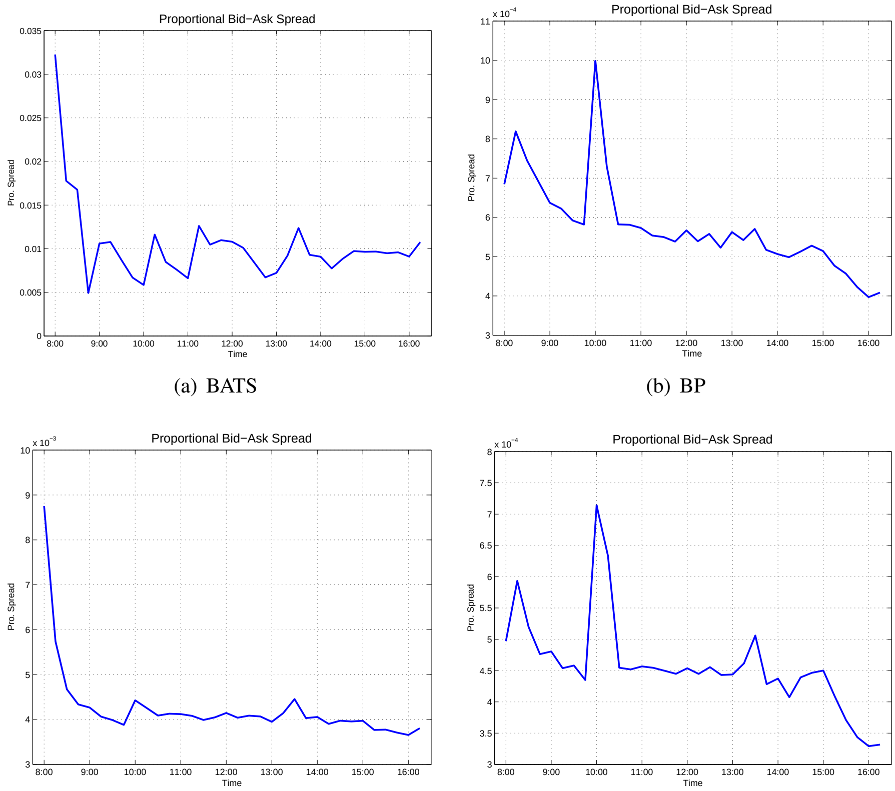
where i is the level of the order book. Depth is averaged over each 15-minute interval and over the whole sample period. As shown in Figure 3, the near market depth increases over the trading day for all four stocks.

In summary, the analysis of bid-ask spread and market depth for a number of large cap shares on the LSE indicates that spread and depth are not constant across the trading day. Neither is it clear from the analysis that all equities follow the same evolution of these characteristics during the trading day. This suggests that the liquidity profile (and order book shape) needs to be analysed at an individual equity level.

4 Self-Organising Maps

Self-organising maps (SOMs) [22–25] are loosely inspired by the self-organising capability of neurons in the cortex. Specifically, SOMs are artificial neural nets which use unsupervised learning to adapt (organise) themselves in response to signal inputs. SOMs have been utilised for a large variety of clustering and classification applications in domains as diverse as speech recognition, medical diagnosis and finance [13, 19,24].

A SOM consists of two layers, an input layer which serves as a holding point for the input data, and a *mapping* layer represented as a low-dimensional grid (typically two dimensions are used). The two layers are fully connected to each other with each



(a) BATS

(b) BP

(c) GSK

(d) HSBC

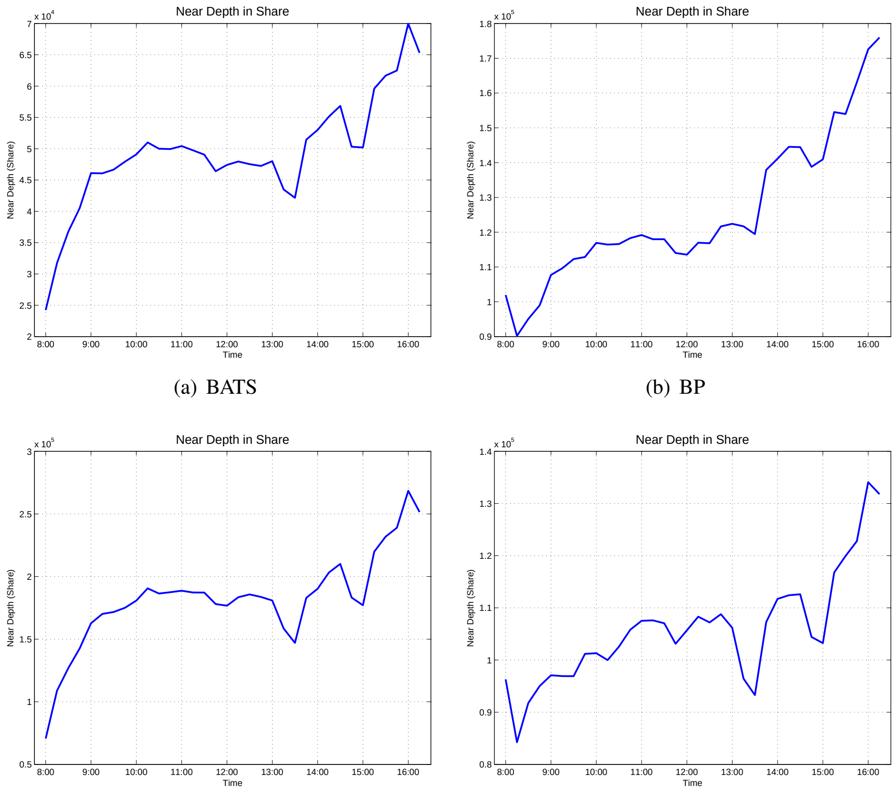
Fig. 2 Intraday Proportional Bid-ask Spread

