Application of self-organizing maps to clustering of high-frequency financial data

Adam Blazejewski **Richard Coggins**

School of Electrical and Information Engineering
University of Sydney,
Sydney, NSW 2006, Australia,
Email: adamb@ee.usyd.edu.au, richardc@ee.usyd.edu.au

Abstract

This paper analyzes the clustering of trades on the Australian Stock Exchange (ASX) with respect to the trade direction variable. The ASX is a limit order market operating an electronic limit order book. The order book consists of buy limit orders (bids) and sell limit orders (asks). A trade takes place if a new order arrives which matches an existing order in the limit order book. If the matched order is a bid (ask) then the trade is considered to be seller(buyer)-initiated and the trade direction variable assumes a corresponding value. We employed self-organizing maps (SOMs) to perform unsupervised clustering and visualization of four dimensional trade level data for the ten stocks on the ASX with the largest market capitalization. Trade size, the best bid and ask volumes, and a variable capturing previous trade directions were used as input variables. The visualization of the data using the SOM transformation reveals that buyer-initiated and seller-initiated trades form two distinct clusters in correspondence with non-equilibrium market conditions and elicits the main structural features of the clusters.

Keywords: High-frequency financial data, trade clustering, trade direction, self-organizing map, equities.

1 Introduction

The Australian Stock Exchange (ASX) is a limit order market operating an electronic limit order book. The electronic limit order book, further referred to as the limit order book or the book, is a mechanism for collecting, storing, and matching of buy and sell limit orders submitted by market participants, as well as a mechanism for trade execution. Apart from the ASX, the limit order book is employed by many stock exchanges around the world, for example the Paris Bourse, the Tokyo Stock Exchange, and the Singapore Stock Exchange. In the first part of this paper, covered in section two, we explain the operation of the limit order book market as implemented by the ASX. In the second part of the paper, starting from section three, we conduct an exploratory analysis of trade level data. Our research hypotheses are formulated in section three. The trade dataset is described in section four. Section five gives details on the self-organizing map algorithm applied to the trade clustering task. The charts and tables with results are presented in section six. Final remarks and potential directions for further research are given in the conclusions.

2 The Australian Stock Exchange

This section presents the main principles of the limit order book mechanism operated by the ASX. A more thorough account can be found in (Aitken, Frino, Jarnecic, McCorry, Segara & Winn 1997). The limit order book consists of two queues, called a buy (bid) side and a sell (ask) side, which store buy and sell limit orders, respectively. Each stock (security) traded on the exchange has its own set of two queues. Limit orders are orders to trade, either to buy or to sell, with a specified size (number of shares) and a limit price, which is a constraint imposed on the actual trade price. Buy orders are called bids, while sell orders are called asks (offers). The limit order price constraint requires that, in the case of bids, a trade may happen at a price no higher than the specified limit price. For asks, on the other hand, the actual trade price may not be lower than the specified limit price. A bid with the highest limit price is called the best bid. An ask with the lowest limit price is called the best ask. The best bid and the best ask have priority to trade first. If two bids (asks) have the same limit price, then the one which was entered into the book first has priority. These two rules together constitute a price and time priority rule. For reasons of either clarity or brevity limit orders may be further referred to as orders, order's size as order size or number of shares, total number of shares as share volume or volume, and limit price as price.

The price difference between the best ask and the best bid in the book is called the spread. If the spread is positive then the price in the middle of the spread, equal to an average of the best ask price and the best bid price, is defined and called the mid-point price. That price is quoted as a stock's price. Bids and asks are entered into the limit order book by market participants throughout a trading day, with prices and sizes of their choice. The orders are stored in the book until they are amended, deleted or traded. Figure 1 depicts a limit order book before, during, and after trade execution. A trade takes place when a new bid (ask) arrives with a limit price equal to or higher (lower) than the limit price of the best ask (bid) in the limit order book. Such an order, called a marketable limit order[1] and considered a trade initiator, will create an overlap between the prices of the best bid and the best ask in the book, thereby changing the spread from positive to zero or negative. The non-positive

[1] A marketable limit order is different from a market order. The latter is an unpriced order which executes immediately against the best order, and if more volume is needed, the next best orders, on the opposite side of the limit order book, until all of its volume has been traded. In practice market orders on the ASX are implemented via appropriately priced marketable limit orders.

