A local non-parametric model for trade sign inference ⋆

Adam Blazejewski [a,*], Richard Coggins [b]

[a] *School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia*

[b] *Discipline of Finance, School of Business, University of Sydney, NSW 2006, Australia*

Abstract

We investigate a regularity in market order submission strategies for twelve stocks with large market capitalization on the Australian Stock Exchange. The regularity is evidenced by a predictable relationship between the trade sign (trade initiator), size of the trade, and the contents of the limit order book before the trade. We demonstrate this predictability by developing an empirical inference model to classify trades into buyer-initiated and seller-initiated. The model employs a local non-parametric method, k-nearest-neighbor, which in the past was used successfully for chaotic time series prediction. The k-nearest-neighbor with three predictor variables achieves an average out-of-sample classification accuracy of 71.40%, compared to 63.32% for the linear logistic regression with seven predictor variables. The result suggests that a non-linear approach may produce a more parsimonious trade sign inference model with a higher out-of-sample classification accuracy. Furthermore, for most of our stocks the observed regularity in market order submissions seems to have a memory of at least 30 trading days.

Key words: Order submission, Trade classification, K-nearest-neighbor, Non-linear, Memory
PACS: 89.65.Gh, 05.45.Tp, 05.45.-a

⋆ This is an extended and revised version of a paper presented at the 16th Australasian Finance & Banking Conference (Sydney, Australia, December 2003).

* Corresponding author. Tel.: +61-2-9351-3229; fax: +61-2-9351-3847.
E-mail address: adamb@sedal.usyd.edu.au (A. Blazejewski).

1 INTRODUCTION

Most trades which take place on stock exchanges with a continuous double auction and a limit order book can be divided into two categories, buyer-initiated and seller-initiated. A buyer-initiated trade (a buy) is a trade triggered by a buy market order matched against one or more sell limit orders in the order book. The opposite holds for a seller-initiated trade (a sell), where a sell market order is matched against one or more buy limit orders in the order book. Submitters of market orders are called liquidity demanders, while submitters of limit orders stored in the book are called liquidity providers. Trades for which a buyer and a seller are the same entity (crosses), as well as trades executed during single price auctions (exchange-initiated) represent a small fraction of all trades and are excluded from our analysis. The side of the market which submits a market order and thereby initiates a trade is called a trade initiator. The trade initiator is the same as the initiator of the market order which triggered the trade, which allows to use the former to determine the latter [1]. The trade initiator can be treated as another trade attribute in the form of a binary variable, beside price, size, and others. The trade initiator variable is alternatively referred to as a trade sign, trade direction, trade indicator, or buy/sell indicator. We will use the second term, *trade sign*, throughout the rest of this paper.

To date the trade sign variable has been employed in such areas of market microstructure research as price formation, order and trade imbalance, order flow and order submission strategies, market impact, and trade classification. Our study primarily belongs to the research on order submission strategies and trade classification. Refs. [1–9] represent a small sample of studies on order flow and order submission strategies in various markets with a limit order book. Ref. [1] finds a positive autocorrelation in the order flow in the Paris Bourse, where a probability of a buy(sell) market order conditional on the previous buy(sell) market order is greater than an unconditional probability of such an order. Ref. [2] reports the same regularity for the Tokyo Stock Exchange. The observed positive autocorrelation in the order flow is considered to be caused by breaking up of large orders, momentum trading, and similar reactions to news releases [1]. The autocorrelation of order signs is also claimed to be a long-memory process, by Ref. [10] for the Paris Bourse, and by Ref. [11] for the London Stock Exchange. As far as order submission strategies are concerned, traders are found to monitor the state of the order book and choose their actions accordingly [3, 4, 7, 8, 12]. In particular, there is some evidence that large volume on the same side of the book makes submissions of market orders more frequent [7]. On the other hand, some authors speculate that large

[1] A trade aggregation procedure, described later in the text, is necessary for mapping trades into market orders.

volume on the opposite side may encourage the submission of a large market order [8].

There exist a number of studies on trade classification [13–16]. The proposed methods, however, have been primarily designed for quote-driven markets. They aim to recover the trade sign with as high an accuracy as possible, relying on quote and trade prices. Perhaps the closest to our work are studies by Porter [17] and Aitken et al. [18]. Porter [17] uses logistic regression to classify trades and finds systematic temporal patterns in interday and intraday probabilities of trading at the asking price on the US and Canadian exchanges. Aitken et al. [18] analyze the intraday probability of trading at the asking price on the Australian Stock Exchange. They use limit order book and other data to build a logistic regression model for a set of over 3 million trades, and manage to correctly classify 53.3% of trades, while 51.58% of all trades in their data set are at the asking price.

Our study explores a regularity in market order submission strategies on the Australian Stock Exchange (ASX). The regularity is evidenced by a predictable relationship between the trade sign, size of the trade, and the contents of the limit order book before the trade. We demonstrate this predictability by developing an empirical trade sign inference model to classify trades into buyer-initiated and seller-initiated. The model is based on a local non-parametric method, k-nearest-neighbor (k-NN). This method has been successfully applied by other researchers to forecasting chaotic time series [19–21], as well as various financial time series, like for example currency exchange rates [22] and a stock index [23]. We use transaction level data for twelve large stocks on the ASX. The trade sign classification is conditional on contemporaneous and past volumes in the order book, trade sizes, and past trade signs. Quote and trade prices are not used. Classification accuracy is determined through out-of-sample testing. The classification performance of the k-NN classifier is compared against the performance of three other classifiers: linear logistic regression, trade continuation, and majority vote. We show that the k-NN classifier is superior to the other classifiers and can separate buyer-initiated and seller-initiated trades in our data set with an average accuracy of over 71%. Furthermore, for most of our stocks the observed regularity in market order submissions seems to have a memory of at least 30 trading days.