mapping layer node having an associated weight vector, with one weight for each connection with the input layer. The vector associated with each mapping layer node (referred to as a ‘codebook vector’) positions (or ‘embeds’) each of these nodes in the input space. Accordingly, a mapping layer node can be considered as a form of ‘detector’ with its weight vector ‘pointing to’ a location in the space of input data.

A SOM acts to project (compress) the input data vectors onto the low-dimensional mapping layer. The aim of the SOM is to group ‘like’ input data vectors close together in the mapping layer and the method is therefore *topology preserving*. An input vector will be mapped to the mapping layer node whose weights are most similar to the values of the input vector.

4.1 SOM Algorithm

Algorithm 1 outlines the general training algorithm for a SOM. In this algorithm:



(a) BATS

(b) BP

(c) GSK

(d) HSBC

Fig. 3 Intraday Near Depth (Shares)

- n is the dimension of the input data vectors: $x = (x_0, x_1, \ldots, x_{n-1})$, so there are n input nodes;
- there are m mapping layer nodes: m is determined during the initialisation phase;
- each node $i \in \{1, \ldots, m\}$ in the mapping layer has a weight vector $w_i = (w_{i0}, w_{i1}, \ldots, w_{i,n-1})$, so there are $n \times m$ weights in total;
- α is the learning rate of the map; and
- $h_{i^*,k}$ defines a neighbourhood function from mapping layer centre i^* to neighbour k, e.g., a radially symmetric Gaussian $h_{i^*,k}(t) = \exp\left(\frac{-\|r_k - r_{i^*}\|^2}{2\sigma(t)^2}\right)$ where $\|r_k - r_{i^*}\|$ is the distance between node k and the best matching unit (BMU) i^* in the two dimensional grid (r_k is the position vector of mapping node k in $\mathbb{R}^2$).

Note that eq. (2) in the SOM algorithm refers to a vector update: all n components of w_k are updated simultaneously. Both the neighbourhood size and the learning rate decay during the training run, in order to fine-tune the developing SOM. The neighbourhood size controls how many mapping layer nodes are adjusted (learn) as each data vector is presented during training.

Algorithm 1: Self-Organising Map Algorithm

```
Choose the number $m$ of mapping layer nodes;
Initialise the weight vectors for each of these nodes;
$t$=1 (time step counter) ;
repeat
    for each vector $x = (x_0, x_1, \ldots, x_{n-1})$ in the training dataset do
        for each mapping layer node $i \in \{1, \ldots, m\}$ do
            Calculate the distance between the training vector $x$ and the weight
            vector $w_i$ using
                $$d_i = \sum_{j=0}^{n-1} (x_j - w_{ij}(t))^2 \qquad (1)$$
        end
        Select the mapping node $i^*$ that has the minimum value of $d_i$;
        for each neighbouring mapping node $k$ of $i^*$, including $i^*$ itself do
            Update the weight vector for node $k$ using
                $$w_k(t+1) := w_k(t) + \alpha(t) h_{i^*,k}(t)(x - w_k(t)) \qquad (2)$$
        end
    end
    $t$=$t$+1;
until weight vectors stabilise;
```

4.1.1 Training the SOM

Initially a large neighbourhood size is used and most mapping layer nodes respond as data vectors are presented. The neighbourhood size reduces as the algorithm iterates and fewer mapping layer nodes respond to each input vector. Therefore, in the early stages of the algorithm, a broad brushstroke picture of the input data distribution is learned with finer details being filled in as the neighbourhood size shrinks in later iterations of the algorithm.