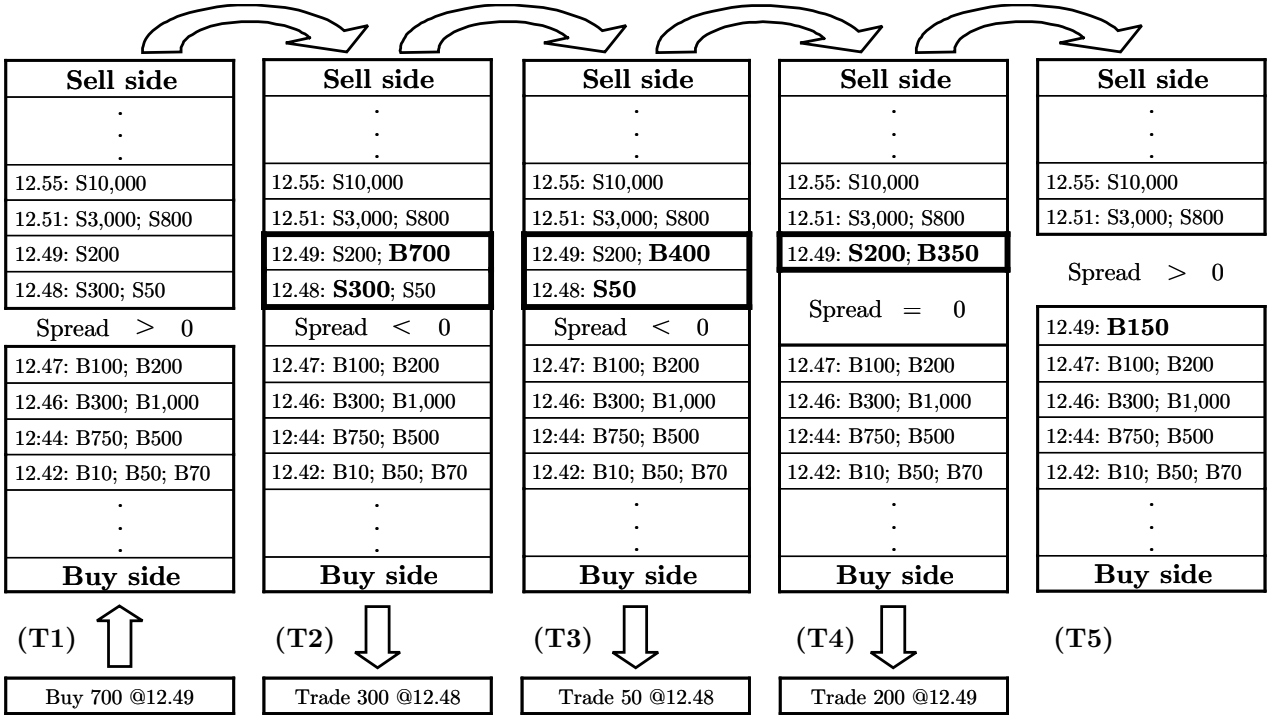

Figure 1: States T1 to T5 of the limit order book for a fictitious stock and currency.

spread triggers the order matching mechanism, which will then execute a trade at a price equal to the limit price of an existing order in the book. The trade is called buyer-initiated if the initiating order was a bid (buy order), and seller-initiated if the initiating order was an ask (sell order). If both the trade initiating order and the existing order have the same size then this will be the actual size of the executed trade, and subsequently both orders will be removed from the limit order book. If the sizes of the two orders are different then the trade size will be equal to the smaller of the two sizes, while the smaller order and the matched part of the larger order will be removed from the limit order book. Any remaining shares of the trade initiating order will be subject to the continued matching process as long as the spread is not positive, potentially generating more trades. Once the spread becomes positive the remaining shares stay in the book as another limit order.