2 DATA SET

The Australian Stock Exchange is an order driven market with an electronic limit order book and without designated market makers. It operates a continuous double auction throughout the day, Monday to Friday, from around 10

a.m. [2] to 4 p.m., except for single price auctions at market opening, after market closing, and occasionally during the day, after a trading halt. Our data set consists of tick-level information on all trades, orders and full limit order book contents for twelve stocks with large market capitalization [3] on the ASX, for the period from 11 November 2002 to 27 August 2003, comprising 199 trading days. During the investigated period the selected stocks belonged to the top 30 stocks ranked by market capitalization (8 stocks were in the top 10), were actively traded on each day, and did not undergo any major price revisions or splits.

The data were collected by the stock exchange and represent the highest resolution and most complete transaction data set available. In particular, every transaction (trade or order) has an exact time stamp, accurate to 1 second, while transactions occurring within the same second are arranged in the original sequential order. Each trade, apart from price and size information, has a trade sign attribute and an on/off market flag. The trade sign attribute can assume one of four values: buyer-initiated, seller-initiated, crossing, and exchange-initiated. The on/off market flag indicates whether a trade was transacted through the limit order book (on-market) or outside of it (off-market). We analyze buyer-initiated and seller-initiated on-market trades only. Crossing trades and exchange-initiated trades were excluded. As far as the limit order book is concerned, the data set includes complete price and size information for each bid and ask in the book throughout a trading day. There are 2,355,334 trades in the whole data set. A subset with buyer-initiated and seller-initiated trades represents 92.73% of the data set and contains 2,184,046 trades, with 50.44% of them being buyer-initiated.

3 METHODS

We develop an empirical inference model of the trade sign variable for a single trade. The first 30 days in the data set are used to select the best predictor variable sets out of a collection of 71 sets. Variable sets are ranked by classification accuracy across all stocks, and the best sets are selected for the logistic regression and the k-nearest-neighbor. The remaining 169 days in the data set serve as a test set to evaluate the classification accuracy of the models with the best predictor variable sets. Two simple classifiers, a trade continuation

[2] Groups of stocks take part in the morning single price auction between 9:59:45 a.m. and 10:15:00 a.m., in an alphabetic order by stock name, with groups earlier in the alphabet opening before groups later in the alphabet.

[3] The codes of the stocks, ordered by market capitalization in a decreasing order, are as follows: NAB, BHP, CBA, ANZ, WBC, NCP, RIO, WOW, FGL, SUN, SGB, MIG.

and a majority vote, based on lagged values of the trade sign only, are used for performance comparison. The models are estimated and tested with a moving window method.

Before the analysis we construct a market order sign proxy by aggregating together trades resulting from the same order. We apply two simple rules to aggregate trade sequences. Firstly, a change of the spread in the limit order book from positive to non-positive signals a beginning of a new trade sequence. Secondly, the time when the spread becomes positive again marks the end of that trade sequence. The trade sequence found is then aggregated into a single trade, with its size being equal to the sum of all constituent trade sizes. The process is repeated for all trades in the data set. This approach will work even during periods of concentrated trading, where there are orders and trades with the same time stamp (accurate to 1 second) and the duration between transactions seems to be zero, because no new or amended orders will be accepted until the market is cleared. The process of aggregation reduced the total number of buyer-initiated and seller-initiated trades in the data set to 1,542,205, out of which 51.78% are buyer-initiated.

The complete history of trade and order flow could potentially allow an exhaustive search approach to find the best predictor variables for the specific target variable. By the exhaustive search we mean estimating models and testing their classification accuracy for all possible combinations of predictor variables. Unfortunately, the large amount of data and dimensions in our data set would make this approach prohibitively expensive (in terms of computational time). On the other hand, variable selection methods, like for example those based on the Akaike information criterion [24], are not applicable for the k-NN classifier. The k-nearest-neighbor belongs to a class of memory-based classifiers and requires out-of-sample testing to assess its generalization performance. Our variable selection procedure is a constrained version of the exhaustive search. The number of possible predictor variable combinations, further referred to as variable sets, was restricted to 71 by introducing a set of candidate variables and a set of rules for combining these variables into variable sets. The set of candidate predictor variables, $\mathbf{V}$, consists of the following variables:

- $a^{\mathrm{p}}_{\mathrm{n-g}}$ - lag g of the total volume in the limit order book at the ask price level p, captured just before an order which triggered the $(n-g)$-th trade; $g \in \mathbb{Z}$, $g = 0 \ldots 3$; $p \in \mathbb{Z}$, $p = 1 \ldots 3$; n indexes over the aggregated daily trades.
- $b^{\mathrm{p}}_{\mathrm{n-g}}$ - lag g of the total volume in the limit order book at the bid price level p, captured just before an order which triggered the $(n-g)$-th trade; $g \in \mathbb{Z}$, $g = 0 \ldots 3$; $p \in \mathbb{Z}$, $p = 1 \ldots 3$.
- $s_{\mathrm{n-g}}$ - lag g of the trade size; $g \in \mathbb{Z}$, $g = 0 \ldots 5$.