As the calculation of a distance metric is required in SOM training, input data vectors are typically standardised before the algorithm commences. During training, a sample vector is drawn randomly from the input data set. The nodes in the mapping layer *compete* for the input data vector and the winner is the mapping node whose vector of incoming connection weights most closely resembles (is nearest to) the input data vector. The winner, or *best-matching unit* (BMU), has the values of its weight vector adjusted to move them towards the values of the input data vector, thereby moving the location of the BMU towards the location of the input data item. An important component of the training process is that not only the BMU, but also its neighbouring nodes on the mapping layer are adjusted in each training iteration. These neighbouring nodes also have their weight vectors altered to become more like the input data vector. The training process is unsupervised as it does not use any explicit outputs. The process is based solely on measures of similarity between the input data vectors and weight vectors associated with each of the nodes on the SOM's mapping layer.

As more input data vectors are passed through the network, the weight vectors of the mapping layer nodes will self-organise. By the end of the training process, different parts of the mapping layer will respond strongly to specific regions of input space. The self-organisation process also encourages the mapping layer weight vectors to congregate to regions of the input space where the training data is concentrated, with relatively few (if any) weight vectors being located in sparsely populated regions of the input space. The self-organising map therefore tends to approximate the probability density function of the input data.

5 Experimental Design

In this section we outline how the SOM methodology was applied in this study.

5.1 Data For SOM Experiments

Data was drawn for the four equities discussed in Sect. 3 (British American Tobacco, BP, Glaxo Smith Kline, and HSBC Holdings) for the month April 2010. Table 2 provides some descriptives concerning the number of limit orders placed and the number of transactions over the chosen one-month period for the selected stocks. From this it is clear that many more limit orders are placed than are ever executed (i.e., many limit orders are cancelled prior to execution). For instance, 1 166 898 limit orders were recorded for BATS during the month whereas only 116 413 trades were executed. Across the entire LSE database of about 2 000 equities, some 282 877 693 limit orders were placed placed during the month with 11 872 593 trades being executed (i.e., overall only 4.2% of all orders placed on the order book were executed). The volume of orders and trades across all equities in the LSE results in a large dataset with some 23GB of data for the month.

Ticker	Name	Number of Limit Orders	Number of Trades Executed
BATS	British American Tobacco PLC	1 166 898	116 413
BP	BP PLC	2 211 549	288 062
GSK	Glaxo Smith Kline PLC	822 722	119 813
HSBC	HSBC Holdings PLC	2 119 339	282 979
Entire Market		282 877 693	11 872 593

Table 2 Order Vs. Trade Data

5.2 Order Book Shape

The shape of the order book is defined as follows. For a given time t, a snapshot of the current order book at time t contains two lists, the list of the bid orders and the list of ask orders, placed on the market and not deleted until time t. Orders are aggregated in such a way that the orders with the same price are displayed together with the total volume (separately for each bid and ask list). Let h_B be the number of aggregated price levels in the bid list and $(p_1^{(B)}, v_1^{(B)})$, $(p_2^{(B)}, v_2^{(B)})$, …, $(p_{h_B}^{(B)}, v_{h_B}^{(B)})$ denote the aggregated list of price levels concerning bid orders, sorted from the highest price $p_1^{(B)}$ to the lowest price $p_{h_B}^{(B)}$, where $p_i^{(B)}$ is the price and $v_i^{(B)}$ is the volume of the i-th aggregated bid order. Similarly, let h_A be the number of aggregated price levels in the ask list and $(p_1^{(A)}, v_1^{(A)})$, $(p_2^{(A)}, v_2^{(A)})$, …, $(p_{h_A}^{(A)}, v_{h_A}^{(A)})$ denote the aggregated list of price levels concerning ask orders, sorted from the lowest price $p_1^{(A)}$ to the highest price $p_{h_A}^{(A)}$, where $p_i^{(A)}$ is the price and $v_i^{(A)}$ is the volume of the i-th aggregated ask order.

We define the order book shape of the size $2h$, for some $h \in \mathbb{N}$, as the vector

$$\mathbf{s} = (s_h^{(B)}, s_{h-1}^{(B)}, \dots, s_1^{(B)}, s_1^{(A)}, \dots, s_{h-1}^{(A)}, s_h^{(A)}) \in \mathbb{R}^{2h}, \tag{3}$$

where

$$s_k^{(B)} = \frac{v_k^{(B)}}{\bar{v}}, \quad s_k^{(A)} = \frac{v_k^{(A)}}{\bar{v}}, \quad \text{for } k = 1, 2, \dots, h, \tag{4}$$

and

$$\bar{v} = \sum_{i=1}^{h_B} v_i^{(B)} + \sum_{i=1}^{h_A} v_i^{(A)}. \tag{5}$$

This characterisation of the order book allows the SOM to consider both relative volume at each price level (tick) and the degree of symmetry (or ‘imbalance’) between both sides of the book.