Figure 1 depicts states of the limit order book for a fictitious stock and currency, around an arrival of a new bid (buy order). For each side of the book only orders at several prices closest to the spread have been shown. Each rectangular cell on the buy side and the sell side represents orders at a single price in the order book. The price is shown on the left of a cell, with corresponding orders on its right. Buy limit orders (bids) are marked with B, while sell limit orders (asks) are marked with S, followed by their sizes, respectively. The left-most order at a given price has the time priority. The first state, at time T1, shows the book with a positive spread and a new bid arriving. The bid is for 700 shares with a limit price of 12.49. After it arrived the book changed its state to T2. Because of the way it is priced the new order is a marketable limit order, which means that at least a part of its volume can be matched and traded immediately. A thick frame indicates an area of overlapping bids and asks, with a non-positive spread. At that moment the order matching mechanism matches orders S300 and B700, executing a trade of 300 shares at 12.48. Then order S300 is removed, while order B700 is amended to B400. Subsequently two more trades are executed as illustrated by the transitions of the order book through states T2 to T5. The orders are matched in a sequence determined by the price and time priority rule. All three trades are classified as buyer-initiated. Eventually the positive spread will be restored, while the unmatched 150 shares of the new buy order will stay in the book as a bid at the price of 12.49. Consequently, the prices of the best bid and the best ask in the book will change too.

It should be noted that out of the three trades generated, the first two were executed at a price of 12.48, while the third one at 12.49. The volume (size) weighted average price for the whole sequence is calculated by dividing the total dollar value by the total trade volume (total number of shares) for the three trades, and is equal to 12.4836. Had the new buy order been for no more than 350 shares, which is the total volume of all asks at the price of 12.48, it could have been executed at a single price of 12.48.

The foregoing are the main principles of the limit order book operation. In the second part of the paper we will conduct an exploratory analysis of the trade dataset collected on the Australian Stock Exchange.

3 Hypotheses

In the previous section we described an aggregated trade whose volume weighted average price was worse than the best price before that trade. As can be seen, the limited bid and ask volumes in the order book may introduce additional costs for market participants who want to achieve an immediate trade execution. Traders, in general, try to avoid incurring extra costs, by splitting a large order into a series of smaller ones and submitting them gradually over the trading day. This approach, however, takes time and exposes the traders to potentially adverse price movements. Alternatively they may decide to trade fast, even at an extra cost, if they have to meet a deadline or they feel that a price change is imminent. The extra costs incurred due to either trading too fast or too slowly are recognized as hidden transaction costs.

The above considerations make it clear that the formulation of an optimal order submission strategy is not a trivial task. The problem is the subject of

very intense research, both in the academia and in the finance industry. The literature on practical order submission solutions is rather scarce, possibly due to their potential for commercialization. Most of the papers published deal with the US markets, mainly the New York Stock Exchange (NYSE) and NASDAQ, both of which, unlike the ASX, are not pure limit order markets. The topic of transaction costs and large trades has been studied in the US context by, among others, Holthausen *et al.* (Holthausen, Leftwich & Mayers 1990), Chan and Lakonishok (1995), and Keim and Madhavan (1998). In Australia large trades were analyzed by Aitken and Frino (1996). A measurement methodology for the transaction costs was proposed by Perold (1988). As far as order submission strategies are concerned some analytical solutions are presented in Bertsimas and Lo (1998) and Almgren and Chris (2000).

In this paper we provide evidence of various order submission strategies by analyzing a historical record of executed trades. We will study individual trades and the states of the limit order book around them. Our qualitative approach will employ histograms and unsupervised clustering as the main techniques. We divided the data exploration process into four tasks by formulating the following four hypotheses (labels in brackets):

- (H1) Marketable limit orders do not request more volume than there is available at the best relevant price in the order book.
- (H2) Seller-initiated trades are more frequent than buyer-initiated trades for high values of an order imbalance, defined in the next paragraph, in the order book.
- (H3) Seller-initiated trades are more frequent than buyer-initiated trades when a recent market pressure, defined in the next paragraph, was a selling pressure rather than a buying pressure.
- (H4) The above relationships have stock-specific and time-varying characteristics.

The order imbalance in the limit order book is defined as a ratio of the total volume at the best ask price and the best bid price, respectively. We normalize these volumes as well as the volume ratio via a natural logarithm transformation to reduce the skewness of the data. The market pressure captures a recent history of the trade initiator variable. If more than half of a selected number of recent trades were seller(buyer)-initiated then the market pressure is said to be a selling (buying) pressure.