The symbol $\mathbb{Z}$ denotes the set of integers. The lagged trade sign is denoted as $\epsilon_{\mathrm{n-g}}$ and does not belong to $\mathbf{V}$. The sign of the current trade, ϵ_{n} (lag 0), is the target variable for the inference. Throughout the rest of the paper the index n of the current trade will be omitted, simplifying variable symbols to $a_{\mathrm{g}}^{\mathrm{p}}$, $b_{\mathrm{g}}^{\mathrm{p}}$, s_{g}, and ϵ_{g}, respectively. Consequently, ϵ_0 will denote the target variable. The trade sign variable ϵ_{g} can assume two values only, $+1$ for buyer-initiated trades, and -1 for seller-initiated trades. Total volumes and the trade size are measured in units of shares. The first ask(bid) price level ($p = 1$) corresponds to the price of the best ask(bid) in the limit order book, while subsequent price levels correspond to prices $p - 1$ price ticks above(below) the first ask(bid) price. For all the stocks in the data set one price tick is equal to one cent [4]. The largest value of g was set to 5 (for the trade size variable). Consequently, the first five trades on each day are used as lagged trades only. To further reduce the number of predictor variables and their combinations an additional set of constraints, $\mathbf{C}$, was imposed. It specifies rules which must be satisfied by any variable set $\mathbf{X}$, where $\mathbf{X} \subseteq \mathbf{V}$:

(1) Number of elements: $\#\mathbf{X} = n_x$, and $n_x = 2 \dots 7$.
(2) Bid-Ask symmetry: if $a_{\mathrm{g}}^{\mathrm{p}} \in \mathbf{X}$ then $b_{\mathrm{g}}^{\mathrm{p}} \in \mathbf{X}$.
(3) Mandatory variables: $\{a_0^1, b_0^1\} \subseteq \mathbf{X}$ or $\{a_1^1, b_1^1\} \subseteq \mathbf{X}$.
(4) Price priority: if $x^{\mathrm{p}} \in \mathbf{X}$ then $\forall i \in \mathbb{Z}$: $x^{\mathrm{p-i}} \in \mathbf{X}, p - i \geq 1$.
(5) Lag priority: if $x_{\mathrm{g}} \in \mathbf{X}$ then $\forall i \in \mathbb{Z}$: $x_{\mathrm{g-i}} \in \mathbf{X}, g - i \geq 1$.

The unary operator "#" determines the number of elements in a set. The maximum number of variables in a set was limited to seven due to our preference for parsimonious models and a need for a sufficient ratio of cases (trades) to predictor variables [25]. The introduction of the two sets, $\mathbf{V}$ and $\mathbf{C}$, reduced the total number of predictor variable sets to 71. Before model estimation all predictor variables are pre-processed by calculating their natural logarithms. After this transformation lagged values of the trade size s_{g} are signed with the corresponding values of the lagged trade sign ϵ_{g} ($g > 0$). The contemporaneous value of the trade size s_0 is not signed because the contemporaneous trade sign ϵ_0 is the target variable. The signing procedure incorporates the trade sign into the trade size. This avoids a potential problem of how to include binary variables in the distance metric of the k-NN classifier, if lagged trade signs were included as predictor variables.

An instance of a set of values for a given variable set, including the current trade sign, corresponds to a single trade and is called a data point. The terms data point and trade will be used synonymously, but for clarity one may sometimes be preferred over the other. The only target variable is the sign of the current trade, ϵ_0. The process of model estimation (training) and testing is iterative, and employs a moving window method. The models are estimated

[4] Prices on the ASX are quoted in the Australian dollars and cents.

using a given training interval $\mathbf{T}$, which consists of all trades on N_{t} days in $\mathbf{T}$. The models estimated on a given training interval are tested on a test interval $\mathbf{S}$ comprising all trades on a single day (test day) immediately after the training interval $\mathbf{T}$. Selection of more recent data for the test interval than for the training interval is dictated by the time series nature of the data [26]. The result of the testing procedure is a single classification accuracy value for the test day. The estimation and testing are then repeated for a new pair of training and test intervals, obtained by shifting the previous pair of intervals one day forward. The process continues iteratively for N_{s} test days, producing a set of N_{s} daily classification accuracy values for each model.

The models are built for the four classifiers: logistic regression, k-nearest-neighbor (k-NN), trade continuation, and majority vote. The logistic regression classifier is constructed as follows:

$$\begin{aligned} \epsilon_0 &= \begin{cases} -1 & \text{if } \gamma \leq 0 \\ +1 & \text{if } \gamma > 0 \end{cases} \\ \gamma &= ln\left(\frac{P(\epsilon_0)}{1-P(\epsilon_0)}\right) \\ \gamma = f(\mathbf{x}), \quad f(\mathbf{x}) &= A\mathbf{x} + c, \quad \mathbf{x} = (x_1, \ldots, x_{\mathrm{n_x}}) \end{aligned} \tag{1}$$

Function $f(\mathbf{x})$ is a linear regression with n_x predictor variables x_{i}, where $x_{\mathrm{i}} \in \mathbf{X}$. The value γ of the logit function is calculated as a natural logarithm of the ratio of the class membership probabilities $P(\epsilon_0)$ and $(1 - P(\epsilon_0))$ [25]. Input data $\mathbf{x}$ are assigned to the buyer-initiated class ($\epsilon_0 = +1$) if their corresponding logit value γ is above 0 , and to seller-initiated class ($\epsilon_0 = -1$) otherwise.