5.3 Discovering Shape Patterns

As defined in the previous section, the shape of an order book may be represented by a vector of length $2h$, i.e., a point in the data space $\mathbb{R}^{2h}$. Grouping such data points in the data space corresponds to dividing order book shapes into clusters containing similar shapes. Further, finding representatives of each group of data points corresponds to defining common patterns, i.e., order book shapes, which are ‘similar’ to other shapes in their cluster.

In the proposed approach, clustering of order book shapes in the data space $\mathbb{R}^{2h}$ is performed by a SOM composed of M codebook vectors, $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M \in \mathbb{R}^{2h}$ embedded in the data space and a topology structure in a form of a regular two dimensional hexagonal lattice defining a neighbourhood relation for the codebook vectors. Such codebook vectors, connected to each other according to the neighbourhood relation, form an elastic net embedded in the data space, which will be adjusted to the given data sample in the training process in order to fold onto the cloud of given data points.

In developing the SOM, each order book shape was defined using a a vector of length $d = 30$ (i.e., the volumes at 15 price levels on each side of the book), resulting in a data space of $\mathbb{R}^{30}$. The codebook vectors are initially placed in a regular grid on the plane defined by the two first principal components of the given data sample. Next, in successive iterations of the training algorithm, one data point $\mathbf{s}$ is randomly picked from the given data sample, and the distances between each codebook vector $\mathbf{v}$ and the picked data point $\mathbf{s}$ are evaluated and the nearest codebook vector $\mathbf{u}$ (the BMU), is determined. The BMU and its neighbor codebook vectors are moved towards the picked data point $\mathbf{s}$ using the following formula

$$\mathbf{v} := \mathbf{v} + \alpha \cdot \theta(\mathbf{s}, \mathbf{u}, \mathbf{v}) \cdot (\mathbf{s} - \mathbf{v}), \tag{6}$$

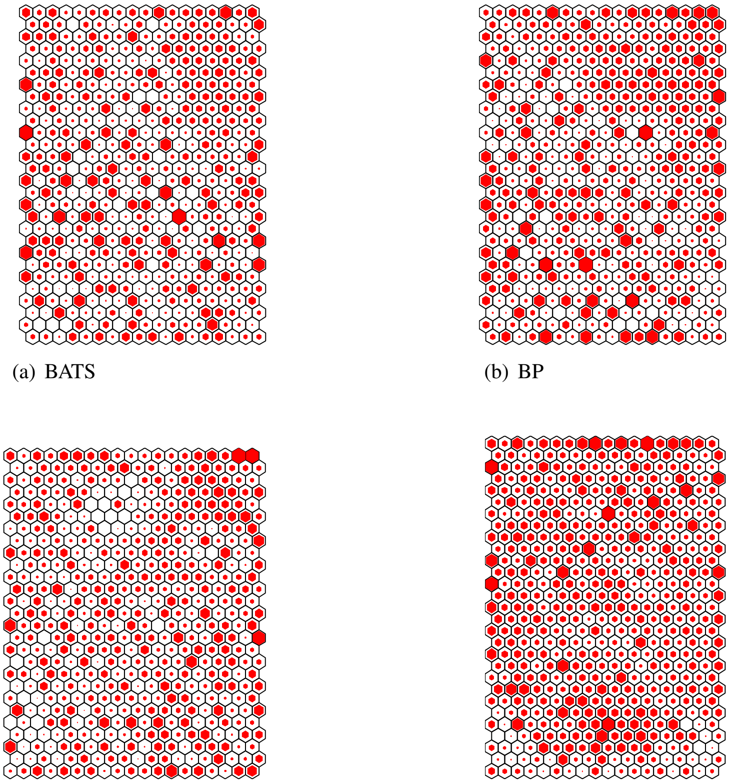
where $\mathbf{v}$ is the codebook vector being updated, $\mathbf{s}$ is the picked data point, α is a learning coefficient decreasing in successive iterations of the training algorithm and $\theta(\mathbf{s}, \mathbf{u}, \mathbf{v})$ is a neighborhood function (Gaussian in this study) determining the size of the movement, usually $\theta(\mathbf{s}, \mathbf{u}, \mathbf{v})$ is larger for $\mathbf{v}$ close to $\mathbf{u}$ according to the neighborhood relation and smaller for other codebook vectors. The SOM algorithm repeats, adjusting the net to the data sample until a terminating condition is reached. Assigning each data point from the data sample to the nearest codebook vector defines a clustering of the given data sample, where data points assigned to the same codebook vector constitute a cluster.