As far as single trade data are concerned to date researchers have focused mainly on the clustering of trade prices, where some prices occur more frequently than others within a given price range. A number of papers, like for example Harris (1991), studied the US markets. Aitken *et al.* (Aitken, Brown, Buckland, Izan & Walter 1996) analyzed price clustering on the ASX, and reported an increased clustering effect for larger trades. Positive serial correlation in the trade direction was found by Biais *et al.* (Biais, Hillion & Spatt 1995) for the Paris Bourse and Hasbrouck and Ho (1987) on the NYSE. We are not aware of any previous research on cluster analysis of joint trade and limit order book data on a single trade level. Blazejewski *et al.* (Blazejewski, Coggins & Aitken 2003) used the same dataset, described in the next section, to develop predictive models for the direction of the next trade. They used supervised learning techniques, however. The models developed included logistic regression and k-nearest-neighbor, with the latter model achieving the highest forecasting accuracy. Our research hypotheses were formulated on the basis of their work.

4 Data

The dataset consists of 1,059,714 trades, covering a period of 103 trading days, from the 2nd January, 2002, to the 31st May, 2002. It includes information on all trades as well as bids and asks in the limit order book in a time range from 10:15am to 4pm, for the ten stocks with the highest market capitalization on the ASX. The complete record of all trades makes it the highest frequency dataset possible, in temporal sense. For each trade there is a trade direction attribute, which can assume one out of four possible values. We are interested in buyer-initiated and seller-initiated trades only. Before further analysis we performed an aggregation of trades triggered by the same order. As was illustrated in Figure 1, a single marketable limit order may result in a sequence of trades. To obtain a one-to-one mapping between marketable orders and trades we need to add up share volumes of all trades for each trade sequence. The aggregation of the original dataset reduced the number of trades from 1,059,714 to 767,612. We also removed all aggregated trades which took place while the normal order matching mechanism was suspended[2]. After pre-processing our filtered dataset contains 689,076 trades, with 49.51% buyer-initiated trades and 50.49% seller-initiated trades. Further details on the aggregation procedure can be found in Blazejewski *et al.* (2003).

5 Methods

We performed an exploratory, mainly qualitative analysis of the trade dataset. The main techniques used were histograms and unsupervised clustering, with results presented as two dimensional charts. As the unsupervised clustering procedure we employed a self-organizing map. A good description of the SOM algorithm can be found in Kohonen (1998), while its applications in exploratory data analysis are presented in Deboeck and Kohonen (1998) and Kaski (1997). Specifically, we use SOM for data reduction and for projection of a four dimensional dataset onto a two dimensional visualization plane. The SOM algorithm was chosen because of its popularity, robustness, minimal assumptions, and performance in unsupervised density mapping of an input space distribution. The number of nodes for the SOM was selected to be proportional to the square root of the number of trades (Vesanto & Alhoniemi 2000). In our dataset of ten stocks, the minimum number of trades for the whole period for a single stock was 37,679, while the corresponding maximum number was 102,130. Consequently we set the number of nodes to 900, arranged in a square grid of 30x30, with rectangular connections between nodes. We used batch learning with a Gaussian neighborhood function. Euclidean distance served as a measure of similarity, with scaling factors enabling significance tuning (additional normalization) for individual variables. Other data projection and reduction methods, like Generative Topographic Mapping (Bishop, Svensén & Williams 1998) or a combination of multidimensional scaling and k-means, and other distance metrics, like L1 (Manhattan distance) for example, could also be used, but were outside the scope of this paper.

[2]During so called single price auctions, which take place on the ASX before and after the trading hours, orders are collected in the limit order book without being executed. After a specific time the batch execution occurs at a single price, as determined by volumes and prices of orders in the book.

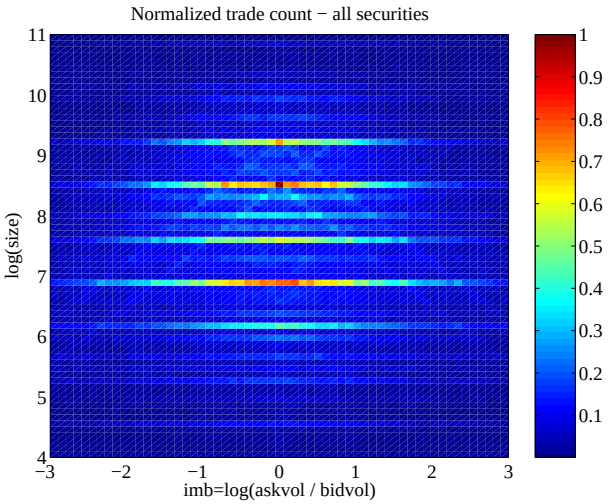

Figure 2: Normalized trade count for all securities for trade size and order book imbalance (60x70 bins).