The k-nearest-neighbor classifier belongs to a class of non-parametric, memory-based classifiers. During the training phase a set $\mathbf{D}_{\mathrm{t}}$ of all data points in a given training interval $\mathbf{T}$ is stored in the classifier's memory. Testing is conducted for a set $\mathbf{D}_{\mathrm{s}}$ of all data points in a test interval $\mathbf{S}$. During an evaluation of a test data point $\mathbf{d}_{\mathrm{s}}$ the classifier computes squared Euclidean distances between $\mathbf{d}_{\mathrm{s}}$ and all the data points $\mathbf{D}_{\mathrm{t}}$ in its memory. Calculation of the Euclidean distance involves all n_x dimensions (predictor variables) in a given set $\mathbf{X}$. Subsequently a set $\mathbf{K}$ of k data points from the classifier's memory with the shortest distances to $\mathbf{d}_{\mathrm{s}}$ is selected. A trade sign ϵ_0 for the test data point $\mathbf{d}_{\mathrm{s}}$ is inferred as equal to the trade sign of the majority of the data points in $\mathbf{K}$, as long as k is an odd positive integer. In our experiment three values of k are used: 1, 5, and 9. For each value of k a separate k-NN model is estimated and tested. The following is a more formal description of the k-NN classifier:

$$
\begin{aligned}
\epsilon_0 &= \begin{cases} -1 & \text{if } \beta_\epsilon \leq 0 \\ +1 & \text{if } \beta_\epsilon > 0 \end{cases} \\
\beta_\epsilon &= \sum_{\mathbf{d}_\mathrm{i} \in \mathbf{K}} \epsilon_0 \\
\forall(\mathbf{d}_\mathrm{i} \in \mathbf{K})\ \forall(\mathbf{d}_\mathrm{j} \in \mathbf{D}'_\mathrm{t})\ &: \|\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{s}\| \leq \|\mathbf{x}_\mathrm{j}, \mathbf{x}_\mathrm{s}\|,\ \mathbf{d}_\mathrm{s} \in \mathbf{D}_\mathrm{s} \\
\mathbf{D}'_\mathrm{t} &= \mathbf{D}_\mathrm{t} \setminus \mathbf{K},\ \#\mathbf{K} = k,\ \mathbf{d} = (x_1, \ldots, x_{\mathrm{n_x}}, \epsilon_0) \\
\|\mathbf{x}, \mathbf{x}'\| &= (\sum_{\mathrm{j}=1}^{\mathrm{n_x}} (x_\mathrm{j} - x'_\mathrm{j})^2)^{\frac{1}{2}}
\end{aligned}
\tag{2}
$$

The brackets "$\|\cdot\ ,\ \cdot\|$" denote an Euclidean distance operator, while the binary operator "$\setminus$" calculates a set difference. A comprehensive treatment of the k-nearest-neighbor classifier can be found in Ref. [27].

The trade continuation classifier exploits the observed autocorrelation in the market order sign. It assumes that the sign of the current trade will be the same as the sign of the previous trade:

$$
\epsilon_0 = \epsilon_1 \tag{3}
$$

The majority vote classifier does not use any information from the test interval. It detects an imbalance between buyer-initiated and seller-initiated trades in the training interval. The classifier determines the sign of the majority and then assigns it to all trades in the test interval:

$$
\begin{aligned}
\epsilon_0 &= \begin{cases} -1 & \text{if } \beta_\epsilon \leq 0 \\ +1 & \text{if } \beta_\epsilon > 0 \end{cases} \\
\beta_\epsilon &= \sum_{\mathbf{d}_\mathrm{i} \in \mathbf{D}_\mathrm{t}} \epsilon_0
\end{aligned}
\tag{4}
$$

The last two classifiers, the trade continuation and the majority vote, do not use any predictor variables from the list $\mathbf{V}$, and consequently their performance does not depend on a choice of $\mathbf{X}$.

As mentioned earlier, the length of a training interval is N_t days, while there are N_s one day test intervals. The choice of N_t is not obvious and may depend on a particular stock. We decided to try various values between 1 and 30 days, starting from one day, and then every even number of days until 30 days. Each of the selected 16 values of N_t defines a separate training timescale. One of the classifiers, the trade continuation, does not depend on a training timescale. We initially intended to estimate and test the four classifiers [5] described above on the whole data set of 199 trading days. The k-nearest-neighbor classifier has a

[5] Separate k-NN models are built for each of the three values of k.

high computational cost and was expected to consume most of the computer time. Preliminary computations revealed that building models for the four classifiers, 16 training timescales, and 71 variable sets, over the whole period in the data set, would take several months on a four node computer cluster [6]. More importantly, however, it became apparent that the size of the data set would be too small, relatively to the total number of 4,561 potential models for each stock [7], to find statistically significant differences in the classifiers' performance. These two factors, the computational cost and the statistical significance of multiple comparisons, led us to impose additional constraints on the experiment.

We divided the whole data set of 199 trading days into two parts. The first 30 days in the data set will be used to select a subset from the collection of 71 predictor variable sets. The subset will include four variable sets, the best two for the logistic regression, and the best two for the k-nearest-neighbor. Out of the two sets for each classifier one set will allow contemporaneous variables, while the other one will not. The trade continuation and the majority vote classifiers will not be affected by these decisions because they do not depend on any of the predictor variables on the list $\mathbf{V}$. The training interval length will be set to one value only, 20 days. This particular value was selected because we have a preference for predictor variable sets performing well on longer timescales, between 10 and 30 days. There will be 10 validation days which will provide 10 classification accuracy values for each stock. To mitigate a potential problem of overfitting [8], due to the small number of validation days, results for all stocks will be pooled together. The variable sets with the highest mean classification accuracy, across individual mean accuracies for the twelve stocks, will be selected as the best predictor variable sets.

The remaining 169 days in the data set, starting from day 31, will serve as a test set to evaluate daily performance of the classifiers with the selected four predictor variable sets. A separate classification accuracy will be determined for each of the 16 training timescales. During this phase the first 30 days in the data set will be re-used to construct training intervals for model estimation, but they will never be used for testing of classification accuracy. The last day of a first training interval, regardless of its length, will always fall on day 30 in the data set. This will ensure that models are estimated with the most recent data available and that they are all tested on the same set of 169 days.

---

[6] Each node was approximately twice as fast as the Intel® Celeron® 2.00 GHz processor with 256 MB of memory.