6 Results from SOM Experiments

In this section we provide the results of our two sets of experiments. Initially, we examine the clustering of order book patterns found in the data. We then examine the evolution of patterns within and between these clusters.

6.1 Frequent Order Book Patterns

An illustration of order book clustering may be found on Figure 4(a). It presents a hits diagram with a hexagonal topology of the size 29×18 for a SOM with 504 codebook vectors created for clustering $10\,641$ unique order book shapes of the BATS company. In the hits diagram, each cell corresponds to one codebook vector and adjoining cells correspond to neighbour codebook vectors. Filling of the cells depends on the number of order book shapes assigned to the codebook vector, such that an empty cell means that no order book shapes were assigned to the codebook vector, while a full cell means that many order book shapes were assigned to the codebook vector. One can see that there are a number of large clusters which group many order book shapes and these define characteristic common order book patterns. There are also a number of small clusters which group a smaller number of order book shapes. Thus, these clusters contain less common or ‘atypical’ order book shapes. Figures 4(b) - 4(d) presents results obtained for other three stocks. Taken together, the results suggest that order book patterns, although infinite in possible shape, can be represented in a relatively compact number of clusters for each stock.



(a) BATS

(b) BP

(c) GSK

(d) HSBC

Fig. 4 Hits diagram for all four stocks. This presents the results of the clustering of the order book shapes of the four companies using a Self-Organizing Map of about 504 codebook vectors with a hexagonal topology. The degree of fill in each cell in the diagram indicates the number of order book shapes in that cluster

Each codebook vector, being a representative of the order book shapes assigned to the cluster, defines an order book 'pattern'. The cluster includes all the order book shapes that are closer to the codebook vector of that cluster than to that of another. Order book shapes within a cluster will vary and not all will necessarily appear identical. Taking one of the stocks (BATS) we illustrate in Figure 5 the results of investigating the largest clusters and examining the degree of similarity between a certain number of order book shapes nearest to the codebook vector of the cluster. The top row refers to the largest cluster, the middle row to the second largest cluster, and the bottom row to the third largest cluster obtained for the BATS company. The three columns in Figure 5 present the 1st, 5th and 10th order book shape nearest to the codebook vector of the cluster. The order book shapes within each cluster bear some similarity and it is notable that there are quite a number of limit orders placed a few price points away from the best bid and best ask respectively.

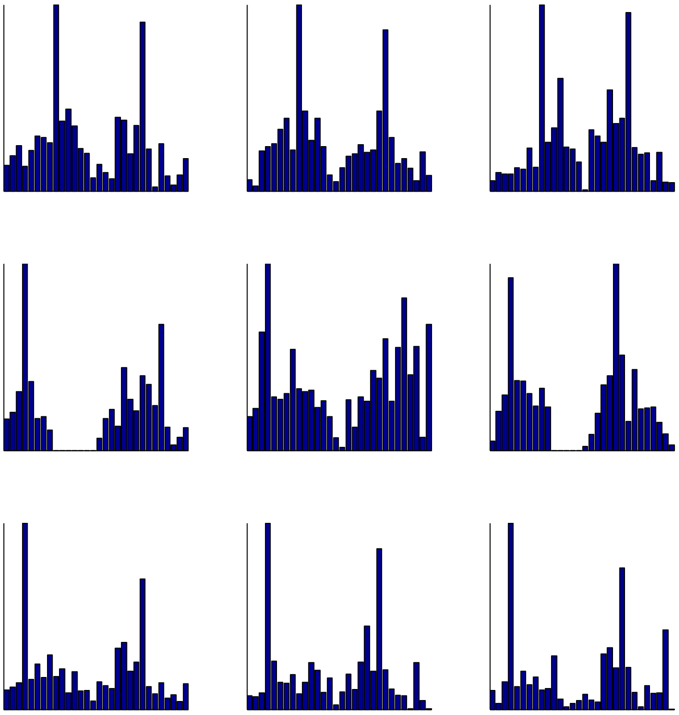


Fig. 5 Order book shapes from the 1st, 2nd and 3rd largest cluster: the 1st, 5th and 10th order book shape nearest to the codebook vector of the cluster for the BATS company.

Although self-organising maps discovered a partition of the data space into the given number of clusters, some of these clusters may be very close and rather similar, so they might be joint together to form larger clusters of order book snapshots. In the proposed approach, we do not focus on optimising the granularity of the clustering, because it would mainly affect the computational complexity of the approach without significant changes to the financial knowledge discovered from order book snapshots.