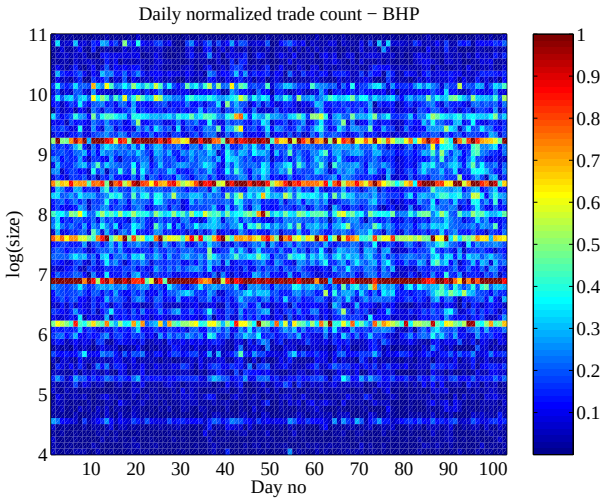

Figure 3: Daily normalized trade count for the BHP stock for trade size and time (103x70 bins).

We chose the following four variables for the input space:

- askvol - total volume at the best ask price in the limit order book just before a trade
- bidvol - total volume at the best bid price in the limit order book just before a trade
- size - size of a trade
- qi_sum - market pressure indicator.

The last variable, qi_sum, is calculated as a sum of the last five trade direction variables, where values of 0 and 1 stand for buyer-initiated and seller-initiated trades, respectively. This is a discrete variable, with six possible values, from 0 to 5. Before employing the clustering procedure we performed data normalization. The raw values of askvol, bidvol, and size variables were transformed via the natural logarithm. This was done to correct for their heavily skewed distributions. To allow for data visualization the output space has two variables, selected depending on the type of chart and marked on a chart's axes. There are two types of charts. The count charts show normalized trade counts, scaled in the range [0, 1]. The trade initiator ratio charts present a ratio of seller-initiated and buyer-initiated trade counts, scaled in the range [-3, 3]. This last range was empirically found to provide the best contrast in the visualization of the trade initiator ratio. The more seller-initiated trades there are, relative to the buyer-initiated ones, the higher the ratio and the closer to red the presentation color is. Both types of charts use color to indicate counts and ratios, respectively. To show the results of the self-organizing map clustering on the charts we used a two dimensional projection of the codebook vectors.

The analysis was performed on a stock by stock basis. This approach was dictated by the fact that even though there can be some correlations between price returns of different stocks, the stocks are traded through their own separate limit order books and they trade at different prices. The last point means that the same dollar value of shares would be represented by different volumes of shares for different stocks. As a result we would observe different trade sizes and different volumes in the order book. Figure 2 shows trade counts for all ten stocks in the dataset, with trade size and order imbalance on the axes. It can be seen that trade sizes form five major and several minor clusters (horizontal lines). When we compare this against Figure 3, which shows daily trade counts for the BHP stock, we can see that some of the BHP clusters are much more pronounced. Also, the trade counts of clusters seem to vary on a daily basis. We find that the data, within the limits of our exploration, indirectly support hypothesis H4. On the other hand, regardless of the trade count values, the existence of similar major clusters for all stocks and for the BHP stock in a size (log) range between 6 and 9.3 seems to indicate a behavioral regularity, with corresponding trade sizes of 500, 1000, 2000, 5000, and 10,000 shares. To properly account for the time-varying aspects of the problem we would need to apply a local modelling approach, which is beyond the scope of this paper. We computed results for each of the ten stocks separately, for the whole period in the dataset. We found that, despite individual characteristics, the BHP stock is representative of the majority (but not all) of our stocks on a qualitative level. In the subsequent section therefore, unless indicated otherwise, we present results for the BHP stock only.