[7] The total of $4,561$ models was calculated in the following way: $3 * 16 * 71$ k-NN, $16 * 71$ logistic regression, 1 trade continuation, and 16 majority vote models.

[8] Overfitting can occur when multiple tests are performed on a relatively small data set. There are a number of remedies for this problem, including cross-validation, bootstrap, or increasing the size of the data set.

Table 1
Classification accuracy (%) for the best predictor variable sets across the twelve stocks - 10 validation days. Abbreviated headings: CV - contemporaneous variables; SetNo - variable set number; SD - standard deviation.

CV	SetNo	Predictor variables	Mean	SD	Min	Max
Logistic regression:						
Yes	53	a_0^1, a_1^1, b_0^1, b_1^1, s_1, s_2, s_3	60.95	3.28	55.70	66.23
No	45	a_1^1, b_1^1, s_1, s_2, s_3, s_4	55.78	2.82	51.32	60.05
k-NN ($k = 1$):						
Yes	3	a_0^1, b_0^1, s_0	69.55	2.89	64.97	73.49
No	6	a_1^1, b_1^1, s_1	54.78	2.43	50.83	58.68
k-NN ($k = 5$):						
Yes	3	a_0^1, b_0^1, s_0	71.63	2.99	66.61	76.23
No	6	a_1^1, b_1^1, s_1	57.10	2.71	52.82	61.53
k-NN ($k = 9$):						
Yes	3	a_0^1, b_0^1, s_0	71.75	3.00	66.99	76.38
No	6	a_1^1, b_1^1, s_1	58.20	2.79	53.90	62.61

The software for the experiment was implemented using the SMARTS® trading and surveillance system, and Matlab® with the NETLAB toolbox [28].

4 RESULTS

The test statistics for the best predictor variable sets within the collection of 71 sets, calculated using the first 30 days in the data set, are presented in Table 1. The statistics represent mean values across individual statistics for the twelve stocks. The means of the four individual statistics are reported: mean, standard deviation, minimum, maximum. The individual statistics were calculated across the 10 validation days. The sets with the highest mean classification accuracy were selected as the best predictor variable sets. Statistical significance was not determined because it is not used by our selection procedure. This approach was adopted to avoid an inconclusive result, if differences between sets were not statistically significant. As can be seen for all classifiers, the sets with contemporaneous variables have higher mean accuracies than the sets without them. The difference is around 5% for the logistic regression, and between 13% and 15% for the k-nearest-neighbor. The most interesting, however, is the difference between the k-NN classifier and the logistic regression, for the sets with contemporaneous variables. Depending on the value of k, the k-nearest-neighbor has the mean accuracy approximately 9% to 11% higher than the logistic regression, while its standard deviation varies from 2.89% to

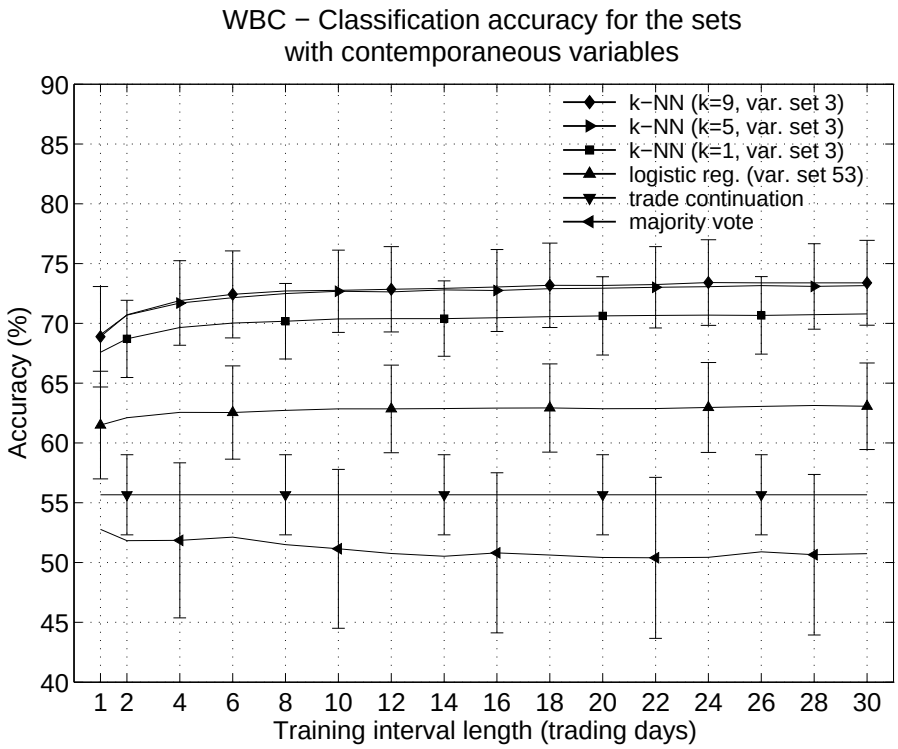

Fig. 1. Classifier performance vs. training interval length for the WBC stock, for the sets with contemporaneous variables - 169 test days. The ranges mark one standard deviation below and above respective test means.

3.28%. Furthermore, the higher the value of k the better the performance of the k-NN classifier, even though an improvement between $k = 5$ and $k = 9$ is minimal. The results for the sets without contemporaneous variables show a difference of only 2.42% between the best k-nearest-neighbor ($k = 9$) and the logistic regression. Another aspect worth pointing out is the number of predictor variables in the best variable sets. In the case of the k-NN classifier the sets have only three variables, and they are identical for all three values of k within a group with contemporaneous variables (set 3), and within a group without contemporaneous variables (set 6), respectively. On the other hand, in the case of the logistic regression, the best variable set with contemporaneous variables contains seven variables, the maximum number allowed.