6.2 Transitions between Order Book Patterns

Next we examine the transitions between order book shapes over time. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{50}$ denote codebook vectors of the 50 largest clusters. Let $\mathbf{s}_1^{(k)}, \mathbf{s}_2^{(k)}, \ldots, \mathbf{s}_{100}^{(k)}$ denote the 100 order book shapes nearest to the codebook vector $\mathbf{v}_k$, for each $k = 1, 2, \ldots, 50$.

At the beginning, for each cluster $k = 1, 2, \ldots, 50$ and for each successive shape $i = 1, 2, \ldots, 100$ in the cluster, one may determine the time $t_0^{(k,i)}$ when the order book shape occurred. Let $t_l^{(k,i)} > t_0^{(k,i)}$ denote the earliest time after $t_0^{(k,i)}$ when any shape of the l-th cluster occurred ($t_l^{(k,i)} = -1$ if no shape from the l-th cluster occurred after $t_0^{(k,i)}$), for each $l = 1, 2, \ldots, 50$. Let $\delta_{i,k,l} = 1$ if $0 < t_l^{(k,i)} - t0^{(k,i)} < 300$, i.e. if the first shape from the l-th cluster occurred within 5 minutes after the i-th shape from the k-th cluster, and $\delta_{i,k,l} = 0$ otherwise.

Therefore, one may evaluate a transition probability matrix $\mathbf{T} \in \mathbb{R}^{50\times 50}$ with elements t_{kl} as

$$t_{kl} = \frac{1}{100} \sum_{i=1}^{100} \delta_{i,k,l}, \tag{7}$$

which corresponds to the frequency of the fact that after a shape from the k-th cluster, a shape from the l-th cluster will occur within 5 minutes.

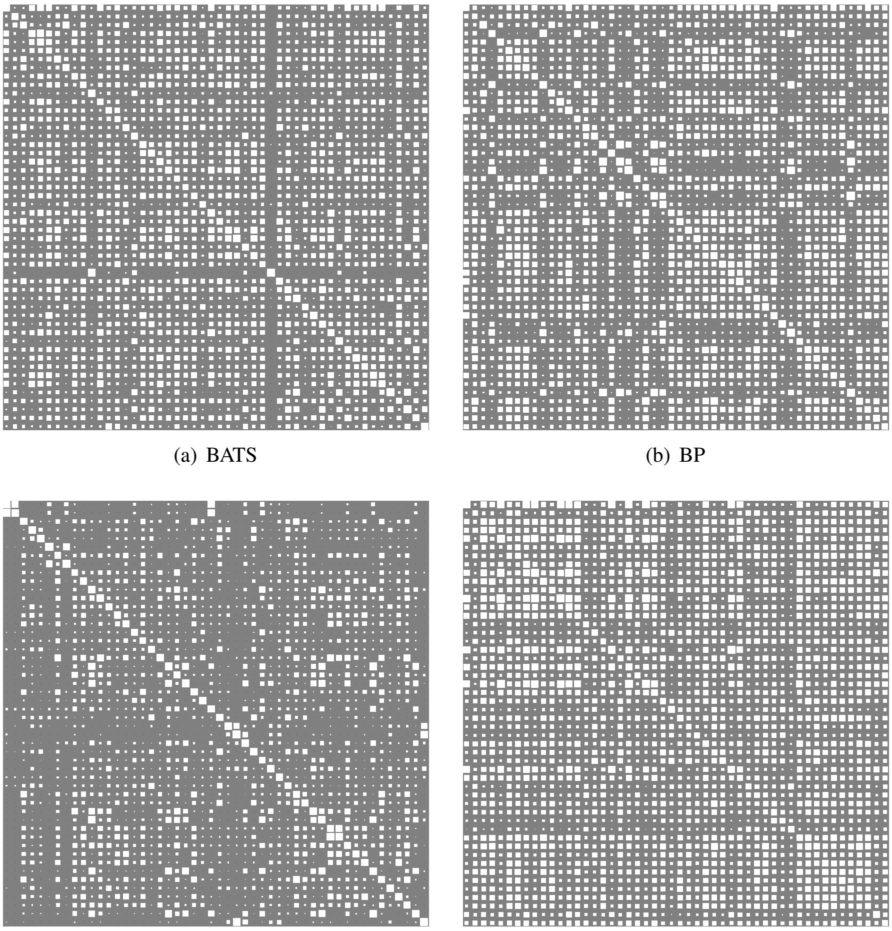
Figure 6(a) presents the transition probability matrix $\mathbf{T}$ for the BATS company in the form of a Hinton diagram, which is a compact way of visualising numerical values in a matrix. A Hinton diagram typically presents a numerical matrix as a matrix of coloured cells, where each cell visualises the value of the corresponding cell of the matrix: the size of each cell is proportional to the value, and the colour (usually black or white) indicates the sign (negative / positive) of the value in the corresponding cell of the original matrix. In this case, all values are constrained into the range $0 \rightarrow 1$ as each cell of the Hinton diagram corresponds to a cell of the transition probability matrix $\mathbf{T}$. The size of the square in the cell of the Hinton diagram corresponds to the value in the cell of the transition probability matrix $\mathbf{T}$ (large white squares indicate values close to 1, whereas small squares indicate values close to 0).