The data processing and modelling software was implemented using SMARTS suite of applications, Matlab^{TM} computing platform, and the SOM toolbox developed at the Helsinki University of Technology, Finland.

6 Results

To address the first hypothesis (H1) we calculated a percentage of trades, relative to the total number of trades, which did not trade more volume than was available at the best relevant price in the limit order book. Table 1 presents results for the BHP stock and for the whole dataset of ten stocks. As can be seen trades which did not meet the condition represented less than 2% of all trades for the whole dataset, and less than 1% for the BHP stock. We find that the data support hypothesis H1.

Code	Si%	Bi%	Total%
BHP	46.60	52.70	99.30
All	49.64	48.75	98.39

Table 1: Percentage of trades with the executed volume no greater than the volume at the best price in the limit order book (Si-Seller-initiated, Bi-Buyer-initiated).

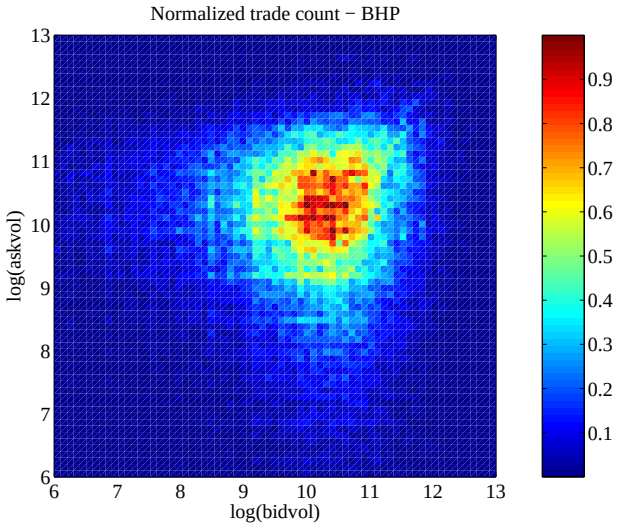

Figure 4: Normalized trade count for the BHP stock for askvol and bidvol (70x70 bins).

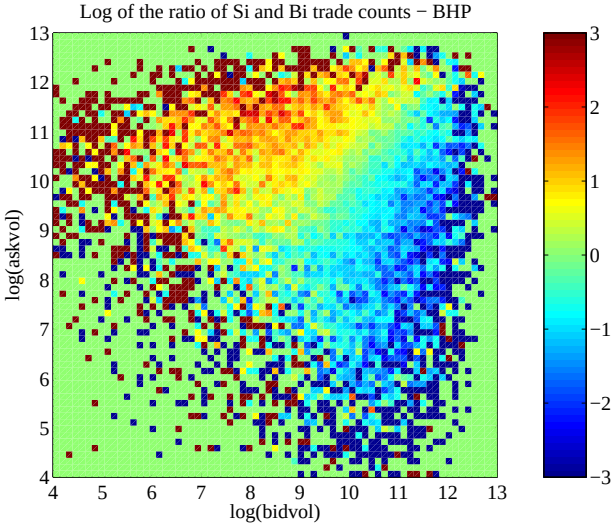

Figure 5: Trade initiator ratio for the BHP stock for askvol and bidvol (70x70 bins).

We are now going to analyze a number of charts produced to obtain a qualitative answer to the question posed by hypothesis H2. Figure 4 depicts a trade count for variables askvol and bidvol. It seems to have a single core with a center in the point (10, 10), and two weak (low count) projections to the sides. Looking at Figure 5 we can see that it is these two weak projections in Figure 4 which correspond to buyer-initiated (green to blue) and seller-initiated trades (green to red), located symmetrically around the line for which askvol=bidvol. The seller-initiated trades, however, are above this line, while the buyer-initiated ones are below it. The separating line has a green color, which stands for a (log) ratio of 0. It seems to represent an equilibrium, where equal volumes at the best ask and the best bid coincide with equal trade counts of each initiator type, meaning an equal probability of observing seller-initiated and buyer-initiated trades. However, there is also an elongated cloud of data between points (10, 6) and (6, 10), which is perpendicular to the askvol=bidvol line. This formation is a mixture of buyer-initiated and seller-initiated trades and could possibly be separated by accounting for the trade size. The foregoing provides support for hypothesis H2.