The selected best predictor variable sets were subsequently used to determine classifiers' performance for the twelve stocks across the 169 test days. Separate classification accuracy statistics were calculated for each of the 16 training timescales. Fig. 1 shows mean accuracy curves for the WBC stock with the variable sets including contemporaneous variables. Each curve represents a single classifier and depicts mean classification accuracies for all training timescales. The chart for the WBC stock is a typical one. Eight other stocks in the data set have charts which qualitatively agree with it. Fig. 2 shows mean accuracy curves of the k-NN ($k = 9$) classifier with contemporaneous

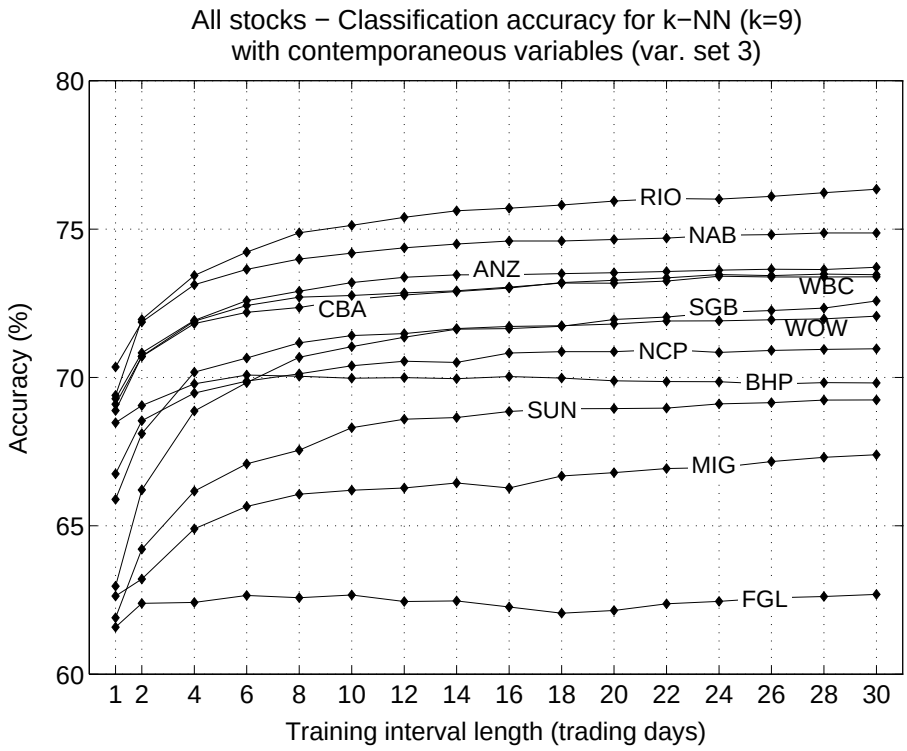

Fig. 2. Classifier performance vs. training interval length for the 12 stocks. The k-nearest-neighbor ($k = 9$) with contemporaneous variables - 169 test days.

variables, for the twelve stocks. The whole set of twelve charts like in Fig. 1 and the chart in Fig. 2 have a number of qualitative characteristics enumerated below, with numbers in brackets specifying how many stocks exhibit a given characteristic. The main characteristics are as follows:

(1) among the k-NN classifiers, the higher the value of k the greater the mean accuracy. The difference between accuracies for $k = 9$ and $k = 5$, however, can be minimal and sometimes negative, but on average $k = 9$ is the best (12).
(2) the mean accuracy of the k-NN classifier, where $k = 9$, is a monotonically increasing function of the training interval length. The rate of the increase, however, rapidly declines. Small, negligible fluctuations are sometimes present (10).
(3) the mean accuracy of the k-NN classifier, where $k = 9$, is greater than the mean accuracy of the logistic regression classifier for all training timescales (8).
(4) the mean accuracy of the k-NN classifier, where $k = 9$, is greater than the mean accuracies of the trade continuation and the majority vote classifiers, for all training timescales (12).

The best test statistics for individual stocks, calculated across the 169 test days, are presented in Table 2. For each classifier, the best statistics were de-

Table 2
Classification accuracy (%) for individual stocks - 169 test days. Abbreviated headings: CV - contemporaneous variables; SD - standard deviation; N_t - best training interval length. * - significant at $\alpha = 0.01$. Statistical significance was tested only for the k-NN ($k = 9$) with contemporaneous variables and the indicated N_t.