Figures 6(b) - 6(d) presents similar results for the other three stocks. The results suggest that the probabilities of transitions between some patterns during a 5 minute time horizon are quite high, which is potentially useful information for trading purposes.

In order to verify that the transition probability matrix in each case is not random, Pearson's chi-squared test was performed with a null hypothesis that the probability matrix is random with the uniform distribution for each transition probability matrix. The values of the test statistics obtained for each company are, 3,126.3680 (BATS), 3,048.4880 (BP), 8,398.2720 (GSK) and 2,763.3360 (HSBC) respectively. In all instances the values significantly exceed the critical value for the chi square test at the 1% significance level, thereby rejecting the null hypothesis.

7 Conclusions and Perspectives

In this paper we undertake an analysis of information from the LSE order book for a number of large-cap equities in order to gain insight into evolution of the order book over time. Initially, we examine the intraday seasonality of a number of aspects of liquidity including the size of the bid-ask spread and order book depth. We also present initial research exploring order book shape using SOMs which seeks to determine whether there are clusters of frequent patterns in order book shapes and whether the order book patterns for individual equities transitions from one cluster to another in a non-random fashion. The findings indicate that order books exhibit notable intraday



(a) BATS

(b) BP

(c) GSK

(d) HSBC

Fig. 6 The Hinton diagram of the transition probability matrix for the 50 largest clusters of order book shapes for each company

seasonality, that clusters of order book shapes can be identified, and that the transition of order book shape is not random. These findings have important implications for the design of trading systems and open up considerable scope for further research

As is commonly the case in an exploratory study, further research questions are opened up. While an SOM methodology can detect clusters of order book shapes, additional investigation is required to improve the clustering approach in order to increase the robustness of the developed clusters and to ensure that they are meaningful to a domain expert. In particular, a pre-processing step to remove outlier order book patterns may be useful. Further work is also required in order to ascertain the most appropriate trade-off between number of clusters and cluster granularity.

We also note that many alternative clustering methodologies exist and future work could usefully investigate the utility of other clustering approaches for order book pattern discovery. In this paper we adopted an SOM methodology to cluster the order book snapshots, because of the complex structure of the data under study. Compared to simple clustering methods, such as the k-means algorithm or its extensions, SOMs are capable of detecting irregular clusters partitioning the data space into a large number of Voronoi cells using learning mechanisms that try to find representatives of even small groups of data points. In contrast, simpler clustering methods are less flexible and less sensitive to such data irregularity. Future research could extend this study by examining the results which can be obtained using other clustering methods, particularly density-based clustering approaches.

As seen in Sect. 3, order books display intraday seasonality concerning liquidity. Future work could also investigate whether this needs to be taken into account in the clustering process in order to best develop transition matrices for order book shape.

Acknowledgements This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 08/SRC/FM1389. Calculations have been carried out using resources provided by Wroclaw Centre for Networking and Supercomputing (http://wcss.pl), grant No. 405. The authors would also like to acknowledge the contribution of the anonymous reviewers to the improvement of this paper and the work of Dr Wei Cui on the analysis of the descriptives in Section 3.