To include additional variables in our analysis we applied the SOM algorithm to the four dimensional input space, with askvol, bidvol, size, and qi_sum as input variables. The codebook of the SOM was subsequently projected back onto a plane in the input space, by selecting only two of its dimensions. The results are presented in Figures 6 and 7. The trade count chart in Figure 6 has the appearance of a skeleton version of the chart in Figure 4. The area with the highest values, the core, is very small, while the sideways projections manage to capture the direction and length of their counterparts in Figure 4. We experimented with various scaling factors for the input variables to assess how a change in the distance measure affects sizes and the separation of clusters on the chart. We obtained the best results with scaling factors equal to one, that is, no scaling. The trade initiator ratio chart in Figure 7 has one sideways projection assigned to the buyer-initiated trades (green to blue), and the other one to the seller-initiated trades (green to red). We note that there are a few blue points in the red area, but in general the separation between the buyer-initiated and seller-initiated trades seems to be very good.

The third hypothesis, H3, can be addressed through an appropriate frequency table. Table 2 shows respective percentages of seller-initiated and buyer-initiated trades associated with the six possible values of the qi_sum variable.

qi_sum	Si%	Bi%	Log Ratio
0	5.34	9.65	-0.257
1	16.51	21.88	-0.122
2	26.62	28.13	-0.024
3	27.91	24.25	0.061
4	17.73	12.73	0.144
5	5.89	3.36	0.244

Table 2: Percentage of the seller-initiated and buyer-initiated trades for the BHP stock after qi_sum seller-initiated trades in the last five trades (Si-Seller-initiated, Bi-Buyer-initiated).

The extreme values of 0 and 5 for qi_sum correspond to an extreme imbalance between the numbers of seller-initiated and buyer-initiated trades, with a strong positive autocorrelation between the majority direction of the last five trades and the direction of the next trade. Log ratio in the table was calculated as a natural logarithm of the ratio of the percentage of the seller-initiated and buyer-initiated trades, respectively. Although the corresponding data for the other securities vary, they also show a positive autocorrelation in the trade direction variable. We conclude that the data support hypothesis H3.

7 Conclusions

We conducted an exploratory analysis of the trade level data for the Australian Stock Exchange. We formulated and then qualitatively confirmed four hypotheses. The main contribution of this paper was an application of the self-organizing map to the unsupervised clustering of trades into buyer-initiated and seller-initiated groups. We used histograms and the SOM algorithm to visualize the trade data in two dimensions. We note that the SOM transformation emphasizes regularly occurring trading patterns conditioned on all dimensions of the trade data, while the simple histograms fail to separate rare and common trading behaviors. The visualization of the data using the SOM transformation reveals that buyer-initiated and seller-initiated trades form two distinct clusters in correspondence with non-equilibrium market conditions and elicits the main structural features of the

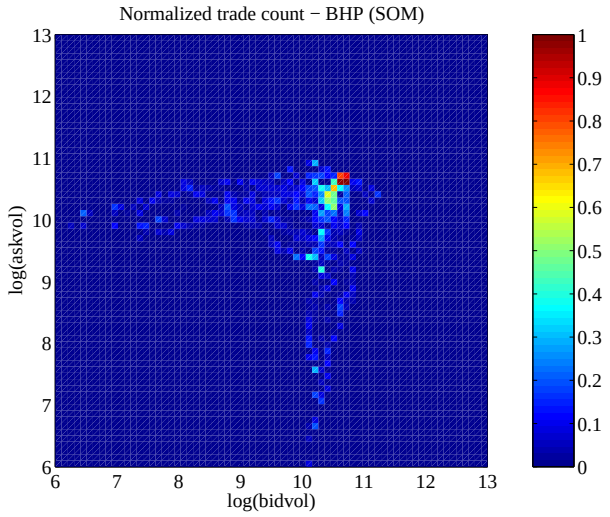

Figure 6: Normalized trade count for the BHP stock for askvol, bidvol, size, and qi_sum, after the SOM transformation (70x70 bins).