Stock	CV	k-NN ($k = 9$)			Logistic reg.			Trade cont.		Major. vote	
		Mean	SD	N_t	Mean	SD	N_t	Mean	SD	Mean	SD
NAB	Yes	74.87*	3.08	30	59.57	3.48	16	54.93	3.39	51.40	5.43
	No	57.95	2.08	30	55.95	3.28	18	-	-	-	-
BHP	Yes	70.08	4.25	6	70.34	3.77	10	59.22	4.30	55.18	9.58
	No	64.99	3.24	30	59.77	5.70	4	-	-	-	-
CBA	Yes	73.48*	2.25	28	58.74	3.36	18	54.32	2.70	52.99	5.54
	No	56.70	2.27	30	55.28	3.73	8	-	-	-	-
ANZ	Yes	73.72*	3.27	30	62.14	3.80	14	55.40	3.25	53.04	6.20
	No	59.01	2.72	30	56.12	3.85	20	-	-	-	-
WBC	Yes	73.41*	3.58	24	63.13	3.67	28	55.67	3.34	52.78	6.14
	No	59.11	2.46	28	56.38	3.30	30	-	-	-	-
NCP	Yes	70.97*	3.66	30	67.95	3.72	18	58.21	3.97	51.80	7.68
	No	63.34	3.03	30	58.82	3.58	18	-	-	-	-
RIO	Yes	76.34*	2.92	30	58.98	4.28	10	55.23	3.63	53.20	7.41
	No	58.27	3.12	22	56.88	4.16	30	-	-	-	-
WOW	Yes	72.07*	4.15	30	62.93	4.12	26	56.86	3.72	52.83	8.00
	No	58.72	3.30	28	57.64	4.13	28	-	-	-	-
FGL	Yes	62.69	6.08	30	66.72	5.67	24	61.38	5.09	56.87	11.15
	No	61.15	5.12	30	62.30	6.17	6	-	-	-	-
SGB	Yes	72.58*	4.91	30	60.78	5.14	24	58.37	4.90	53.55	9.31
	No	56.59	4.46	10	58.74	5.64	24	-	-	-	-
SUN	Yes	69.24*	3.87	30	60.24	4.25	24	57.19	4.11	52.52	8.55
	No	56.51	3.49	30	57.53	4.57	18	-	-	-	-
MIG	Yes	67.39	6.08	30	68.30	6.04	30	61.05	5.02	56.23	11.40
	No	62.06	5.52	28	61.49	5.66	24	-	-	-	-
Av.	Yes	71.40	4.01	27	63.32	4.27	20	57.32	3.95	53.53	8.03
value	No	59.53	3.40	27	58.08	4.48	19	-	-	-	-

termined by finding a training timescale (best interval) corresponding to the highest mean accuracy. The best training interval length is reported for the k-nearest-neighbor and the logistic regression. Statistics for the best k-NN classifier are included for $k = 9$ only. The trade continuation and the majority vote classifiers are not sensitive to a choice of predictor variables and have their duplicate statistics indicated by hyphens. Paired one-tailed t-tests with the Bonferroni adjustment for multiple comparisons ([29]) were applied to establish the statistical significance of differences in the means for each individual stock. The mean accuracy of the k-NN classifier with $k = 9$, contemporaneous variables and a respective best training interval length, was compared against

144 mean accuracies for other combinations of a classifier, variable set, and a training timescale [9] . For most of the stocks the mean accuracy of the specified k-nearest-neighbor model was found to be greater than the mean accuracies of all other models except for some k-NN classifiers with contemporaneous variables. This is true for 9 out of 12 stocks (marked with * in Table 2), at the significance level of 0.01. The lack of significance for comparisons between various k-NN classifiers with contemporaneous variables was most probably due to small differences in their classification accuracy, the large number of statistical tests performed, and the conservative assumptions of the Bonferroni adjustment.

The average best mean for the k-NN $(k = 9)$ classifier with contemporaneous variables and the best training timescale, across all stocks, is 71.40%. It is 8.08% higher than the corresponding average best mean for the logistic regression. It is also 14.08% and 17.87% higher than the corresponding averages for the trade continuation and the majority vote classifiers, respectively. The average statistics for the sets without contemporaneous variables are much less impressive. In particular, the average best means for the k-nearest-neighbor $(k = 9)$ and the logistic regression are both below 60%, just above the average mean for the trade continuation classifier (57.32%). The statistical significance of the individual results for the k-nearest-neighbor without contemporaneous variables was not determined in order to limit the total number of multiple comparisons.

As far as the best training interval length for the k-nearest-neighbor $(k = 9)$ is concerned, the average length is 27 days for both types of variable sets, with a value of 30 days for most of the individual stocks. As we mentioned above, in the case with contemporaneous variables these values are not statistically significant, while in the case without contemporaneous variables the significance tests were not performed. The classification accuracy for most of the stocks, however, appears to be a monotonically increasing function of the training interval length, as depicted in Fig. 2. These two characteristics, 30 days being the best training interval length and the monotonic increase of the classification accuracy, suggest that for most of our stocks the observed regularity in the trade sign may have a memory of at least 30 trading days. This value corresponds to the longest training timescale used in our experiment. A data set covering a longer period would be required to determine an upper bound on the memory duration.

[9] The total number of models constructed for each stock was 145: $2 * 3 * 16$ k-NN, $2 * 16$ logistic regression, 1 trade continuation, and 16 majority vote models.

5 CONCLUSIONS

We investigated a regularity in market order submission strategies on the Australian Stock Exchange. An empirical model for the trade sign inference was developed using transaction level data for twelve large stocks on the ASX. We proposed the k-nearest-neighbor classifier as an alternative to the linear logistic regression. The average classification accuracy of the k-NN ($k = 9$) classifier, across all stocks and allowing contemporaneous predictor variables, was found to be 71.40% (SD=4.01%), or 8.08% higher than the corresponding accuracy of 63.32% (SD=4.27%) for the logistic regression. When compared with the trade continuation and the majority vote classifiers, the k-nearest-neighbor was 14.08% and 17.87% better, respectively. The results for individual stocks show that the proposed k-NN classifier is better than the other three classifiers for most of the stocks, at the significance level of 0.01. The best k-NN model required only three predictor variables: total volumes at the best bid and ask in the order book just before a trade, and the contemporaneous trade size. In contrast, the best logistic regression required seven variables, the maximum allowed. These results suggest that a non-linear approach may produce a more parsimonious trade sign inference model with a higher out-of-sample classification accuracy. Furthermore, for most of our stocks the classification accuracy of the k-nearest-neighbor ($k = 9$) with contemporaneous predictor variables is a monotonically increasing function of the training interval length, with 30 days being the best interval. It appears that for these stocks the investigated regularity in market order submissions may have a memory of at least 30 trading days.