References

1. Al-Suhaibani M, Kryzanowski L (2000) An exploratory analysis of the order book, and order flow and execution on the Saudi stock market. Journal of Banking and Finance 24(5):1323–1357
2. Ben Omrane W, de Bodt E (2007) Using self-organizing maps to adjust for intra-day seasonality. Journal of Banking & Finance 31(6):1817-1838
3. Biais B, Hillion P, Spatt C (1995) An empirical analysis of the limit order book and the order flow in the Paris Bourse. Journal of Finance 50(3):1655–1689
4. Brabazon A, O'Neill M (2006) Biologically Inspired Algorithms for Financial Modeling. Springer, Berlin
5. Brabazon A, O'Neill M, Dempsey I (2008) An Introduction to Evolutionary Computation in Finance. IEEE Computational Intelligence Magazine 3(4):42–55
6. Brabazon A, O'Neill M, McGarraghy S (2015) Natural Computing Algorithms. Springer
7. Cai C, Hudson R, Keasey K (2004) Intraday bid-ask spreads, trading volume and volatility: recent empirical evidence from the London Stock Exchange. Journal of Business Finance and Accounting 31(5):647-676
8. Cao C, Hansch O, Wang X (2008) Order Placement Strategies In A Pure Limit Order Book Market. Journal of Financial Research 26(2):113–140
9. Cao C, Hansch O, Wang X (2009) The information content of an open limit order book. Journal of Futures Markets 29(1):16–41
10. Chen T, Li J, Cai J (2008) Information content of inter-trade time on the Chinese market. Emerging Markets Review 9(2):174-193
11. Cho J, Nelling E (2000) The probability of limit-order execution. Financial Analysts Journal 56(5):28–33
12. Chung K, Van Ness B, Van Ness R (1999) Limit orders and the bid-ask spread. Journal of Financial Economics 53(3):255–287
13. Deboeck G, Kohonen T (1998) Visual Explorations in Finance With Self-Organizing Maps. Springer
14. Duong H, Kalev P, Krishnamurti C (2009) Order aggressiveness of institutional and individual investors. Pacific-Basin Finance Journal 1(4):1–14
15. Easley D, Lopez de Prado L, OHara M (2016) Discerning information from trade data. Journal of Financial Economics 120(2):269–285

16. Goldstein M, Kumar P, Graves F (2014) Computerized and High Frequency Trading. The Financial Review 48(2):177–202
17. Gould M, Bonart J (2015) Queue Imbalance as a one-tick-ahead Price Predictor in a limit Order book. Available at: http://arxiv.org/pdf/1512.03492.pdf
18. Griffiths M, Smith B, Turnbull D, White R (2000) The costs and determinants of order aggressiveness. Journal of Financial Economics 56(1):65–88
19. Gurney K (1997) An introduction to Neural Networks. London: University College London Press
20. Hall A, Hautsch N (2006) Order aggressiveness and order book dynamics. Empirical Economics 30(1):973–1005
21. Harris L, Hasbrouck J (1996) Market vs. Limit Orders - The SuperDOT Evidence On Order Submission Strategy. Journal of Financial and Quantitative Analysis 31(2):213–231
22. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological Cybernetics 43:59–69
23. Kohonen T (1990) The Self-organizing map. Proceedings of the IEEE 78(9):1464–1480
24. Kohonen T (1998) The SOM Methodology. In: Deboeck G, Kohonen, T (eds) Visual Explorations in Finance with Self-organizing Maps, pp 159–167, Berlin: Springer-Verlag
25. Kohonen, T. (2000). Self-Organizing Maps, Springer
26. Lee Y, Fok R, Liu Y (2001) Explaining intraday pattern of trading volume from the order flow data. Journal of Business Finance and Accounting 28(3):199-230
27. Lo A, MacKinlay A, Zhang J (2002) Econometric models of limit-order executions. Journal of Financial Economics 65(1):31–71
28. Lo I, Sapp S (2010) Order Aggressiveness and Quantity: How Are They Determined in a Limit Order Market? Journal of International Financial Markets Institutions and Money 20(3):223–243
29. Moreno D, Marco P, Olmeda I (2006) Self-organising maps could improve the classification of the Spanish mutual fund industry. European Journal of Operational Research 174(2):1039–1054
30. O'Hara M (1995) Market Microstructure Theory. Blackwell, Oxford
31. O'Hara M (2015) High frequency market microstructure. Journal of Financial Economics 116(2):257–270
32. Omura K, Tanigawa Y, Uno J (2000) Execution Probability of Limit Orders on the Tokyo Stock Exchange. http://ssrn.com/abstract=252588, accessed 25 April 2016
33. Pascual R, Verdas D (2009) What pieces of limit order book information matter in explaining order choice by patient and impatient traders? Quantitative Finance 9(1):527–545
34. Pöllä M, Honkela T, Kohonen T (2009) Bibliography of Self-Organizing Map (SOM) Papers: 20022005 Addendum. TKKReports in Information and Computer Science, Helsinki University of Technology, Report TKK-ICS-R24
35. Ranaldo A (2004) Order Aggressiveness in Limit Order Book Markets. Journal of Financial Markets 7(1):53–74
36. Sarlin P, Peltonen T (2013) Mapping the state of financial stability. Journal of International Financial Markets, Institutions & Money 26:46–76
37. Verhoeven P, Ching S, Ng H (2004) Determinants of the decision to submit market or limit orders on the ASX. Pacific-Basin Finance Journal 12(3):1–18
38. Wilinski M, Cui W, Brabazon A, Hamill P (2015) An Analysis of Price Impact Functions of Individual Trades on the London Stock Exchange. Quantitative Finance 15(10):1727–1735
39. Xu Y (2009) Order aggressiveness on the ASX market. International Journal of Economics and Finance 1(1):51–75