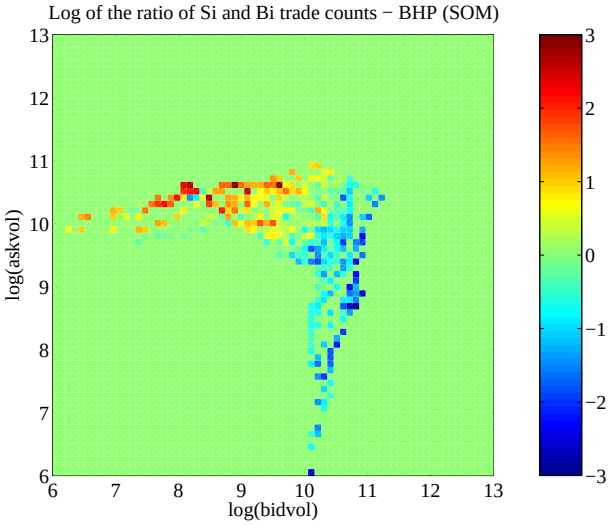

Figure 7: Trade initiator ratio for the BHP stock for askvol, bidvol, size, and qi_sum, after the SOM transformation (70x70 bins).

clusters. In future work we would like to include more variables in our trade direction model, in particular more current and lagged volume and price information from the limit order book.

8 Acknowledgements

The database from which the dataset was extracted as well as software used for extraction were provided by Capital Markets Cooperative Research Centre (CMCRC) and its industry partners, which we gratefully acknowledge. Mr Adam Blazejewski was supported by an Australian Postgraduate Award scholarship and a CMCRC scholarship.

References

Aitken, M., Brown, P., Buckland, C., Izan, H. Y. & Walter, T. (1996), 'Price clustering on the Australian Stock Exchange', *Pacific-Basin Finance Journal* **4**(2-3), 297–314.

Aitken, M. & Frino, A. (1996), 'Execution costs associated with institutional trades on the Australian Stock Exchange', *Pacific-Basin Finance Journal* **4**(1), 45–58.

Aitken, M., Frino, A., Jarnecic, E., McCorry, M., Segara, R. & Winn, R. (1997), *The Microstructure of the Australian Stock Exchange: An Introduction*, SIRCA Occassional Series, 1996 Asia-Pacific Capital Markets Foundation.

Almgren, R. F. & Chriss, N. (2000), 'Optimal execution of portfolio transactions', *Journal of Risk* **3**, 5–39.

Bertsimas, D. & Lo, A. W. (1998), 'Optimal control of execution costs', *Journal of Financial Markets* **1**(1), 1–50.

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**(5), 1655–1689.

Bishop, C. M., Svensén, M. & Williams, C. K. I. (1998), 'GTM: The Generative Topographic Mapping', *Neural Computation* **10**(1), 215–234.

Blazejewski, A., Coggins, R. & Aitken, M. (2003), 'Dynamic non-parametric model for trade direction forecasting', accepted for presentation at the 16th Australasian Finance & Banking Conference, Sydney, Australia.

Chan, L. K. C. & Lakonishok, J. (1995), 'The behavior of stock prices around institutional trades', *Journal of Finance* **50**(4), 1147–1174.

Deboeck, G. & Kohonen, T., eds (1998), *Visual Explorations in Finance with Self-Organizing Maps*, Springer Verlag.

Harris, L. (1991), 'Stock price clustering and discreteness', *Review of Financial Studies* **4**(3), 389–415.

Hasbrouck, J. & Ho, T. S. Y. (1987), 'Order arrival, quote behaviour, and the return-generating process', *Journal of Finance* **42**(4), 1035–1048.

Holthausen, R. W., Leftwich, R. W. & Mayers, D. (1990), 'Large-block transactions, the speed of response, and temporary and permanent stock-price effects', *Journal of Financial Economics* **26**, 71–95.

Kaski, S. (1997), Data exploration using self-organizing maps, Ph.D., Helsinki University of Technology, Finland.

Keim, D. B. & Madhavan, A. (1998), 'The cost of institutional equity trades', *Financial Analysts Journal* **54**(4), 50–69.

Kohonen, T. (1998), 'The self-organizing map', *Neurocomputing* **21**, 1–6.

Perold, A. F. (1988), 'The implementation shortfall: paper versus reality', *Journal of Portfolio Management* **14**, 4–9.

Vesanto, J. & Alhoniemi, E. (2000), 'Clustering of the self-organizing map', *IEEE Transactions on Neural Networks* **11**(3), 586–600.