Further work, which is also the subject of our current research, could involve a detailed analysis of the k-NN classifier's performance for the best variable set found. This analysis should subsequently lead to a development of a parametric non-linear model. Such a parametric approach would provide a more quantitative insight into market order submission strategies employed by market participants. The parametric approach could also be more computationally efficient than the k-nearest-neighbor, which as a consequence should allow us to analyze more stocks on a longer data period. As far as commercial applications are concerned, it is not clear at this stage if the observed regularity in the trade sign can be profitably exploited. Some answers could perhaps be obtained by incorporating our model into the existing models of limit order execution and trading costs.

6 ACKNOWLEDGEMENTS

We acknowledge the provision of the ASX data and the data extraction software by Capital Markets Cooperative Research Centre (CMCRC) and its industry partners. A. Blazejewski gratefully acknowledges a CMCRC scholarship.

References

[1] B. Biais, P. Hillion, C. Spatt, An empirical analysis of the limit order book and the order flow in the Paris Bourse, Journal of Finance 50 (5) (1995) 1655–1689.

[2] Y. Hamao, J. Hasbrouck, Securities trading in the absence of dealers: trades and quotes on the Tokyo Stock Exchange, Review of Financial Studies 8 (3) (1995) 849–878.

[3] H.-J. Ahn, Y.-L. Cheung, The intraday patterns of the spread and depth in a market without market makers: The stock exchange of Hong Kong, Pacific-Basin Finance Journal 7 (5) (1999) 539–556.

[4] H.-J. Ahn, K.-H. Bae, K. Chan, Limit orders, depth, and volatility: Evidence from the stock exchange of Hong Kong, Journal of Finance 56 (2) (2001) 767–788.

[5] K. Hedvall, J. Niemeyer, G. Rosenqvist, Do buyers and sellers behave similarly in a limit order book? A high-frequency data examination of the Finnish stock exchange, Journal of Empirical Finance 4 (2-3) (1997) 279–293.

[6] L. E. Harris, Optimal dynamic order submission strategies in some stylized trading problems, Financial Markets, Institutions & Instruments 7 (2) (1998) 1–76.

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

[8] M. Potters, J. Bouchaud, More statistical properties of order books and price impact, Physica A: Statistical Mechanics and its Applications 324 (1-2) (2003) 133–140.

[9] I. Zovko, J. D. Farmer, The power of patience: a behavioural regularity in limit-order placement, Quantitative Finance 2 (5) (2002) 387–392.

[10] J. Bouchaud, Y. Gefen, M. Potters, M. Wyart, Fluctuations and response in financial markets: the subtle nature of "random" price changes, Quantitative Finance 4 (2) (2004) 176–190.

[11] F. Lillo, J. D. Farmer, The long memory of the efficient market, preprint cond-mat/0311053.

[12] P. Verhoeven, S. Ching, H. G. Ng, Determinants of the decision to submit

market or limit orders on the ASX, Pacific-Basin Finance Journal 12 (1) (2004) 1–18.
[13] C. M. C. Lee, M. J. Ready, Inferring trade direction from intraday data, Journal of Finance 46 (2) (1991) 733–746.
[14] M. Aitken, A. Frino, The accuracy of the tick test: Evidence from the Australian stock exchange, Journal of Banking & Finance 20 (1996) 1715–1729.
[15] K. Ellis, R. Michaely, M. O'Hara, The accuracy of trade classification rules: Evidence from Nasdaq, Journal of Financial and Quantitative Analysis 35 (4) (2000) 529–551.
[16] E. R. Odders-White, On the occurence and consequences of inaccurate trade classification, Journal of Financial Markets 3 (2000) 259–286.
[17] D. C. Porter, The probability of a trade at the ask - An examination of interday and intraday behavior, Journal of Financial and Quantitative Analysis 27 (2) (1992) 209–227.
[18] M. Aitken, P. Brown, H. Y. Izan, A. Kua, T. Walter, An intraday analysis of the probability of trading on the ASX at the asking price, Australian Journal of Management 20 (2) (1995) 115–154.
[19] J. D. Farmer, J. J. Sidorowich, Predicting chaotic time series, Physical Review Letters 59 (8) (1987) 845–848.
[20] D. Kugiumtzis, N. Lingjærde, N. Christophersen, Regularized local linear prediction of chaotic time series, Physica D 112 (1998) 344–360.
[21] J. Jiménez, J. A. Moreno, G. J. Ruggeri, Forecasting on chaotic time series: A local optimal linear-reconstruction method, Physical Review A 45 (6) (1992) 3553–3558.
[22] R. Gençay, Linear, non-linear and essential foreign exchange rate prediction with simple technical trading rules, Journal of International Economics 47 (1) (1999) 91–107.
[23] S. Zemke, Nonlinear index prediction, Physica A: Statistical Mechanics and its Applications 269 (1) (1999) 177–183.
[24] H. Akaike, Information theory and the extension of the maximum likelihood principle, in: B. N. Petrov, F. Csáki (Eds.), Second International Symposium on Information Theory, Akadémiai Kiadó, Budapest, 1973, pp. 267–281.
[25] B. G. Tabachnick, L. S. Fidell, Using multivariate statistics, 3rd Edition, HarperCollins College Publishers, 1996.
[26] B. LeBaron, A. S. Weigend, A bootstrap evaluation of the effect of data splitting on financial time series, IEEE Transactions on Neural Networks 9 (1) (1998) 213–220.
[27] B. V. Dasarathy (Ed.), Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, Los Alamitos, California, 1990.
[28] I. T. Nabney, NETLAB: Algorithms for Pattern Recognition, Springer, 2002.
[29] J. A. Rafter, M. L. Abell, J. P. Braselton, Multiple comparison methods

for means, SIAM Review 44 (2) (2002) 259–278.